Computational Models of Auditory Function

Editors:

Steven Greenberg
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704

Malcolm Slaney
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

To be published by IOS Press, Amsterdam, The Netherlands, in August 2001

This draft printed on August 12, 2001

Series Information from IOS

Title from IOS

ISBN/Copyright from IOS

Contents

Physiology of the Auditory Periphery

Modeling the Cochlea

Sound Localization and Binaural Mechanisms

The Bat as a Computational System

Auditory Processing of Pitch

Temporal Processing and Periodicity Analysis

Auditory Dynamics

Auditory Scene Analysis

Auditory Processing of Speech

PREFACE

"The purpose of computing is insight, not numbers."
Richard Hamming

Homo sapiens is considered to be, above all else, a visually oriented species, with the other senses viewed more as supporting players than as star performers. This "bias" in perceived sensory function has had the historical consequence of casting the auditory modality into the "back seat" of both experimental and computational neuroscience for many decades — or so it would appear.

In actuality, auditory neuroscience has been at the vanguard of both disciplines for many years. Important scientific landmarks include:

(1) the first biological application of Fourier's theorem [5]
(2) the first systematic application of electrical recording technology in sensory neurophysiology [8]
(3) the original use of micro-electrodes for recording the electrical activity of single neurons [2]
(4) the initial application of computer technology for presenting experimental signals [3]
(5) the original utilization of computers to collect and analyze neurophysiological data [4]
(6) the first application of entirely digital technology for signal presentation, data collection and analysis [6][7]
(7) the first apparent application of non-linear modeling to behavioral function [1]

As we start the twenty-first century (and the third millennium) the dawn of a new scientific era is approaching, one that melds traditional experimental and descriptive methodology with the emerging power of computational and quantitative approaches. The current volume serves to define the shape, texture and scope of this important, new field of scientific inquiry, as well as to delineate its likely technological contribution to such fields as telephony, automatic speech recognition, hearing prostheses, speech synthesis, high-quality voice/audio reproduction and transmission.

The volume is divided into nine sections, each focusing on a specific topic germane to computational hearing.

The first section discusses computational approaches to the physiology of the auditory periphery, ranging from the cochlea (the chapter by Gebeshuber and Rattay) to the auditory nerve (Stankovic) and up through the ventral cochlear nucleus (Kalluri and Delgutte).

The second section applies computational approaches to two areas germane to processing in the cochlea. Irino and Unoki describe a model for spectral analysis based on gammachirp filtering, while Bruce and colleagues describe models for processing of sound by cochlear implants used in the profoundly hearing-impairing.

The third section of the book focuses on the localization of sound from a variety of different perspectives. Brungart describes a model for the perception of auditory distance, while Ito and Akagi apply sophisticated computational techniques to the problem of sound localization in general. Hartung and Sterbing, in their chapter, use physiological data to predict behavioral performance.

The following section discusses one particular model system — the echolocating bat — as a means of melding computational approaches to behavior and physiology. Wotton and colleagues consider the cues used for computation of elevation, while Müller and Schnitzler discuss the concept of "acoustic flow" in bats.

Section five focuses on pitch perception from the behavioral (Akeroyd and Summerfield) and physiological (Cai and colleagues) perspectives.

Temporal processing and periodicity analysis has been a controversial area of research for over a century. The sixth section focuses on several issues germane to this topic. The first two chapters (by Heil, and by Bleeck and Langner) focus on the importance of the waveform envelope (particularly at the beginning of a signal) for evoking neural excitation. Unoki and Akagi, in their chapter, model the perceptual phenomenon of "co-modulation masking release," a topic of intense behavioral research over the past two decades. Finally, Cariani discusses the importance of neural networks specialized for extracting timing cues in the perception of pitch and timbre.

The seventh section contains a paper by Miller and colleagues that examines the relationship between the thalamic and cortical regions of the auditory pathway, using dynamic signals to deduce the interconnections between these parts of the brain.

Auditory scene analysis, the ability to pick out specific "objects" from a background based on acoustic cues, has been a topic of keen investigation over the past decade. Baumann describes a model for identification and segregation of musical tones. Denham proposes a model of cortical activity (and inhibition) as the basis for some of the segregation ability observed in human listeners. Meyer and colleagues examine the ability of listeners to segregate two streams of speech as an example of auditory scene analysis.

Much of the interest in auditory computational models pertains to their utility for speech processing. The final section of the book examines three different approaches to speech processing using auditory models. Strope and Alwan are concerned with potential robustness of the speech signal in noisy environments based on pitch-relevant, amplitude-modulation cues. Tian and colleagues apply a model of the auditory periphery for robust speech recognition by computer, while Kawahara uses an auditory-inspired model to create realistic talking voices.

This volume is based on a NATO Advanced Study Institute, held at Il Ciocco, in the mountains of Tuscany, between July 1–12, 1998. Over a hundred scientists, representing 17 countries in Europe, North America and Asia, participated in the meeting (for further details, see http://www.icsi.berkeley.edu/real/comhear). The ASI's intent was to provide a rigorous, scientific overview of auditory function in concert with a critical examination of specific strategic issues that potentially hold the key to understanding how the brain portrays the world in terms of sound. As far we know, this was the first scientific meeting to specifically focus on melding computational approaches with the traditional venues of auditory neuroscience and psychoacoustics.

We would like to express our appreciation and gratitude to the ASI Faculty (Jont Allen, Jens Blauert, Ellen Covey, Dan Ellis, Ted Evans, Phil Green, Hynek Hermansky, Gerald Langner, Roy Patterson, Christoph Schreiner, Shihab Shamma, Jim Simmons, Quentin Summerfield, Marianne Vater, Jeff Winer and Eric Young) for their excellent lectures and to all of the other participants for helping to make the meeting an outstanding success.

We would also like to express our appreciation to NATO, which provided the lion's share of funding required to support the meeting through its Office of Scientific and Environmental Affairs. A particular debt of gratitude is owed to Dr. L. Veiga da Cunha, then-director of the Advanced Study Institute Program at NATO, for his help and guidance. We are also grateful to the Turkish and Portuguese divisions of NATO for providing additional financial support to defray the travel expenses of participants from their respective countries.

The ASI also received generous financial support from the (U.S.) Office of Naval Research and the (U.S.) Air Force Office of Scientific Research, for which we are deeply appreciative. We thank, in particular, Dr. Harold Hawkins of ONR and Dr. John Tangney of AFOSR for their efforts in supporting the ASI. We are also grateful for the support of the (U.S.) National Science Foundation which subsidized the travel of several of the younger American scientists. We thank Ms. Rosa Knox and Dr. Robert Metcalfe of that foundation's Graduate Education and Research Development office for their support and assistance. We also thank Barbara and Tilo Kester of International Transfer of Science and Technology (Brussels) who work closely with NATO to provide the infrastructure support for Advanced Study Institutes. The Kesters suggested Il Ciocco in Tuscany as the site most ideal for the purposes of the ASI and helped to coordinate our initial contacts with the hotel and conference center. We are especially grateful to Bruno Gianassi and his staff at Il Ciocco, who continually went beyond the call of duty to insure that everything ran smoothly during the course of the meeting.

Finally, we would like to express our deepest appreciation to the authors for taking the time to prepare their chapters for this volume, as well to thank them for their patience and understanding during the lengthy preparation of the book.

References

[1] Chistovich, L. A., Kozhevnikov, V. A. and Alyakrinskii, V. V. *Speech, Articulation and Perception*. Nauka: Acad. Sci. U. S. S. R., 1965.

[2] Galambos, R. and Davis, H. "The response of single auditory-nerve fibers in acoustic stimulation." *J. Neurophysiol.,* 6: 39–58, 1943.

[3] Goldstein, M. H. Jr., Kiang, N. Y. S. and Brown, R. M. "Responses of the auditory cortex to repetitive acoustic stimuli." *J. Acoust. Soc. Am.,* 31: 356–364, 1959.

[4] Kiang, N. Y.-S. *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*. Cambridge, MA: M.I.T. Press, 1965.

[5] Ohm, G. S. "Über die definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher Tonbildener Vorrichtungen." *Ann. D. Phys.,* 59: 497–565, 1843.

[6] Rhode, W. S. "Neurophysiological techniques and the minicomputer." In *Use of Minicomputers in Research on Sensory and Information Processing*, M. S. Mayzner and T. R. Dolan (eds.), Hilldale, NJ: Lawrence Erlbaum, pp. 229–260, 1978.

[7] Rhode, W. S. and Olsen, R. *Digital Stimulus System for Auditory Neurophysiology*. Technical Report, Department of Neurophysiology, University of Wisconsin, Madison, WI, 1976.

[8] Wever, E. G. and Bray, C. W. "Action currents in the auditory nerve in response to acoustical stimulation." *Proc. Nat. Acad. Sci.,* 16: 344–350, 1930.

Steven Greenberg and Malcolm Slaney
May, 2001

PHYSIOLOGY OF THE AUDITORY PERIPHERY

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

PHYSIOLOGY OF THE AUDITORY PERIPHERY

Steven Greenberg
International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA

The auditory periphery, encompassing the outer, middle and inner ears, as well as the auditory nerve and ventral cochlear nucleus, has been the focus of intensive scientific investigation over the past half century. Much of the motivation for this research effort derives from an interest in the anatomical, physiological and biochemical bases of hearing impairment and potential methods for its amelioration (cf. the chapter by Bruce and colleagues on cochlear-implant research in this volume). Another reason for this interest lies in the fact that the auditory periphery is the "final common pathway" prior to anatomical and physiological diversification characteristic of the central auditory nervous system.

Despite the abundance of studies performed, many questions pertaining to the function of the auditory periphery remain unresolved. The papers in this section address three separate functional issues using quantitative methods.

Gebeshuber and Rattay use a computational model of the cochlea to investigate the origins of the audibility curve for human listeners. The conventional wisdom cites the middle ear as the primary basis for maximum sensitivity in the region between 2.5 and 5 kHz as well as for the steep decline in audibility below 400 Hz [4]. The authors suggest that other factors, such as firing-interval statistics of the auditory nerve, combined with temporal resolution of neuronal spiking due to Brownian motion in the cochlea proper, may also play an important role in the pattern of audibility observed. The authors also point out that the innervation density of auditory-nerve fibers projecting onto inner hair cells is highly correlated with the threshold sensitivity function; they propose that Brownian motion and stochastic resonance may underlie the extraordinary sensitivity of human listeners in the mid-frequency range of the spectrum. Damage to this functional component of the cochlea may account for certain types of hearing loss.

The chapter by Stankovic addresses another important issue in coding of acoustic signals. It has been known for many years that the input-output (I/O) function of auditory-nerve fibers can be approximated with a saturating form of non-linearity and a fixed threshold [6]. At very low sound-pressure levels the discharge rate of a fiber is governed by spontaneous activity. This discharge level is essentially the same as in the absence of sound. Some fibers exhibit very high levels of spontaneous activity (120 spikes/s), while others fire infrequently (<1 spike/s). The wide range of spontaneous activity may have functional implications, as the low- and medium spontaneous rate (<10 spikes/s) fibers generally have higher thresholds (by ca. 10-20 dB SPL) [1][3] and phase-lock with greater precision to low-frequency signals [2] than their high-spontaneous-rate counterparts. Moreover, there is evidence that at least some of the low-SR fibers exhibit a very broad dynamic range of response, with the saturating component of the I/O curve exhibiting a gently sloping character in contrast to the hard-limiting form of saturation observed in high-SR fibers [7].

In order to provide a more systematic means of characterizing the diversity of auditory-nerve rate–intensity functions, Stankovic develops a sophisticated means of selecting parameters for the equations used to fit a model's output with the empirical data derived from single-unit recordings from the cat auditory nerve. The method provides a simple, clear means with which to fit such I/O functions for a wide range of auditory-nerve fibers and thereby provides insight into the potential functional significance of the diversity associated with rate-intensity functions in the same experimental animal.

The auditory nerve projects to the ventral cochlear nucleus, the first site of manifest functional and anatomical specialization in the auditory pathway. In contrast to auditory-nerve fibers, which fire continuously during the coarse of stimulation, certain cell types in the ventral cochlear nucleus fire primarily at stimulus onset, and otherwise substantially reduce their firing rate or cease activity entirely [5]. Kalluri and Delgutte perform a modeling study of the onset units in the ventral cochlear nucleus, focusing specifically on the nature of auditory-nerve input impinging on such cells and the magnitude of synaptic strength binding the different neuronal cell types. In addition, they model certain properties of the onset (Octopus) cell membrane and show that certain specific properties of these cells' responses can be modeled on the basis of such parameters. The significance of their model pertains to the elegant manner in which it is able to deduce specific anatomical and physiological properties (cf. [5]) on the basis of a relatively sparse initial data set. It is notoriously difficult to obtain empirical data pertaining to cell-membrane characteristics and anatomical connectivity. Using a quantitative model to infer the relevant parameters can serve as an effective means of providing initial hypotheses about the anatomy and physiology which can subsequently be verified (or modified) during the course of focal investigations.

Together, the three chapters in this section provide a representative sample of how computational methods can advance the scientific study of the anatomy and physiology of the auditory periphery.

References

[1] Geisler, C. D., Deng, L. and Greenberg, S. "Thresholds for primary auditory fibers using statistically defined criteria." *J. Acoust. Soc. Am.*, 77: 1102–1109, 1985.

[2] Greenberg, S. "Possible role of low and medium spontaneous rate cochlear nerve fibers in the encoding of waveform periodicity." In *Auditory Frequency Selectivity*, B. Moore and R. Patterson (eds.), New York: Plenum, pp. 241–248, 1986.

[3] Liberman, M. C. "Auditory-nerve response from cats raised in a low-noise chamber." *J. Acoust. Soc. Am.,* 63: 442–455, 1978.

[4] Pickles, J. O. *An Introduction to the Physiology of Hearing* (2nd ed.). London: Academic Press, 1988.

[5] Rhode, W. S. and Greenberg, S. "Physiology of the cochlear nuclei." In *The Mammalian Auditory Pathway: Neurophysiology*, R. R. Fay and A. N. Popper, (eds.). New York: Springer Verlag, pp. 94–152, 1992.

[6] Sachs, M. B.and Abbas, P. J. "Rate versus level functions for auditory-nerve fibers in cats: Tone-burst stimuli." *J. Acoust. Soc. Am.*, 80: 1359–1363, 1974.

[7] Winslow, R. L. Barta, P. E. and Sachs, M. B. "Rate coding in the auditory nerve." In *Auditory Processing of Complex Sounds*, W. A. Yost and C. S. Watson (eds.), Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 212–224, 1987.

CODING EFFICIENCY OF INNER HAIR CELLS AT THE THRESHOLD OF HEARING

Ilse C. Gebeshuber[1] and Frank Rattay[2]

[1]Institut für Allgemeine Physik and [2]TU-BioMed
University of Technology, Vienna
Wiedner Hauptstraße 8-10/1145
A-1040 Wien, Austria

1. Introduction

The human hearing threshold curve for pure tones is a nonlinear function of frequency (Figure 1). The minimum sound pressure required for an audible sensation to occur is frequency dependent and spans approximately five orders of magnitude. In the highly sensitive region of the spectrum between 2 and 5 kHz an intensity as low as 10^{-12} W/m^2 is sufficient to evoke an audible sensation.

The current modeling study investigates the efficiency with which the mechanical stage of transduction in the inner ear is transformed into a neural form (i.e., mechano–electric transduction) for low-intensity sinusoidal signals across a range of frequencies. Our goal is to gain insight into the basic physiological properties that underlie the human hearing threshold curve.

The inner hair cells (Figure 2) play a key role in mechano-electric transduction. In the human inner ear are three rows of outer hair cells (OHCs) and one row of inner hair cells (IHCs). A primary function of the OHCs is to amplify the magnitude of low-intensity signals [1][16][24]. IHCs serve as the primary (if not the sole) conduit of frequency-selective information to the brain via their innervation of the auditory nerve. Deflection of the hair cell stereocilia modulates the probability of the cell's transduction channels opening and closing and is responsible for the voltage fluctuations observed within the cell. Such fluctuations provide a low-pass filtered "image" of the stereociliary displacement (along with additional stochastic components resulting from channel gating). In the "active" zones, at the base of the cell, sufficient depolarization of the receptor potential results in Ca^{2+}-induced neurotransmitter release. This transmitter release, if of sufficient magnitude, results in depolarization of proximal auditory-nerve fibers, resulting in the generation of action potentials that are propagated into the auditory brainstem.

The transduction mechanism is so sensitive that displacements resulting from stereociliary Brownian motion contribute significantly to the spontaneous discharge activity observed in highly sensitive (i.e., high spontaneous rate) auditory-nerve fibers [8][10][31]. Acoustically generated displacements less than the thermal motion of the stereocilia are sufficient to cause an audible sensation. The Brownian motion in this instance enhances the detection of weak signals via a mechanism known as "stochastic resonance" [9][10][31]. Stochastic resonance is based on nonlinear statistical dynamics through which information flow in a multistate system (such as the transduction channel of the inner hair cell or the all-or-none process of spike generation) is enhanced by the presence of optimized random noise [23].

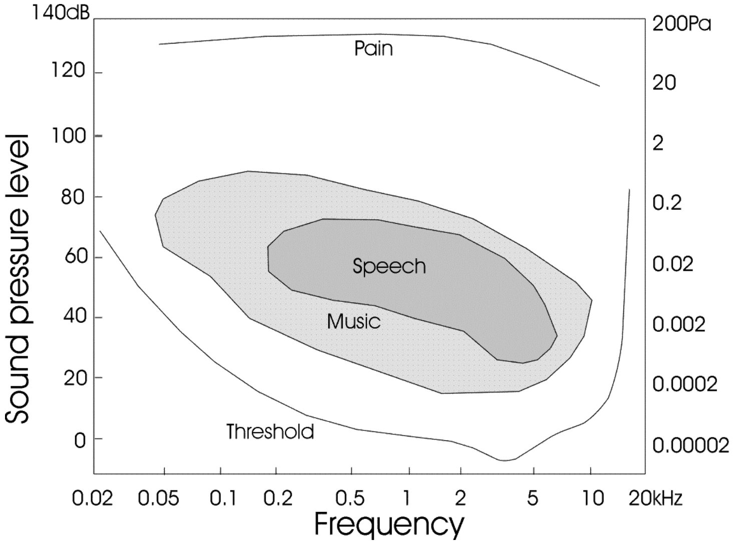


Figure 1 Region of human audibility (i.e., the range between the threshold of hearing and of pain). The intensity is scaled in dB, while the sound pressure is shown in Pascals. Note that the range of pressure variation covers 7 orders of magnitude. Adapted from [33].

Figure 3 shows the number of auditory-nerve fibers innervating an "average" IHC of a normal human cochlea. The innervation density varies as a function of cochlear-frequency position in a manner comparable to the hearing threshold curve (Figure 1). In the mid-frequency region the innervation density reaches a maximum of ca. 15 fibers per IHC. About sixty percent of these fibers are highly sensitive and exhibit relatively high rates of spontaneous activity (18–120 spikes/s).

The highly sensitive nerve fibers change their spiking patterns for low- and mid-frequency signals close to the threshold of hearing as follows: the first sign of influence on the firing of many spontaneously active fibers by a pure tone is phase-locking of the spikes [28] [29]. This may occur at an intensity far below that required to evoke an increase in mean firing rate [12]. This phase-locking effect does not occur for high-frequency signals as a consequence of the jitter associated with interspike times. Afferent fibers may respond differently each time a stimulus of a given amplitude is presented since fluctuations in excitability and latency are directly associated with fluctuations in the membrane resting potential [3].

Endogenous noise in the resting neural membrane potential of nerve fibers decreases with increasing diameter. The noise is on the order of 1 mV root-mean-square (r.m.s.) for myelinated fibers of small diameter and less than 1 mV for larger-diameter myelinated fibers [4]. The mean inner diameter of central axons of human auditory-nerve fibers has an unimodal distribution and ranges between 2.7 and 3.1 μm (with the exception of smaller fibers at the base of the cochlea) [30].

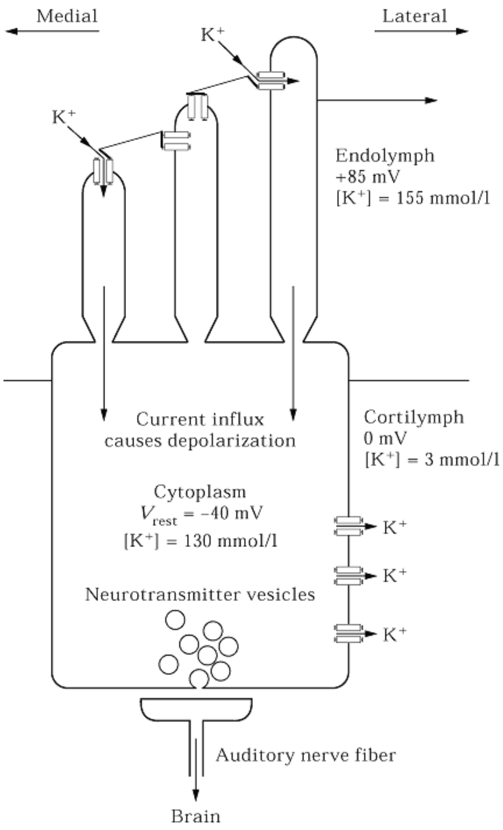


Figure 2 Schematic illustration of an inner hair cell. Endolymphatic fluid motion caused by the movement of the middle ear ossicles induces displacement of the stereocilia of the auditory receptor cell. The stereocilia of an inner hair cell are interconnected by links (elastic protein filaments). The open-close kinetics of transduction channels located close to the top of each stereocilium depend on stereociliary deflection (Figure 4). Even in the resting state the transduction channel open probability is about 15%. Due to potential gradients, ion currents (mainly potassium) enter the cell through the transduction channels and leave through ion channels in the cell body membrane, resulting in a resting potential of ca. –40 mV in the unstimulated hair cell and potential changes of several mV following stereociliary displacement. A potential change as low as 0.1 mV may cause neurotransmitter release and thereby evoke a spike in an auditory-nerve fiber. Note the tapering of the bottom portion of the stereocilia endings. In humans the inner hair cell stereocilia are arranged in a 20 by 3 matrix, with 20 short, 20 intermediate-length and 20 long elements. Each stereocilium behaves like a rigid rod pivoting around its insertion point into the cuticular plate.

2. Materials and Methods

2.1 Brownian Motion

The IHC stereocilia are interconnected by tip-links and horizontal links, and act like stiff rods capable of pivoting around their insertion point into the cuticular plate (Figure 2). The Brownian motion of the stereociliary tips is calculated using a reduced version of the stereocilia linear chain model [31]. The r.m.s. value of the modeled intrinsic bundle noise is ca. 2 nm, which is in accordance with experimental data [2]. The small amplitude of the fluctua-

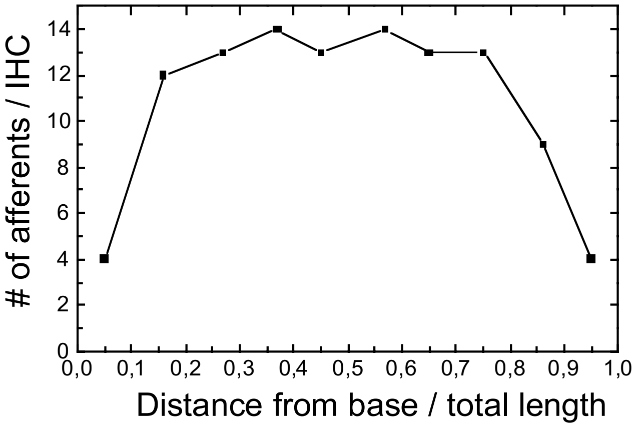


Figure 3 Innervation density per inner hair cell in a normal human cochlea (adapted from [7]). Sixty percent of the afferent nerve fibers are highly sensitive. The transformation from normalized distance, *d*, to the characteristic frequency, *f*, of the nerve fiber obeys the following relation in the human: f=200(10^{2d}-0.7) [13].

tions due to Brownian motion can be appreciated by comparing them to the dimensions of a single stereocilium (ca. 0.2 μm in diameter) or to the bundle's displacement-response relationship (Figure 4). In vestibular hair cells of the frog, viscous drag acting on the bundle limits Brownian motion to relatively low frequencies (200–800 Hz) [2]. However, theoretical considerations suggest a corner frequency of ca. 4 kHz for thermal fluctuations in mammalian hair cells [31], which implies that stochastic resonance may also be effective in the mid-frequency range of audition.

The overall displacement of the hair bundle in response to low-intensity signals is the sum of the bundle movements resulting from Brownian motion as well as from the signal-induced displacement. The signal-to-noise ratio is defined as the ratio of the r.m.s. magnitude of the signal and the r.m.s. level of the stereociliary deflections attributable to Brownian motion. Our

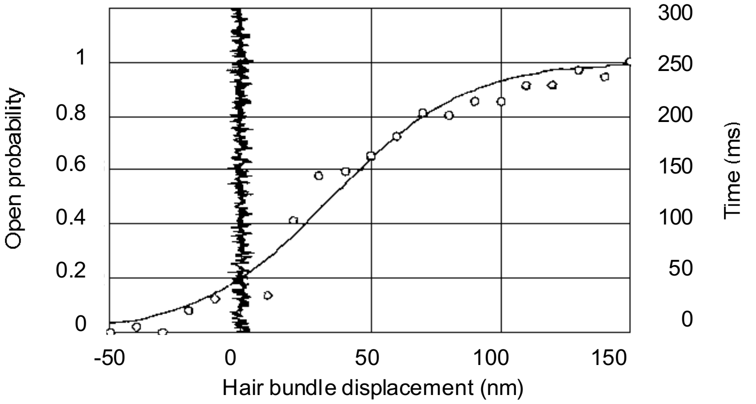


Figure 4 The relation of a 300-ms trace of simulated Brownian motion (low-pass filtered, 2 nm r.m.s. white noise) to a cell's displacement-response behavior. This function relates the probability of transduction channels being open (left y-axis) to the hair bundle displacement (x-axis). Note that in this specific case the transduction-channel, resting-open probability is 0.2. Adapted from [18].

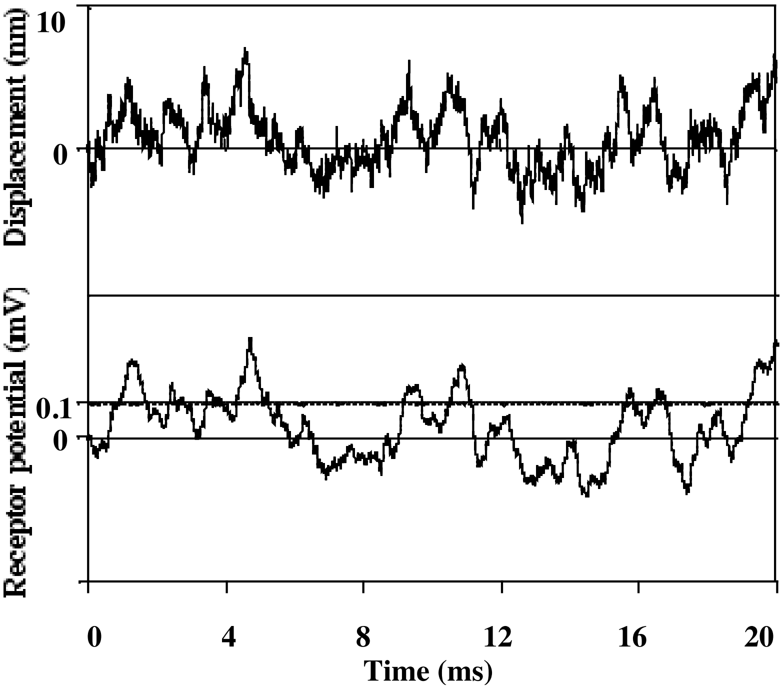


Figure 5 Modeled mechanical and electrical fluctuations due to Brownian motion: the intracellular receptor potential changes (bottom trace) are a low-pass filtered version of stereociliary displacements (top trace) with an additional amount of noise resulting from transduction channel kinetics.

simulation investigates the effects of signal-to-noise ratio (whose normalized range is between 0 and 1 — equivalent to stimulation of the hair bundle with an amplitude between 0 and 2.12 nm r.m.s.). The frequency of the stimulating, deterministic signals ranges between 0.2 and 20 kHz. Figure 5 shows a 20-ms time series of hair-bundle displacements resulting from Brownian motion.

2.2 Endogenous Transduction Channel Noise

The receptor potential fluctuations in the IHC are calculated using a model for the mechano–electrical transduction in inner hair cells [25]. The model uses equivalent electric circuits for cell membrane and cytoplasm (i.e., RC components and batteries). The kinetics of the transduction channels are modeled as Markov processes without memory: whether the channel stays open or closed depends only on its current open probability and not on the length of time the channel has already been open or closed.

For displacements in the range of a few nm, the relation between the stereociliary displacement and the open probability of the transduction channels is linear (Figure 4) [22]. For zero displacement the open probability of the transduction channel is about 0.15. For small displacements to the lateral side the transduction channel open probability increases, resulting in an influx of potassium ions. This influx causes a depolarization of the receptor potential from its resting state (ca. –40 mV). Displacement to the medial side decreases the open probability, resulting in fewer potassium ions entering the cell and a concomitant hyperpolarization of the membrane potential. Since the model's inner-hair-cell membrane time constant, τ, equals 0.255 ms [25], the IHC potential can be thought of as a low-pass-filtered version of the stereociliary displacement pattern combined with additional noise resulting from the stochastic components in channel gating (Figure 5).

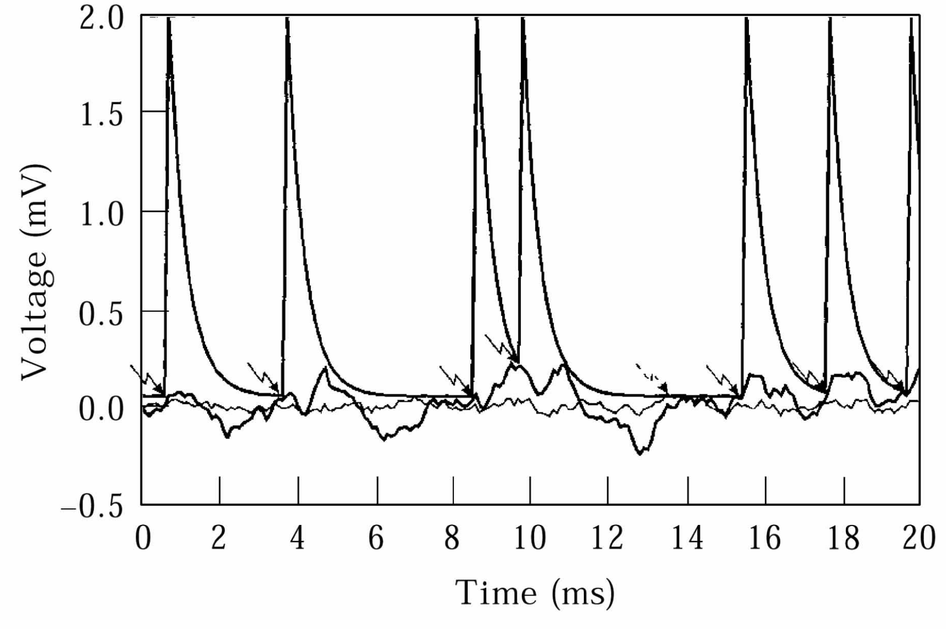


Figure 6 Simulated receptor potential changes and resulting firing behavior. The noise in the voltage fluctuations evoked by a weak 500-Hz signal alone (thin line, hypothetical case without Brownian motion) is a consequence of the endogenous transduction channel noise. Only in one instance (marked by a dashed arrow at 13.5 ms) are the fluctuations large enough to reach the threshold of spiking at 0.1 mV. The compound fluctuations caused by the same sinusoidal signal, the endogenous transduction channel noise and the thermal fluctuations with a signal-to-noise ratio of 0.2 show the enhancing effect of the noise: Seven spikes may occur within 20 ms among associated auditory-nerve fibers. The recovery behavior after spiking is modeled by an exponential decay of the threshold curve. As soon as the voltage fluctuations exceed threshold a new spike can occur.

2.3 Jitter in the Spiking Times: Refractory Period

The spike-generation process is modeled in the following way. Whenever the voltage fluctuations of the IHC exceed a threshold of 0.1 mV (a value sufficient for neurotransmitter release in hair cells [15]), a spike may be generated in an afferent fiber. Because of the stochastic nature of spike generation the probability for spiking is adjusted to obtain a mean spontaneous discharge rate of ca. 100 spikes/s in the resting state [27]. Jitter in the firing pattern is modeled by a single-sided, normally distributed time shift whose standard deviation is 50 μs. Since the absolute refractory period of an auditory-nerve fiber is ca. 0.8 ms (in the cat [19]), the time constant of the exponentially decaying threshold curve is set to 0.25 ms and the maximum value for the height is set to 2 mV (Figure 6).

The spike rate associated with a just-supra-threshold signal does not exceed the spontaneous rate of 100 spikes/s. However, nerve impulses become increasingly phase-locked to the acoustic signal as the signal level increases [10][11][12][14][28][29].

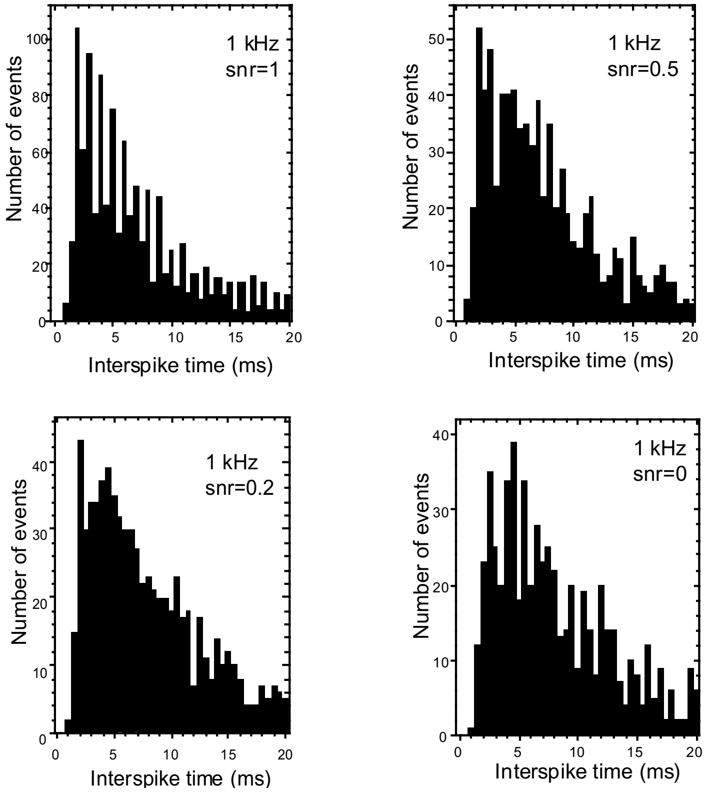


Figure 7 ISIHs for a 1-kHz signal at several signal-to-noise ratios. With increasing SNR, the proportion of spikes in a specific half of the signal period increases relative to the other half. Signal duration is 1s. The model's output represents the activity of 12 highly sensitive nerve fibers. Histogram binwidth is 0.5 ms.

3. Results

In this section we present an analysis of the frequency information encoded in the interspike interval histograms (ISIHs) of simulated auditory-nerve firing patterns induced by low-intensity, sinusoidal hair bundle deflections.

When the histogram binwidth is precisely half the period of the stimulating signal, phase-locking of the interspike times can be readily observed in the ISIHs as an up-down-up-down pattern. The distribution of spikes is non-uniform across time, being concentrated in a restricted portion of the stimulus cycle. This phase-locked behavior is manifested in the interspike interval histogram in the form of modes associated with intervals that are integral multiples of the stimulus period. The maximum interspike time considered in our model is 20 ms. Figure 7 shows ISIHs for stereociliary displacements at 1 kHz. With increasing SNR, the spikes tend to occur increasingly in the first half of the stimulus period (i.e., the phase-locking effect becomes increasingly apparent). A means to assess the information contained in the ISIH is to measure the ratio of spikes occurring during the positive half-wave of the stimulating signal relative to the total number of spikes. This ratio is a measure of the proportion of informative spikes, and has been used as a metric of phase-locking performance [28].

With decreasing signal frequency the number of bins (and therefore the fine structure information) associated with the ISIH decreases (Figure 8). Although the number of infor-

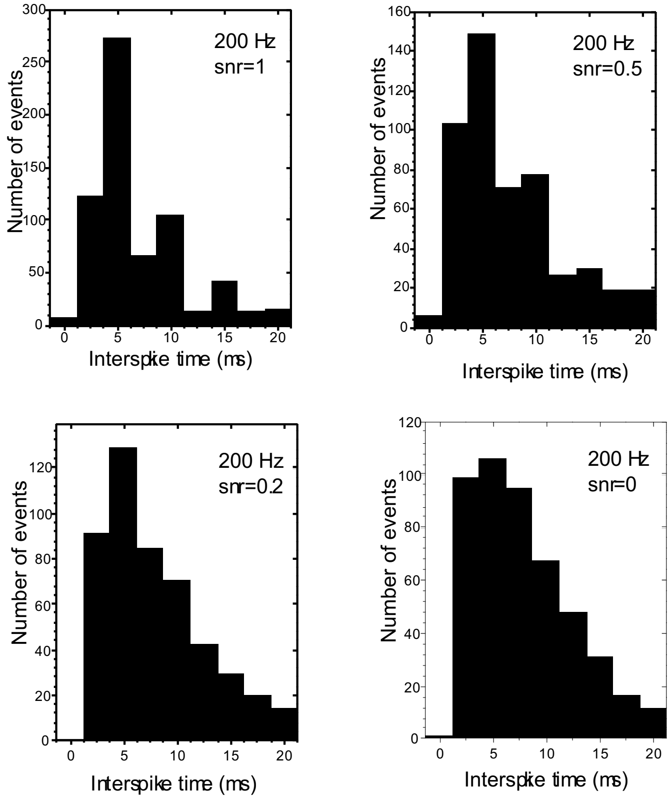


Figure 8 ISIHs for a 200-Hz signal at several signal-to-noise ratios. Signal presentation time is 1s. 5 highly sensitive nerve fibers, binwidth is 2.5 ms.

mative spikes is still greater than 50% for a 200-Hz signal with an SNR of 0.2, the number of events in each bin tend to decrease exponentially as occurs when the SNR is 0. The fine structure in the histogram, with valleys associated with integral multiples of the negative half-wave of the stimulating signal, is lost.

In a previous study we had analyzed the information contained in the ISIHs with artificial neural networks [9] [26]. In the mid-frequency range the neural net accurately detected the signal more than 75% of the time for signal-to-noise ratios as low as 0.1. Taking into consideration the parallel information transfer from several IHCs to the central nervous system considerably reduces the signal duration required to accurately detect the presence of a signal.

The information contained in the ISIHs is evaluated by calculating the number of informative spikes over a range of frequencies and signal-to-noise ratios. The resulting frequency-response-efficiency tuning curves are illustrated in Figure 9. The curves for low SNRs may be thought of as analogous to human hearing threshold curves for pure tones, as they reflect the combined effects of stereociliary Brownian motion, endogenous hair cell noise, the stochastic nature of neurotransmitter release and the innervation density of primary auditory afferents in various frequency bands. Note that this simulation study models threshold curves as they are reflected in the transduction of small sinusoidal displacements of the stereocilia. Furthermore, any possible effect of inhibitory efferent innervation of afferent nerve fibers (see e.g. [5] [6]) on the spiking pattern is neglected due to an absence of experimental data.

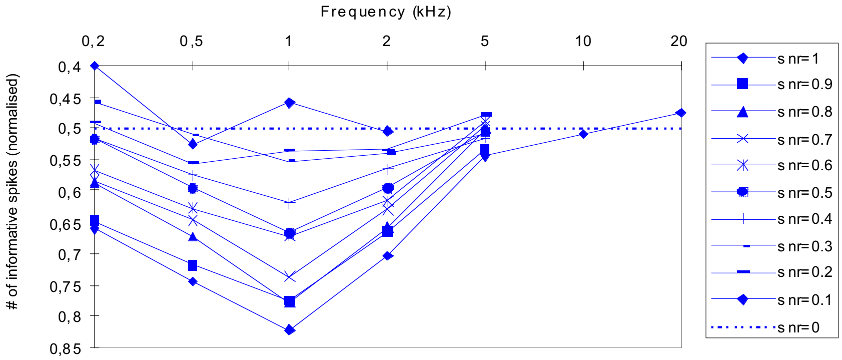


Figure 9 Frequency-response efficiency tuning curves for a multicellular model of peripheral auditory coding, (i.e. normalized number of informative spikes for several frequencies and signal-to-noise ratios). For an SNR of 0 there is no signal present and half of the spikes are phase-locked at chance level (i.e. the normalized number of informative spikes is 0.5). For frequencies between 0.2 kHz and 2 kHz, the phase-locking effect increases with increasing SNR (i.e., the normalized number of informative spikes increases well above 0.5). In the 5 kHz case there is virtually no apparent phase-locking. The 10 kHz and 20 kHz cases show no effect of phase-locking at all. In such instances only the normalized number of informative spikes for an SNR of 1 is presented. For high signal-to-noise ratios the curves are V-shaped. When the signal is reduced in amplitude and the influence of noise increases, the curves broaden and eventually invert at 1 kHz. Stimulus duration for each data point is 1s. For information concerning frequency-dependent innervation density cf. Figure 3.

At 1 kHz (the frequency which can be encoded and decoded optimally under the present conditions) a reversal of the shape of the curve appears at very low signal-to-noise ratios. When the signal level increases, the curves invert and become sharper. This effect corresponds to experimental results observed in noise-induced tuning-curve changes in mechanoreceptors of the rat foot [17]. Modeling the transduction channel kinetics as a Markov process results in a frequency-dependent peak-to-peak receptor potential. For low and high frequencies, the sub-threshold deterministic stimuli elicit voltage changes further from threshold than ones evoked by mid-frequency stimuli. Therefore, the optimal noise level is also frequency-dependent and the inversion of the tuning curve for low SNR stimuli is directly related to the threshold shift (cf. Figure 9 in [17]).

The 2-kHz case is comparable to the 1-kHz case. Increasing the SNR increases the number of informative spikes from approximately (but just higher than) chance level for an SNR of 0.1 to over 70% for an SNR of 1.

For high-frequency signals the jitter in the nerve firing pattern destroys the fine structure in the ISIH. However, statistics of the discharge pattern over a longer period would still contain some temporal information germane to 5 kHz, at least for signal-to-noise ratios close to one. For signals in the range of 10–20 kHz, increasing the signal does not further increase the number of informative spikes since the jitter completely destroys the phase-locking information. Therefore, the psychophysical hearing threshold data for the high-frequency portion of the spectrum cannot be attributed to phase-locking. This means that for high-frequency signals frequency information must be coded in a different way. The increase in spike rate is the most likely candidate for providing this information.

4. Discussion

In this study we have shown that a compound model of coding efficiency of inner hair cells at the threshold of hearing accounts for certain properties of the psychophysically measured human hearing threshold curve (Figure 1). Through the mechanism of stochastic resonance the Brownian motion of IHC stereocilia makes otherwise undetectable low-intensity signals audible. The jitter in auditory-nerve fiber spike times accounts for the steep slope in the threshold curve at high frequencies. In the low-frequency portion of the spectrum the long interspike times prevent detection of the signal, especially at low signal-to-noise ratios.

Changes in hair-bundle morphology also affect the pattern of thermal fluctuations of the stereocilia and therefore exert some influence on spontaneous activity in auditory-nerve fibers. In milder instances of acoustic trauma, morphological changes are only found in the rootlets of the stereocilia (which appear less dense in electron micrographs) [21]. In more severe instances of trauma (typically resulting in permanent damage) kinks or fractures at the rootlet of the stereocilia, and the packed actin filaments (which impart the stereocilia with their rigidity) are depolymerized [20] [32]. Within the IHC tuft the damage to the tall, outer row of stereocilia is often selective; the shorter rows may remain ultrastructurally normal even when the tallest row is completely missing. Moreover, the tip links remain intact on the shorter stereocilia, suggesting that such IHCs may be capable of transduction, but with reduced sensitivity. Auditory-nerve fibers associated with such IHCs exhibit much lower rates of spontaneous activity [21]. Following acoustic overstimulation, tuning curves with elevated "tips" and "tails" are associated with significant decreases in mean spontaneous discharge rate, whereas tuning curves with elevated tips but hypersensitive tails are associated with a clear elevation of the mean spontaneous rates [21]. Our model, in which altered Brownian motion patterns of the stereocilia lead to changes in the spiking pattern, may help to account for the occurrence of such pathological spiking patterns. However, one should bear in mind that in hearing loss of cochlear origin there are other noise-induced changes, such as different steady-state Ca^{2+} concentrations, that are the result of altered Ca^{2+} pump kinetics. Such changes may also be responsible for the pathological spiking patterns.

Future studies of the coding efficiency of inner hair cells at the threshold of hearing should take into consideration the possibility that Brownian motion of the stereocilia changes along the tonotopic axis and may be tuned in such a way as to enhance the audibility of specific frequencies. Such studies should also carefully consider the potential significance of the adaptation process of mammalian-transduction-channel kinetics, as well as the stochastic-resonance phenomena that have recently been demonstrated in transduction channels [18] and in calcium-activated potassium channels in the basolateral IHC membrane (Jaramillo, personal communication).

Acknowledgements

We thank Tony Wenzelhuemer for critically reading the manuscript and Alice Mladenka, Richardson Naves Leao, Juliana Pontes Pinto and Wolfgang Andreas Svrcek-Seiler for their participation in the preliminary studies. We also thank Jonathan Ashmore, Fernán Jaramillo, Steve Greenberg and Ted Evans for discussions.

References

[1] Ashmore, J. F. and Kolston, P. J. "Hair cell based amplification in the cochlea." *Curr. Opin. Neurobiol.*, 4: 503–8, 1994.

[2] Denk, W., Webb, W. W. and Hudspeth, A. J. "Mechanical properties of sensory hair bundles are reflected in their Brownian motion measured with a laser differential interferometer." *Proc. Nat. Acad. Sci.*, 16: 5371–5375, 1989.

[3] Derksen, H. E. "Axon membrane voltage fluctuations." *Act. Physiol. Pharmacol. Neerl.*, 13: 373–466, 1965.

[4] Derksen, H. E. and Verveen, A. A. "Fluctuations of resting neural membrane potential." *Science*, 151: 1388–1389, 1966.

[5] Ehrenberger, K. and Felix, D. "Glutamate receptors in afferent cochlear neurotransmission in guinea pigs." *Hear. Res.*, 52: 73–80, 1991.

[6] Felix, D. and Ehrenberger, K. "The efferent modulation of mammalian inner hair cell afferents." *Hear. Res.*, 64: 1–5, 1992.

[7] Felix, H., Gleeson, M. J., Pollak, A. and Johnsson, L. "The cochlear neurons in humans." *Progr. Hum. Audit. Vestib. Histopath.,* S. Iurano and J. E. Veldman (eds.), Amsterdam: Kugler, pp. 73–79, 1997.

[8] Gebeshuber, I. C., Mladenka, A., Rattay, F. and Svrcek-Seiler, W. A. "Computational demonstration that Brownian motion of inner hair cells stereocilia may enhance the ability to detect low level auditory tones from auditory nerve spiking patterns." *J. Physiol. (Lond.)*, 504P: 127P–128P, 1997.

[9] Gebeshuber, I. C., Mladenka, A., Rattay, F. and Svrcek-Seiler, W. A. "Brownian motion and the ability to detect weak auditory signals." In *Chaos and Noise in Biology and Medicine,* M. Barbi and S. Chillemi (Eds.), World Scientific: Singapore, pp. 230–237, 1999.

[10] Gebeshuber, I. C. "The influence of stochastic behavior on the human threshold of hearing." *Chaos, Solitons & Fractals*, 11: 1855–1868, 2000.

[11] Gebeshuber, I. C., Pontes Pinto, J., Naves Leao, R., Mladenka, A. and Rattay, F. "Stochastic resonance in the inner ear: the effects of endogenous transduction channel noise and stereociliary thermal motions on the human hearing threshold in various frequency bands." *ARGESIM Rep. 10: Proc. TU-BioMed Minisymp. 1998 "Brain Modelling"* F. Rattay (Ed.), pp. 40–44, 1998.

[12] Gleich, O., Narins, P. M. "The phase response of primary auditory afferents in a songbird (Sturnus vulgaris L.)." *Hear. Res.*, 32: 81–91,1988.

[13] Greenwood, D. D. "A cochlear frequency–position function for several species — 29 years later." *J. Acoust. Soc. Am.*, 87: 2592–2605, 1990.

[14] Hind, J. E. "Physiological correlates of auditory stimulus periodicity." *Audiology*, 11: 42–57, 1972.

[15] Hudspeth, A. J. "How the ear's works work." *Nature*, 341: 397–404, 1989.

[16] Hudspeth, A. J. "Mechanical amplification of stimuli by hair cells." *Curr. Opin. Neurobiol.*, 7: 480–486, 1997.

[17] Ivey, C., Apkarian, A. V. and Chialvo, D. R. "Noise-induced tuning curve changes in mechanoreceptors." *J. Neurophysiol.* 79: 1879–1890, 1998.

[18] Jaramillo, F. and Wiesenfeld, K. "Mechanoelectrical transduction assisted by Brownian motion: A role for noise in the auditory system." *Nature Neurosci.*, 1: 384–388, 1998.

[19] Javel, E. "Acoustic and electrical encoding of temporal information." In *Cochlear Implants — Models of the Electrically Stimulated Ear,* J. M. Miller and F. A. Spelman (eds.), New York: Springer–Verlag, pp. 247, 1990.

[20] Liberman, M. C. "Auditory-nerve response from cats raised in a low-noise chamber." *J. Acoust. Soc. Am.*, 63: 442–455, 1978.

[21] Liberman, M. C. and Dodds, L. W. "Single-neuron labeling and chronic cochlear pathology. II. Stereocilia damage and alterations of spontaneous discharge rates." *Hear. Res.*, 16: 43–53, 1984.

[22] Markin, V. S., Jaramillo, F. and Hudspeth, A. J. "The three-state model for transduction-channel gating in hair cells." *Biophys. J.*, 64: A93, 1993.

[23] McNamara, B. and Wiesenfeld, K. "The theory of stochastic resonance." *Phys. Rev. A*, 39: 4854–4869, 1989.

[24] Nobili, R., Mammano, F. and Ashmore, J. F. "How well do we understand the cochlea?" *Trends Neurosci.*, 21: 159–167, 1998.

[25] Rattay, F., Gebeshuber, I. C. and Gitter, A. H. "The mammalian auditory hair cell: A simple electric circuit model." *J. Acoust. Soc. Am.*, 103: 1558–1565, 1998.

[26] Rattay, F., Mladenka, A. and Pontes Pinto, J. "Classifying auditory nerve patterns with neural nets: A modeling study with low level signals." *Sim. Pract. Theor.*, 6: 493–503, 1998.

[27] Relkin, E. M. and Doucet, J. R. "Recovery from prior stimulation. I: Relationship to spontaneous firing rates of primary auditory neurons." *Hear. Res.*, 55: 215–222, 1991.

[28] Rose, J. E., Brugge, J. F., Anderson, D. J., Hind, J. E., "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey." *J. Neurophysiol.*, 30: 769–793, 1967.

[29] Rose, J. E., Hind, J. E., Anderson, D. J., Brugge, J. F., "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey." *J. Neurophysiol.*, 34: 685–699, 1971.

[30] Spoendlin, H. and Schrott, A. "Analysis of the human auditory nerve." *Hear. Res.*, 43: 25–38, 1989.

[31] Svrcek-Seiler, W. A., Gebeshuber, I. C., Rattay, F., Biró, T. and Markum, H. "Micromechanical models for the Brownian motion of hair cell stereocilia." *J. Theor. Biol.*, 193: 623–630, 1998.

[32] Tilney, L. G., Saunders, J. C., Egelman, E. and DeRosier, D. J. "Changes in the organization of actin filaments in the stereocilia of noise-damaged lizard cochleae." *Hear. Res.*, 7:181–197, 1982.

[33] Zwicker, E. *Psychoakustik*. Berlin: Springer–Verlag, p. 34, 1982.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

A METHOD FOR EVALUATION OF MULTIPARAMETER NONLINEAR MODELS ILLUSTRATED ON A COMPUTATIONAL MODEL OF AUDITORY-NERVE RATE–LEVEL CURVES

Konstantina M. Stankovic

Harvard–M.I.T. Division of Health Sciences and Technology, Harvard Medical School
Eaton Peabody Laboratory of Auditory Physiology, Massachusetts Eye and Ear Infirmary
243 Charles Street, Boston, Massachusetts, 02114, USA

1. Introduction

In many areas of computational hearing, selection of mathematical models is typically based on predictive power and physiological plausibility. Many of these models have a relatively large number of parameters, many of which are physiologically interpretable. A vital issue in establishing utility of a model is whether parameters can be reliably estimated from available experimental data. Traditionally, this has been addressed by studying the sensitivity of model outputs to variations in individual parameters. Such analyses have been performed using iterative simulations, which are computationally intensive — especially for models with a large number of parameters — and often offer little insight.

A method has been recently developed [11][1] to quantify the ability to estimate parameters in models that use a least-square-error criterion and a nonlinear parameterization. (A linear parameterization constitutes a special case that was solved earlier, e.g. [3]). This chapter describes an application of the so-called component-wise condition numbers for nonlinear least-squares problems to a commonly employed model in computational hearing: the model of auditory-nerve fiber (ANF) rate–level curves proposed by Sachs and coworkers [7][8]. The model was applied to responses from cat ANFs to tones at the fiber's characteristic frequency (CF). When compared with common practice, the subset selection method has clear advantages in evaluating model-parameters in nonlinearly parameterized problems that use the minimum least-squares criterion.

2. Background

Parameter evaluation is a key step in establishing the validity of a mathematical model. The *n*-dimensional estimate $\hat{\xi}$ of the (unknown) parameter vector ξ is chosen to minimize the discrepancy between model predictions $y(\hat{\xi})$ and actual measurements y (both N-dimensional). This optimization problem is affected by both the model structure and the error criterion, $r(\xi)=y(\hat{\xi})-y$. Very often, models are *nonlinear* functions of the parameters and the error measure, known as the "error-criterion value" is the sum of squared residual errors, $V(\xi) = \frac{1}{2}\|r(\hat{\xi})\|^2$. In such cases, estimation procedures usually involve iterative methods. The most frequently used such method is Gauss-Newton iterated linearized least squares.

A major concern in nonlinear least-squares estimation is that the measurements may not be rich enough to adequately reflect the individual effects of all parameters. This property —

referred to as ill-conditioning — may manifest itself in a slow convergence of the Gauss–Newton procedure. (Note that this ill-conditioning is intrinsic to the problem at hand — because it arises from the mismatch between the model detail and the measurement richness — and is not influenced by the numerical method.) A promising strategy for overcoming this problem is the *subset-selection method* [11][1], which partitions model parameters into: (1) well-conditioned ones (i.e., those that are likely to be estimated reliably from the given measurements), and (2) ill-conditioned ones (i.e., those whose estimates are likely to be unreliable, and whose presence makes the estimation problem very sensitive).

An efficient solution is to fix the ill-conditioned parameters at prior estimates (determined from e.g., physiological reasoning), and to solve a reduced-order problem containing only the well-conditioned parameters. While this procedure can introduce bias, the bias is often more than compensated for by the improvement in estimation of the remaining parameters.

2.1 Case Illustration: Failure of the Iterative Procedure

As an introduction to the subset-selection method we first illustrate a case where a traditionally used iterative procedure fails in unambiguously determining parameters from the data [1]. Since the Gauss-Newton method involves iterated linearization of the original nonlinear problem around the current best guess for parameter values, the N × n gradient (or Jacobian) matrix $J(\xi)= \partial r(\xi) / \partial\xi$ is a key mathematical object. It is used for determination of the Gauss-Newton direction in parameter space during the iterative procedure, and for (approximate) calculation of the matrix of second derivatives (Hessian) which is — for small residual error, $r(\xi)$ — well approximated with $H = J(\xi)' J(\xi)$ (where prime denotes the transpose). The eigenvalues of H correspond to the curvature of the error criterion in the directions of associated eigenvectors. If H (which is positive semidefinite by definition) happens to be singular, with exactly one of its eigenvalues equal to zero, this is equivalent to J having only n-1 independent columns. Then the iteration step can be varied in the direction of the eigenvector corresponding to the zero eigenvalue of H without affecting the error criterion (except possibly in the higher order terms). In parameter estimation, this implies that parameters cannot be determined unambiguously from the data. An inability to unambiguously estimate parameters from data may exist even if the model predictions fit the measurements exactly; this inability involves second partials of the error-criterion function V, not the function V itself. In practice, H is usually not exactly singular, but it may be nearly singular. This corresponds to near indeterminacy of parameters, and also complicates the estimation process.

Nearness to singularity is often characterized by the *condition number*, κ, which is the ratio of the largest (σ_1) and smallest (σ_n) singular values of $J(\xi)$: $\kappa = \sigma_1/ \sigma_n$. We refer to the condition number as the *maximal ratio of singular values* to make the term more intuitive. If some singular values of J are exceedingly small (making κ very large), the error criterion varies very slowly in the corresponding direction and, consequently, the parameter vector is poorly determined in that direction. A high maximal ratio of singular values can also result in a large number of iterations required for convergence of the Gauss-Newton algorithm.

The subset selection algorithm [11][1] determines which parameter axes lie closest to the ill-conditioned directions of the Hessian, and fixes these parameters to prior estimates. Assuming that there are ρ well-conditioned parameters (and n-ρ ill-conditioned ones), the parameter estimation is based on the reduced Jacobian, J_ρ. Since a combinatorial search to obtain a Hessian with the minimal condition (given ρ) is computationally prohibitive, it is replaced in the subset-selection algorithm [11][1] with the determination of a permutation matrix (from the QR decomposition) that yields the ρ-dimensional set of good parameters.

2.2 Multiparameter Computational Model of Auditory-nerve Rate–level Curves

A common way to characterize responses from ANFs is to plot the discharge rate of action potentials as a function of increasing sound level. These "rate–level curves" are monotonic for cat ANFs whose characteristic frequency (CF) is > 9 kHz [5]. Shapes of the monotonic rate–level curves span a continuum ([7][12]) that ranges from "flat saturation" (where the firing rate saturates at high sound levels) through "sloping saturation" (where the firing rate continues to grow, albeit slowly, even at the highest tested sound levels) to "straight" rate–level curves (where firing rate continues to grow steeply even at the highest tested sound levels).

A computational model that successfully fits a wide range of monotonic experimental rate–level curves was developed by Sachs and coworkers [7][8], and has since been commonly employed in the field of computational hearing (reviewed by [2]). As noted by the authors, the model can be viewed "strictly as phenomenological curve-fitting" [8]. However, the original authors, as well as later investigators (e.g., [9]) who analyzed an expanded version of the model (e.g., [10]), noted that some aspects of the model may be related to the underlying motion of the basilar-membrane and hair-cell transduction. Nevertheless, before parameter estimates of this (or any other) model are analyzed for their potential physiological significance, it is critical to identify which parameters can be reliably estimated from the data. The subset selection method is an invaluable tool in accomplishing this goal.

The computational model of rate–level curves assumes that the dependence of firing rate, R, on sound-pressure level at the tympanic membrane, P, can be described as:

$$R = R_{SP} + R_M \frac{(\widehat{P}/\Theta_E)^{\gamma}}{1 + (\widehat{P}/\Theta_E)^{\gamma}}$$

$$\widehat{P} = P\left[\frac{1}{1 + \left(\frac{P}{\Theta_1}\right)^{\beta}}\right]^{\alpha} \qquad (1)$$

where R_{SP} is spontaneous rate of firing (spikes/s), R_M is the maximal range of firing (saturation rate — spontaneous rate) in spikes/s, θ_E is the sound pressure (dynes/cm^2) at which driven rate reaches 1/2 of its maximal value, and θ_I is the sound pressure (dynes/cm^2) of the "compression threshold"; α, β and γ are dimensionless exponents that are set to $\alpha = 1/3$, $\beta = 2$ and $\gamma = 1.77$ in the original model [7][8].

This chapter provides an analysis of parameter estimation for both the original rate–level curve model [7][8] that depends on four parameters [R_M, R_{SP}, θ_E, θ_I], as well as the most general form of the model that depends on seven parameters [R_M, R_{SP}, θ_E, θ_I, α, β, γ].

3. Methods

3.1 Auditory-nerve Fiber Recording

Rate–level curves were recorded from cat auditory-nerve fibers in response to tones at the fibers' characteristic frequency. Details of the stimulation paradigm and recording set-up are provided in [4]. The rate–level curves analyzed in this study are those in the absence of efferent stimulation of Guinan and Stankovic [4] that are averages of repeated measurements (≥ 2 for a given fiber, so to decrease inherent noise in the data). All (total of 53) but one of these rate–level curves had monotonic growth of firing rate with sound level.

3.2 Subset Selection Method

The subset selection algorithm [11] [1]: (1) identifies a direction in a parameter space along which the error-criterion function does not change, (2) identifies which parameter(s) are most closely aligned with this direction, and (3) fixes these parameters to prior estimates. Key steps of the subset-selection algorithm for nonlinear least-squares parameter estimation [11][1] can be summarized as follows:

- Given an initial parameter-vector estimate ξ_0, compute the singular value decomposition of the Jacobian $J(\xi_0)$, yielding J=U S V', where U and V are unitary matrices, and S is a diagonal matrix with non-negative entries.
- *Determine ρ such that the first ρ singular values of J (i.e., diagonal entries of S) are much larger than the remaining n-ρ ones.* This decision is somewhat subjective; the main aim is to pick ρ as large as possible, while maintaining the maximal ratio of singular values (of the reduced-order problem) sufficiently small. For the model analyzed here, ρ was selected to meet two criteria: (1) the maximal ratio of singular values is $\leq 10^6$, and (2) the smallest of the ρ singular values is more than twice as large as the largest of the n-ρ singular values; the latter criterion prevented artificial splitting — that was occasionally imposed by the former criterion - of similar singular values.
- *Make the partition $V=[V_\rho\ V_{n-\rho}]$, with V_ρ denoting the first ρ columns of V.*
- *Determine a permutation matrix P by constructing a QR decomposition with column pivoting, i.e., determine $V_\rho{}'P = Q\ R$, where Q is an orthogonal matrix, and the first ρ columns of R form an upper triangular matrix.*
- *Use the matrix P to reorder the parameter vector $\tilde{\xi} = P\xi$; the first ρ parameters in $\tilde{\xi}$ should be estimated, while the remaining n-ρ should be fixed to prior estimates.*
- *Solve the reduced nonlinear least-squares problem involving ρ parameters.*

3.3 Covariance Method

In addition to using the subset-selection algorithm to analyze the rate–level curve model, we also used the more standard “covariance method.” In this method standard deviations of parameters were estimated through calculation of the diagonal entries of the covariance matrix $C=(J'J)^{-1}$ ([6]). This estimation procedure assumes that the actual optimization problem is locally well approximated by its linearized version. The standard deviations were used to construct an n-dimensional polyhedron in parameter space centered around the point in parameter space representing the optimized solution, with each side equal to twice the parameter standard deviation in the corresponding direction. For each of the vertices of the polyhedron (which are extremal points in the parameter space) we calculated the corresponding rate–level curve. The envelopes of a family of rate–level curves thus generated represent the uncertainty in the fit, which we refer to as the “separation of the envelopes of fit.”

4. Results

The results are described in terms of the interplay of three quantities, which are different (and complementary) measures of the goodness of the fit:

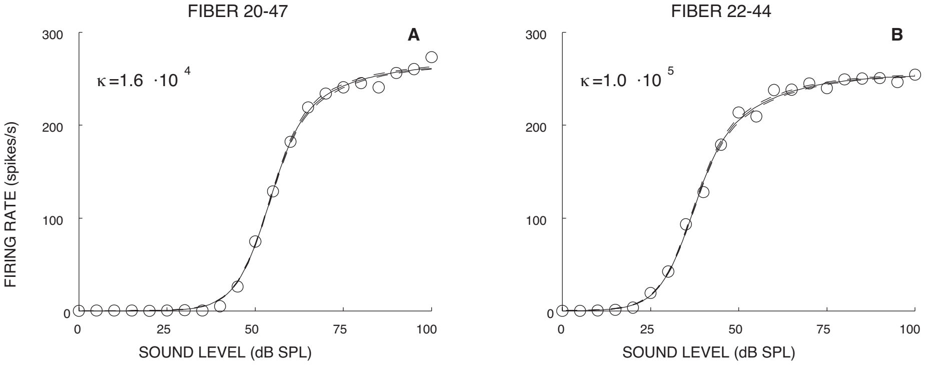


Figure 1 Four-parameter fit [R_M, R_{SP}, θ_E, θ_I] to rate–level curves with sloping saturation from two different auditory-nerve fibers *(α, β, γ* are fixed to $\alpha = 1/3$, $\beta = 2$, $\gamma = 1.77$). Circles: original data. Solid line: fit determined by a nonlinear least-squares optimization procedure. Dashed lines (almost overlapping the solid line): uncertainty in the fit, as determined from the covariance method (details provided in the text). κ: maximal ratio of singular values. **A:** Fiber 20-47. Stimulus: tone at CF=3.55 kHz. Estimated parameters: R_M = 265.9±0.8 spikes/s, R_{SP} = 0.0+0.3 spikes/s, θ_E = 54.78±0.07 dB SPL, θ_I = 62.0±0.3 dB SPL. Error-criterion value = 467.4. Singular values: [$1.98*10^4$, $1.66*10^3$, 3.44, 1.23]. **B:** Fiber 22-44. Stimulus: tone at CF = 12.6 kHz. Estimated parameters: R_M = 254.6±0.7 spikes/s, R_{SP} = 0+0.4 spikes/s, θ_E = 37.99±0.08 dB SPL, θ_I = 43.4±0.2 dB SPL. Error-criterion value = 544.1. Singular values: [$1.38*10^5$, $1.52*10^4$, 3.66, 1.36].

(1) The *maximal ratio of singular values*, κ, which characterizes geometry of the Hessian matrix at optimized parameters, and is a critical feature of the subset-selection method. In general, very large maximal ratios of singular values are associated with the presence of severely ill-conditioned parameters. In contrast, small maximal ratios of singular values imply that all parameters can be estimated with similar precision because curvatures of the error-criterion function are comparable in all directions in the parameter space;

(2) The *separation of the envelopes of the fit*, which is assessed through the covariance method (as described in the Methods section) and represented by dashed lines in Figure 1;

(3) The *error-criterion value*, which characterizes how close the measured values are to predictions generated by the model, where model parameters are determined in the optimization procedure. Mathematically, the error-criterion value is the sum of squared residual errors, as defined in the Background section.

All parameter values presented in this section are results of a nonlinear least-squares optimization procedure, and thus may correspond to local minima. The optimization procedure was restarted with different starting points whenever existence of another local minimum was suspected based on (1) a poor fit (characterized by a large error-criterion value, i.e., the sum of squares of errors), or (2) unusual parameter estimates. Therefore, it is hoped that our results are close to the global minima.

4.1 Four-Parameter Auditory-Nerve Rate–Level Curve Model

Analysis of the original four-parameter rate–level curve model [7][8] by the subset-selection method showed that — for rate–level curves with sloping saturation (total of 37) — all four parameters of the model [R_M, R_{SP}, θ_E, θ_I] were well conditioned, and therefore could

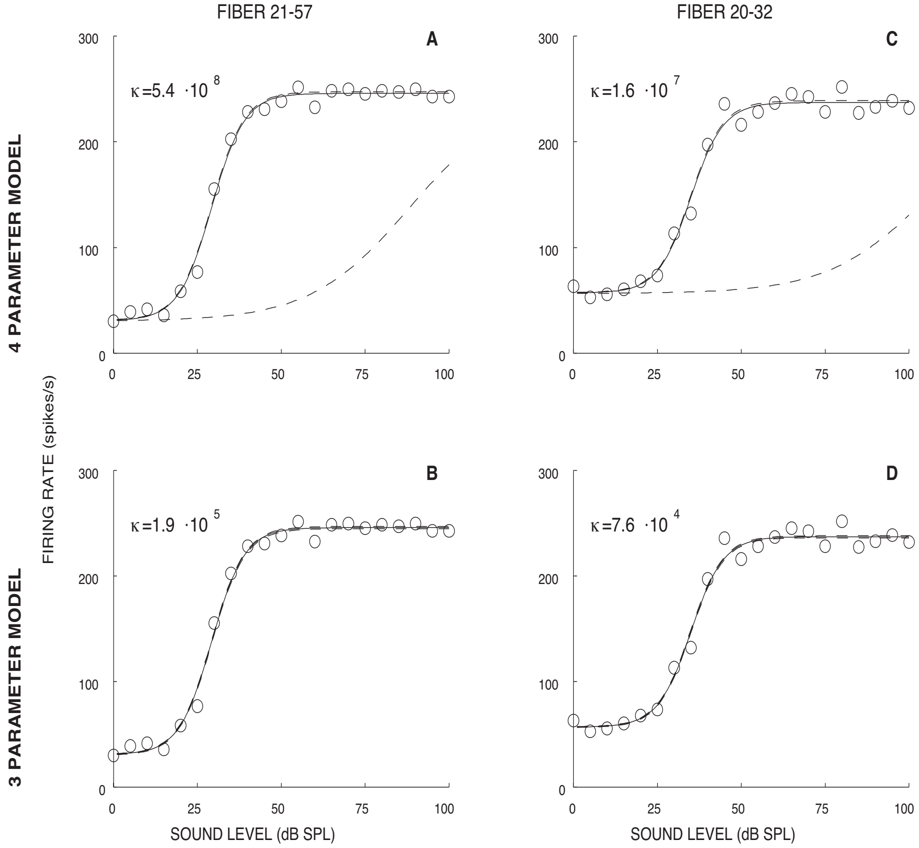


Figure 2 Four-parameter fits [R_M, R_{SP}, θ_E, θ_I] (A, C) and 3-parameter fits [R_M, R_{SP}, θ_E] (B, D) to rate–level curves with flat saturation from two different auditory-nerve fibers. Symbols as in Figure 1. **A, B:** Fiber 21-57. Stimulus: tone at CF = 5.9 kHz. **A:** Estimated parameters: R_M = 215.2±1.0 spikes/s, R_{SP} = 30.7±0.5 spikes/s, θ_E = 29.28±0.08 dB SPL, θ_I=92.2+67.7-92.2 dB SPL. Error-criterion value = 930.2. Singular values: [2.64*10^5, 5.05, 1.37, 4.89*10^{-4}]. **B:** Estimated parameters: R_M = 215.2±0.6 spikes/s, R_{SP} = 30.7±0.5 spikes/s, θ_E = 29.28±0.07 dB SPL, θ_I fixed to 92 dB SPL. Error-criterion value = 930.2. **C, D:** Fiber 20-32. Stimulus: tone at CF = 25.7 kHz. **C:** Estimated parameters: R_M = 180.3±1.3 spikes/s, R_{SP} = 56.9±0.4 spikes/s, θ_E=34.9±0.1 dB SPL, θ_I=86.2+50.8-86.2 dB SPL. Error-criterion value = 1566.4. Singular values: [1.16*10^5, 4.88, 1.52, 7.07*10^{-3}]. **D:** Estimated parameters: R_M = 180.3±0.5 spikes/s, R_{SP} = 56.9±0.4 spikes/s, θ_E = 34.93±0.08 dB SPL, θ_I fixed to 86 dB SPL. Error-criterion value = 1566.4.

be reliably estimated from experimental data. Examples from two fibers are shown in Figure 1. For both fibers, the subset selection method identified a relatively small spread in singular values (e.g., singular values for the fiber in Figure 1A were [1.98*10^4, 1.66*10^3, 3.44, 1.23]. These results were concordant with the outcome of the covariance method, which found the separation of the envelopes of the fit to be very small (dashed lines in Figure 1 almost overlap the solid line). Of note is that our data set did not have clear examples of straight rate level curves; the curves that may have appeared to be straight — based on a quick visual inspection — were found to have steeply sloping saturations.

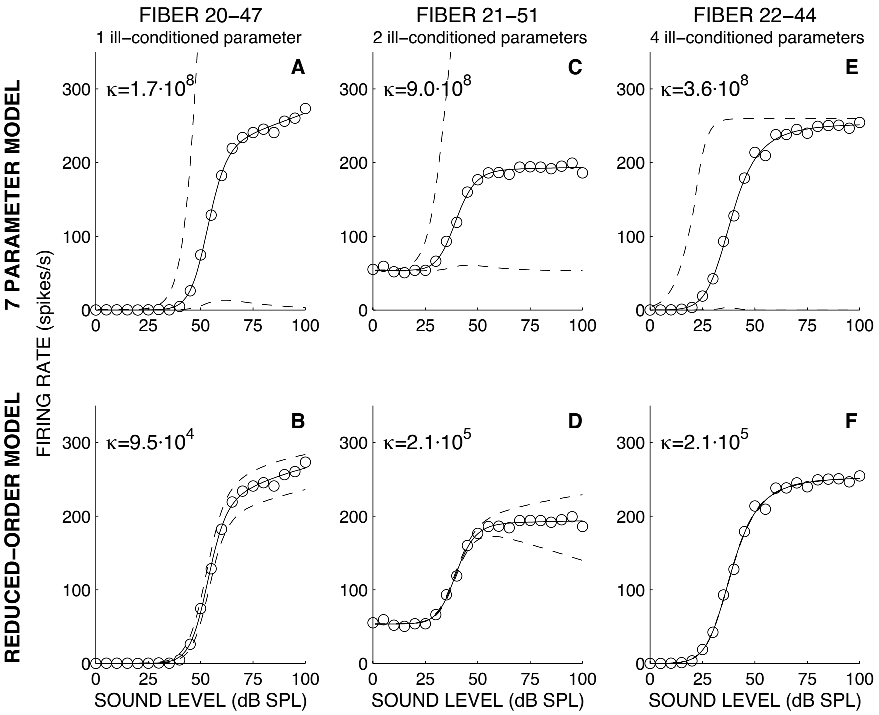


Figure 3 Seven-parameter fit [R_M, R_{SP}, θ_E, $\theta_{I,,}$, α, β, γ]. **(A, C, E)** and reduced-order fit **(B, D, F)** to rate–level curves from three different auditory-nerve fibers. Figure layout as in Figure 2. **A, B**: Fiber 20-47 (same as in Figure 1-A). **A**: Estimated parameters: R_M =875.2$\pm$657.3 spikes/s, R_{SP}=0+0.4 spikes/s, θ_E=57.5+3.2-5.1 dB SPL, θ_I =53.3+0.39-0.40 dB SPL, α=0.53$\pm$0.05, β=1.83$\pm$0.14, γ=2.28$\pm$0.16. Error-criterion value=245.0. Singular values: [2.62*10^5, 2.43*10^4, 458.0, 32.7, 18.0, 2.76, 1.52*10^{-3}]. Ill-conditioned parameter: R_M. **B**: Estimated parameters: R_{SP} =0+0.4 spikes/s, θ_E =54.3$\pm$0.9-1.0 dB SPL, θ_I =58.5$\pm$0.2 dB SPL, α=0.49$\pm$0.01, β=1.84$\pm$0.04, γ=2.09$\pm$0.09. R_M fixed to 300 spikes/s. Error-criterion value=292.3. **C, D**: Fiber 21-51. Stimulus: tone at CF=1.26 kHz. **C**: Estimated parameters R_M=429.9$\pm$297.8 spikes/s, R_{SP}=53.6 $\pm$ 0.5 spikes/s, θ_E=41.5+3.3-5.3 dB SPL, θ_I =38.76$\pm$0.02 dB SPL, α=0.60$\pm$0.07, β=1.66$\pm$0.18, γ=2.45$\pm$0.28. Error-criterion value=280.2. Singular values: [3.02*10^6, 1.87*10^5, 727.0, 29.4, 14.3, 2.27, 3.36*10^{-3}]. Ill-conditioned parameters: R_M, R_{SP}. **D:** Estimated parameters: θ_E =39.8$\pm$0.1 dB SPL, θ_I =42.71$\pm$0.01 dB SPL, α=0.61$\pm$0.02, β=1.62$\pm$0.06, γ=2.27$\pm$0.05. R_M fixed to 200, and R_{SP} fixed to 53.6. Error-criterion value=281.5. **E, F**: Fiber 22-44 (same as in Figure 1-B). **E:** Estimated parameters: R_M=252.5$\pm$6.7 spikes/s, R_{SP}=0$\pm$0.5spikes/s, θ_E=31.9+4.8-11.4 dB SPL, θ_I=36.5$\pm$6.7*10^{-4} dB SPL, α =1.00$\pm$1.03, β=0.76$\pm$0.47, γ=2.31$\pm$0.53. Error-criterion value=463.3. Singular values: [5.37*10^7, 3.53*10^5, 259.0, 24.5, 3.29, 1.93, 0.15]. Ill-conditioned parameters: [R_M, R_{SP}, α, γ]. **F:** Estimated parameters: θ_E=31.9$\pm$0.04dB SPL, θ_I =36.5$\pm$1.3*10^{-4} dB SPL, β=0.76$\pm$0.005. R_M, R_{SP}, α, γ set to the estimates from E. Error-criterion value=463.3.

In contrast to the rate–level curves with sloping saturation, all 4 parameters were not well conditioned for rate–level curves with flat saturation (total of 16); θ_I was the ill-conditioned parameter, and therefore had to be fixed to a prior estimate to improve model performance. This is illustrated in Figure 2 for two auditory-nerve fibers. For both fibers the smallest singular values encountered in optimizations were substantially (i.e., 3-4 orders of magnitude) smaller than the first proximal singular value. Consequently, the maximal ratio of singular values was very large (Figure 2-A, C). The subset selection method identified θ_I as the ill-conditioned parameter associated with the smallest singular values. By fixing θ_I to a sound level in the saturating region of rate–level curves, the 3-parameter model so derived demonstrated a substantial reduction in the maximal ratio of singular values (Figure 2-B, D). Results of the covariance method were similar: θ_I was identified as the parameter with the largest standard deviation (see caption of Figure 2) and, by fixing θ_I, the separation of the envelopes of the fit significantly decreased (e.g., compare dashed lines in Figure 2A and 2B).

4.2 Seven-parameter Auditory-Nerve Rate–Level Curve Model

It is interesting to consider what happens to the performance of the rate–level curve model when it is expressed in its most general form with 7 parameters (model (1)). As expected, by increasing the dimension of the search space the closeness of the fit improved, as evidenced by a decrease in the error-criterion value. For example, for fiber 20-47, the error-criterion value of 467.4 in the 4-parameter model (Figure1A) decreased to 245.0 in the seven-parameter model (Figure 3A). Similarly, for fiber 22-44, the error-criterion value decreased from 544.1 (Figure 1B) to 463.3 (Figure 3E).

However, this improvement in the error-criterion value is deceptive because it is accompanied by a substantial increase in the spread of singular values of the Jacobian, leading to a substantial increase (by several orders of magnitude) in the maximal ratio of singular values, κ. For example, for fiber 20-47, κ increased from $1.6*10^4$ in the 4-parameter model (Figure 1A) to $1.7*10^8$ in the 7-parameter model (Figure 3A). Similarly, for fiber 22-44, κ increased from $1.0*10^5$ (Figure 1B) to $3.6*10^8$ (Figure 3E). This increase in the maximal ratio of singular value implied the presence of ill-conditioned parameters.

For the seven-parameter model, the number of ill-conditioned parameters varied from zero to four, depending on a data set. Figure 3 illustrates this point on examples from three fibers. For fiber 20-47 (Figure 3A,B), R_M was identified as the ill-conditioned parameter, with the unreliable (and nonphysiological) estimate of 875 spikes/s. For that fiber, the corresponding singular values were $[2.62*10^5, 2.43*10^4, 458.0, 32.7, 18.0, 2.76, 1.52*10^{-3}]$. After R_M was fixed to a physiologically plausible value of 300 spikes/s (given the overall trend in the data), model performance improved as evidenced by a substantial decrease in the maximal ratio of singular values. For this fiber, the covariance method gave similar results: R_M was associated with the largest parameter uncertainty (875.2 ± 657.3). In addition, by fixing R_M, the separation of the envelopes of the fit substantially decreased, and standard deviations of other parameters also decreased.

Although for most data the subset-selection method allowed easy estimation of ill-conditioned parameters — based on a large gap between the smallest singular value and the next closest singular value (e.g., fiber 20-47 in Figure 3-A, B, also fibers 21-57 and 20-32 in Figure 2) — for other data this estimation was less obvious. For example, for fiber 22-44 (Figure 3E, F) the gap between the smallest singular values was relatively small, so that the ill-conditioned parameters were identified based on the *overall* large spread ($>10^6$) in singular values. Consequently, based on our criteria for the well-conditioned parameters (see section 3.2), the subset-selection method identified 4 ill-conditioned parameters $[R_M, R_{SP}, \alpha, \gamma]$. For comparison, the covariance method found that only two of those parameters (α, γ) were associated with relatively large standard deviations.

Another example that illustrates a difficulty when imposing rigid (and somewhat arbitrary) criteria to identify well-conditioned parameters using the subset-selection method is fiber 21-51 (Figure 3C, D). As a consequence of using the criterion that the maximal ratio of singular values be $\leq 10^6$, two ill-conditioned parameters, R_M and R_{SP} were identified. However, based on the singular values of $[3.02*10^6, 1.87*10^5, 727.0, 29.4, 14.3, 2.27, 3.36*10^{-3}]$, it is clear that the gap between the sixth and seventh singular values is much larger than the gap between the fifth and sixth singular values. Therefore, if a different criterion for identification of well-conditioned parameters had been used (e.g., that the maximal ratio of singular values be $\leq 1.4*10^6$), only one ill-conditioned parameter, R_M, would have been identified. For comparison, the covariance method found that only one parameter, R_M, had very large standard deviation.

Interestingly, the ill-conditioned parameters of the 7-parameter model were not limited to only α, β and γ, i.e., to the parameters that were fixed in the 4-parameter model. On the contrary, the most common ill-conditioned parameter of the 7-parameter model was R_M (e.g., Figure 3A, C, E), which was always well-conditioned in the in the 4-parameter model.

5. Discussion

This chapter illustrates the usefulness of the subset selection method for model-parameter identification in nonlinearly parametrized problems that use the minimum least-squares criterion [11][1]. The method has been used before in electrical engineering (for applications in electrical machines and in remote sensing), but here we illustrate that it is also applicable in computational neuroscience.

For the original computational model of auditory-nerve [7][8], there is a systematic difference between the number of ill-conditioned parameters depending on the type of saturation — curves with sloping saturation have no, and those with flat saturation have one ill-conditioned parameter (θ_I). This result is not surprising, since θ_I is a measure of the onset of sloping saturation, which is ill-defined for rate–level curves with flat saturation. Specifically, θ_I describes a "compressive threshold" for pressure mapping that allows transformation of a regular sigmoid with flat saturation into a sigmoid with sloping saturation; for rate–level curves with flat saturation θ_I could be at any sound level in the saturating region. It is reassuring that Sachs et al.[8] — using iterative procedures to explore model sensitivities to one parameter at a time — arrived at the same conclusion as we did that the largest parameter uncertainty is associated with θ_I for rate–level curves with flat saturation.

We further demonstrated that the generalized 7-parameter model is typically overparametrized. For a vast majority of fibers, it was not possible to reliably estimate all 7 parameters from the data. By identifying ill-conditioned parameters, and fixing them to physiologically-reasonable estimates, it was possible to significantly improve estimation of well-conditioned parameters. However, it is worth noting that the number of ill-conditioned parameters is a direct consequence of the numerical value adopted as a criterion for the acceptable spread in singular values. Such a value is needed so that all data are treated in a uniform fashion. However, it may not be straightforward to select this number so that results agree with intuition for all cases (e.g., Figure 3C).

The fact that the number of ill-conditioned parameters varied from 0 to 4, and that any parameter (except θ_E) could be ill-conditioned, suggests that a single model is not applicable to all data. Instead, it may be advisable to consider a whole *family* of models (with up to 7 parameters), and to use the method presented here to identify the most suitable model for a given data set. An important member of the family is certainly the standard 4-parameter model that performs well for sloping saturations, and needs to have one parameter (θ_I) restricted for flat saturations. The main advantages of the proposed method are that it can detect overparametrization in a systematic way, and that it can handle more complicated cases when *sets* of parameters need to be fixed to prior estimates, instead of being determined through optimization.

We have shown that results of the subset selection method are typically consistent with the covariance method; ill-conditioned parameters (determined by the subset selection method) tend to be associated with large standard deviations of parameters (determined by the covariance method). It is certainly advisable to use either of the two methods to assess quality of parameter estimates obtained through least-squares optimization procedures. The goal in both methods is to determine the subset of well-conditioned parameters that can be reliably estimated from the data. They also identify ill-conditioned parameters whose parameter estimates are unreliable, and whose presence interferes with estimation of other parameters (that may be well-conditioned in a reduced order model). However, the subset selection method is numerically more robust than the covariance method since the covariance method requires inversion of a matrix that may be close to singular.

A limitation of the current form of the subset selection method is that, for each direction in the parameter space along which the error-criterion function does not change (i.e., a "bad" direction), a *single* parameter that is most aligned with that direction is identified as ill-conditioned. An improvement entails identifying and constraining *combinations* of parameters that are most aligned with "bad" directions; this would amount to reparameterizing the model.

6. Conclusions

The subset-selection method is a powerful technique that allows improved parameter estimation in multiparameter nonlinear models. It is worth emphasizing that the sole attention to the error-criterion value may be misleading when evaluating model performance — a decrease in the error-criterion value may be associated with a substantial increase of the maximal ratio of singular values, which renders some parameter estimates unreliable. When compared with the common practice of performing sensitivity analyses using iterative simulations to determine parameter uncertainty, the subset selection method offers a distinct advantage — it determines the subset of parameters that can be reliably estimated from the data, thus improving numerical conditioning of the optimization problem. Although we presented here only one application of the subset selection method in computational neuroscience, it is certainly relevant whenever models with nonlinear parametrizations and a least-squares-error criterion are used.

Acknowledgments

Insightful comments of Drs. John J. Guinan, Jr. and Christopher A. Shera are greatly appreciated. This work was supported (in part) by training grant T32 DC0038, and grant DC00235 (awarded to JJG) from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, and the MSTP grant T32 GM07753 from the National Institutes of Health.

References

[1] Burth, M., Verghese, G. C. and Velez-Reyes, M. "Subset selection for improved parameter estimation in on-line identification of a synchronous generator." *IEEE Trans. Power Systems,* 14(1): 218–225, 1999.

[2] Delgutte, B. "Physiological models for basic auditory percepts." In *Auditory Computation*, H. L. Hawkins, H. L., T. A. McMullen, A. N. Popper and R. R. Fay (eds.), New York: Springer-Verlag, 1996.

[3] Golub, G. H., and VanLoan, C. F. *Matrix Computations* (2nd ed.). Baltimore: Johns Hopkins University Press, 1989.

[4] Guinan, J. J. Jr. and Stankovic, K. M. “Medial efferent inhibition produces the largest equivalent attenuations at moderate to high sound levels in cat auditory-nerve fibers.” *J. Acoust. Soc. Am.*, 100: 1680–1690, 1996.

[5] Liberman, M. C. and Kiang, N. Y. S. “Single-neuron labeling and chronic cochlear pathology, IV: Stereocilia damage and alterations in rate- and phase-level functions.” *Hear. Res.* 16: 75–90, 1984.

[6] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in C (2nd ed.)*. New York: Cambridge University Press, 1992.

[7] Sachs, M. B. and Abbas, P. J. “Rate versus level functions for auditory-nerve fibers in cats: Tone-burst stimuli.” *J. Acoust. Soc. Am* 56: 1835–1847, 1974.

[8] Sachs, M. B., Winslow, R. L. and Sokolowski, B. H. A. “A computational model for rate–level functions from cat auditory-nerve fibers.” *Hear. Res.*, 41: 61–70, 1989.

[9] Yates, G. K., Winter, I. M., and Robertson, D. “Basilar membrane determines auditory nerve rate–intensity functions and cochlear dynamic range.” *Hear. Res.*, 45: 203–220, 1990.

[10] Yates, G. K. “Basilar membrane nonlinearity and its influence on auditory-nerve rate–level functions.” *Hear. Res.*, 50: 145–162, 1990.

[11] Velez-Reyes, M. *Decomposed Algorithms for Parameter Estimation*. Ph.D. Thesis, Massachusetts Institute of Technology, 1992.

[12] Winter, I. M., Robertson, D. and Yates, G. K., “Diversity of characteristic frequency rate-intensity functions in guinea pig auditory-nerve fibers.” *Hear. Res*, 45: 191–202, 1990.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

CHARACTERISTICS OF COCHLEAR NUCLEUS ONSET UNITS STUDIED WITH A MODEL

Sridhar Kalluri[1, 2] and Bertrand Delgutte[1, 2, 3]

[1] *Harvard-MIT Division of Health Sciences and Technology*
243 Charles Street, Massachusetts Eye and Ear Infirmary
Boston, MA 02114, USA

[2] *Eaton-Peabody Laboratory*
Massachusetts Eye and Ear Infirmary

[3] *Research Laboratory of Electronics*
Massachusetts Institute of Technology

1. Introduction

Onset neurons are characterized by their preferential response to onset transients in acoustic signals. These neurons have long been of interest to auditory scientists because of the importance of onset transients for the perception of speech and music, as well as for sound localization and stream segregation [6][43][48].

Onset units are found throughout the central auditory system, beginning with the ventral cochlear nucleus (VCN) [17]. VCN onset units are particularly interesting from a modeling perspective because their response properties differ sharply from those of their auditory-nerve (AN) inputs. Intracellular labeling of physiologically characterized cells shows that VCN onset units form a heterogeneous group in that they are morphologically associated with several anatomical classes, including stellate cells, bushy cells and octopus cells [31] [34][40][41]. Given this heterogeneity, it is not surprising that current knowledge of the neuronal characteristics associated with onset unit responses is still very much incomplete. In particular, there is no universally accepted scheme for classifying onset units into subtypes based on their response properties nor for associating physiological subtypes with anatomical cell types [5][13][35][36]. A long-term goal of our research is to determine the neuronal properties underlying the responses of different subtypes of onset units to acoustic stimulation and thereby clarify the correspondence between cell types and physiological subtypes of onset units. As a first step towards that goal, this chapter describes a mathematical model used to identify a minimum set of neuronal characteristics required to obtain key response properties common to all VCN onset units.

Two types of models have been used for investigating the underlying mechanisms of onset unit response patterns. One approach has been to construct a detailed biophysical model, including active membrane channels [38], and electro-anatomical characteristics of the cell body and dendritic tree [7][25]. Such models point to the importance of fast membrane dynamics and weakly excitatory synapses requiring coincidence of many inputs to obtain onset response patterns to tonal signals. However, at present there is insufficient information pertaining to ion channels and synaptic distributions in most VCN neurons to adequately constrain these biophysical models. A further difficulty is that different parameters would be required for each of the different cell types that give rise to onset response properties. Because of the drawbacks associated with detailed biophysical models, we have developed a very general phenomenological model that contains the essential elements without attempting to model detailed biophysical properties specific to a given neuronal class.

Previous phenomenological models of onset units [2][9] have suggested the need for a high-pass filtering mechanism such as depolarization block, threshold accommodation, or receptor desensitization that would decrease the probability of discharge during sustained depolarization. However, these models only examined in detail the neuronal responses to a limited set of stimuli. We extend the results of these earlier models by investigating a relatively large set of stimuli and neuronal response characteristics. A major result of our study is that three separate response properties are identified that strongly constrain the model. If these properties are correctly predicted then the model is successful in predicting responses to a broader range of stimuli. This chapter focuses on these three crucial, physiological properties:

(1) The onset peri-stimulus time (PST) histogram for high-frequency tone bursts consisting of a prominent peak followed by little or no response during the on-going part of the stimulus [32]. This property contrasts with the sustained response patterns of auditory-nerve inputs.
(2) Entrainment of spike discharges to low-frequency (< 1 kHz) tones [35] (i.e., the occurrence of one spike on every cycle of the stimulus). AN-fibers rarely entrain, in that multiple tone cycles typically occur between successive spikes.
(3) Similar thresholds for broadband noise and characteristic frequency (CF) tones [46]. Again, this property contrasts with that of AN fibers, where CF tone thresholds are always significantly lower than noise thresholds.

Our model of a VCN onset cell is based on an integrate-to-threshold point neuron whose inputs are AN fibers acting via excitatory synapses. We use a two-part strategy for identifying model characteristics necessary for obtaining realistic onset response properties. First, the dynamic properties of the cell membrane are fit to intracellular measurements of voltage responses to current injections in octopus cells (which are the cells most convincingly associated with VCN onset-responding neurons [31][34][40]). Second, synaptic and input properties of the model (specifically synaptic strength, number of independent AN inputs, and CF distribution of these inputs) are constrained by the three response properties enumerated above.

2. Methods

2.1 Model of Auditory-Nerve Fibers

AN-fiber responses were computed using a model that simulates the primary features of temporal discharge patterns for tones and noise [8]. The model includes the following features.

(1) The bandpass tuning of an AN fiber is modeled as a linear gammatone filter [18].
(2) A second-order, low-pass filter with a 1.1 kHz cutoff frequency models the reduction of synchronization to high-frequency tones.
(3) Adaptation is described using a model of the inner-hair-cell synapse [45].
(4) To describe the statistical properties of discharge patterns, the spike train is modeled as a non-stationary renewal process whose instantaneous probability of discharge is the product of a component representing excitatory drive from the hair cells and a component

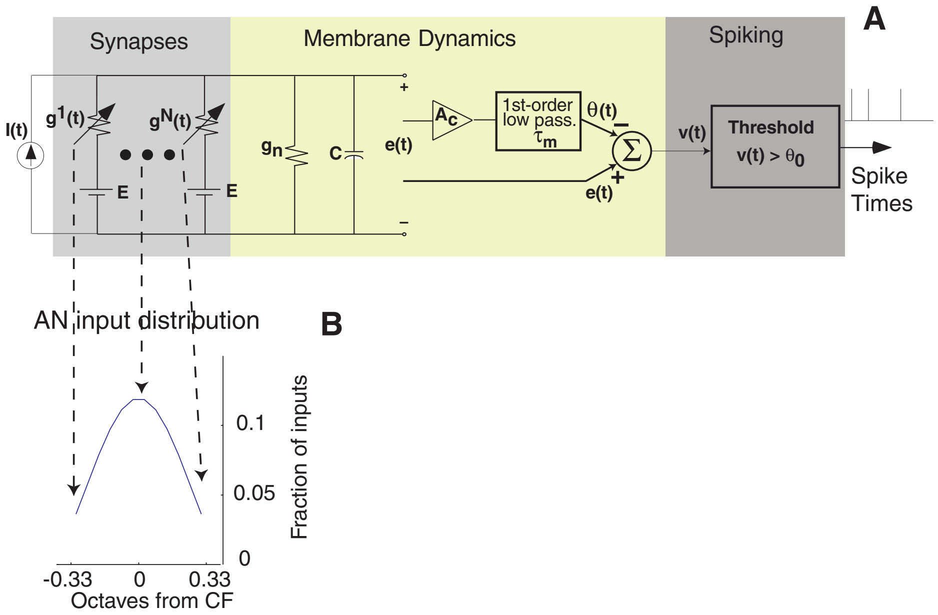


Figure 1 The model for an onset responding neuron. A. Schematic of the integrate-to-threshold point-neuron model. B. The distribution of characteristic frequencies (CFs) of auditory nerve (AN) inputs to the model neuron. In this case the total CF range of the AN inputs is 2/3 of an octave.

representing refractory properties of the fiber [19]. Spikes for each AN fiber are produced using a set of independent random number generators.

2.2 *Point-Neuron Model of VCN Onset Cells*

Figure 1A illustrates a schematic representation of the model onset neuron. The model is deterministic. We divide it into two components, input and membrane dynamics. All of the model parameters are listed in Table 1.

2.2.1 Input

The model contains only excitatory synapses that are driven by spikes from AN fibers. A spike causes a smooth transient increase in conductance of the corresponding synapse. The duration of the conductance change is 500 microseconds for all synapses. Synaptic strength (or magnitude of conductance change) is the same for all synapses. For illustrative purposes synaptic strength is normalized to the unitary synaptic strength (defined as the threshold strength of a synapse such that an isolated input spike gives rise to an output spike).

The number of independent AN inputs to the model neuron is also a parameter. We examine values of this parameter between 1 and 128. The distribution of CFs of the AN inputs is described by a Gaussian-like density function (Figure 1B). The function is symmetric about the CF of the model onset unit on a log-frequency axis. In all figures, the CF of the model onset unit is 6 kHz. The total frequency range spanned by the inputs is a parameter of the model.

Table 1. Summary of model parameters. Asterisk denotes that the parameter is varied in some illustrations.

Membrane parameters	
Membrane time constant: $\tau_m = C/g_n$	0.39 ms
Accommodation time constant: τ_θ	0.67 ms
Accommodation gain: A_c	0.49
Absolute refractory period	0.75 ms
Synaptic and input parameters	
Normalized synaptic strength	0.16*
Number of AN inputs	100*
CF range of AN inputs	2/3 octave*
Duration of synaptic conductance change	0.5 ms

2.2.2 Membrane Dynamics and Refractoriness

The dynamic properties of the membrane are based on the two-factor model for membrane electrical excitability first investigated by Hill and others [16][29][33]. In the model (Figure 1A), the membrane voltage v(t) is the difference between an integrative factor e(t) and an accommodation factor, $\theta(t)$. The integrative factor results from low-pass filtering (temporally integrating) the summed synaptic inputs. The time constant of the integrative factor, τ_m, is very short (< 1 ms), and is determined by the membrane capacitance. The accommodation factor is itself a low-pass filtered version of the integrative factor. Since $\theta(t)$ is subtracted from e(t), it effectively acts as a high-pass filter that emphasizes transients in the synaptic inputs. The time constant, τ_θ, of the accommodation factor is longer than that of the integrative factor, but is still very short. The accommodation factor results in a decrease in membrane voltage during sustained depolarization. An accommodation gain, A_c, controls the amount of accommodation relative to the integrative factor.

The neuron produces a spike discharge whenever the membrane voltage exceeds a fixed threshold, θ_0. For a fixed refractory period following a spike, the neuron cannot fire, and both the integrative process and the accommodation process are undefined. Thus, no attempt is made to model the spike waveform. At the end of the refractory period, the integrative process is reset to zero, while the accommodation process resumes the value it had prior to the spike.

3. Results

3.1 Membrane Electrical Properties

Our model for membrane electrical excitability is based on properties of octopus cells derived from *in vitro* experiments. We chose octopus cells because they are the most convincingly identified onset responders in the VCN [13][31][34][40]. Parameters of the membrane model were estimated using voltage responses of octopus cells to current injections.[1]

Only responses of the octopus cell to depolarizing current injections were used for estimating parameters. Figure 2 shows responses of an octopus cell to step current injections of

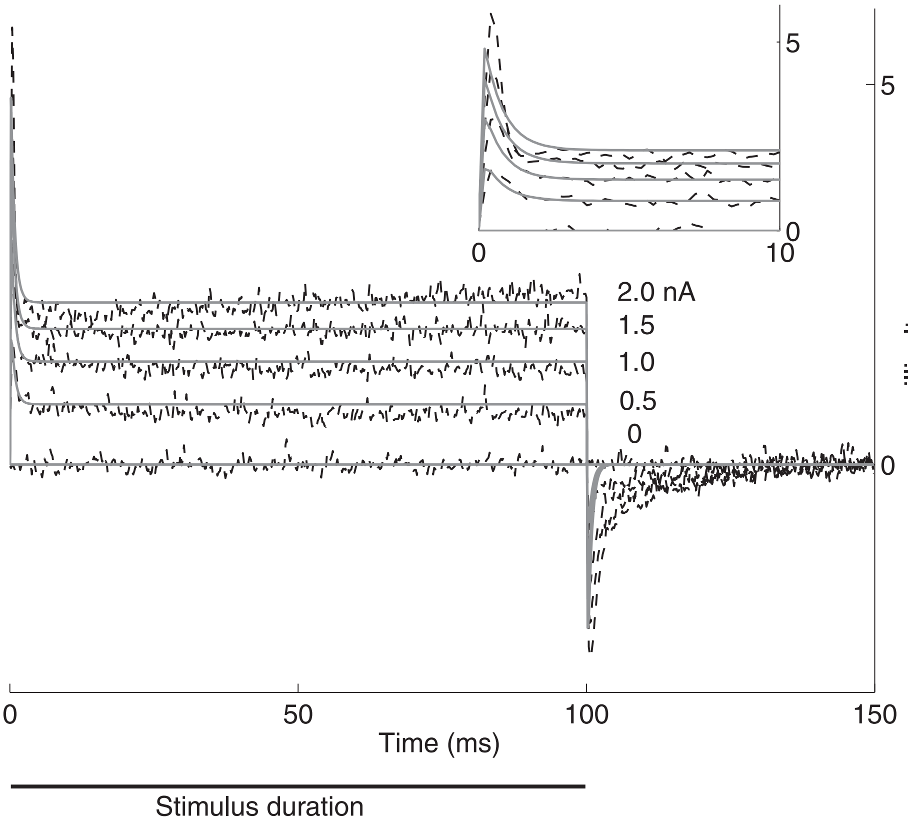


Figure 2 Model and octopus cell responses to step current injections. Solid curves represent the model and dashed curves indicate the octopus cell responses [14]. Inset shows the first 10 ms of the response.

different amplitudes and model responses that best fit the data according to a least-squares-error criterion [14].

The sub-threshold octopus cell response exhibits a rapid rise and a slower decline to a constant steady-state level in the first 10 ms after the onset of the stimulus. A model with a single dynamic process (i.e., with a single time constant) can fit either the rise or the fall of the octopus cell response, but not both. At least two dynamic processes (i.e., with two time constants) are required to concurrently capture both properties of the response. Figure 2 shows that our model successfully captures the main features of the octopus cell response; the best-fit parameters are $\tau_m = 0.25$ ms, $\tau_\theta = 0.64$ ms, and $A_c = 0.62$. The two dynamic processes, membrane integration and accommodation, confer a very brief maximum in membrane voltage for positive current injections that is reflected in the single spike elicited at onset by supra-threshold current steps in the model and in octopus cells (not shown). The model predicts spikes at the offset of large negative current injections; such spikes are also observed in octopus cells.

There are some systematic deviations between model and data evident in the responses to large current injections. In particular, the model underestimates the height of the initial peak in the response and the rapidity of the early decline of the response (inset of Figure 2). At least one additional time constant in the accommodation factor is required to accurately describe these properties of the octopus cell response.

We estimated model parameters using responses of five different octopus cells to current injections. The residual variance of the best-fit model was always less than 7% of the total variance of the data. The model parameters were similar across these five cells. In the following sections are described the model's response to acoustic signals using the median parameter values for the membrane model listed in Table 1.

3.2 Synaptic Strength and Number of Independent AN Inputs

We examined, in the model, the effect of synaptic strength and the number of independent AN inputs on temporal discharge patterns in response to tone-burst stimulation. More specifically, we looked at how the gross shape of the PST histogram associated with CF-tone responses and the synchronization and entrainment to a low-frequency tone (600 Hz) vary with these model parameters.

3.2.1 PST Histograms for Tones

A defining property of VCN onset units is the shape of their PST histograms of high-frequency tone burst responses. These histograms have a prominent peak at stimulus onset, followed by a low discharge rate during the steady-state portion of the stimulus over a wide range of stimulus levels [4][5][13][35][46]. Figure 3 illustrates how PST histograms of CF tone burst responses at 20 dB above threshold depend on synaptic strength and on the number of independent AN inputs to a model neuron.

As a reference, a PST histogram is shown for a 6-kHz (CF) tone burst response of an AN input (Figure 3D), which has a sustained response with fast adaptation. The model response has different PST histogram shapes depending on the number of inputs and synaptic strength (Figure 3A, 3B, 3E, 3F). When the model exhibits weak synapses and many (100) independent AN inputs, its PST histogram has an onset shape with no sustained activity (Figure 3A). Increasing the synaptic strength moderately (Figure 3E) produces a PST histogram of the model response that is more sustained, similar to the PST histogram of the AN input. When the strength of synapses is moderate, but the number of AN inputs is reduced, the PST histogram of the response is still of the onset variety, but with more sustained activity (Figure 3B). Figure 3C shows how the shape of PST histograms varies as a function of synaptic strength and the number of independent AN inputs. We identify three types of model responses (criteria shown in the caption of Figure 3):
(1) a region where there is little or no response,
(2) a region where the PST histogram has an onset shape, and
(3) a region where the PST histogram has a sustained shape.

Although the shape of the PST histogram depends on both the number of independent AN inputs and on synaptic strength, the latter is more important. In order to obtain the onset form of PST histogram associated with high-frequency tone bursts, the synaptic strength must be weak.

Spontaneous rate depends in a similar way on synaptic strength and the number of independent AN inputs (with synaptic strength being more important). The model has a spontaneous rate magnitude (< 2 spikes/s) appropriate for an onset unit when the synaptic strength is weak. These observations are similar to those associated with previous models of onset neurons [23] [38].

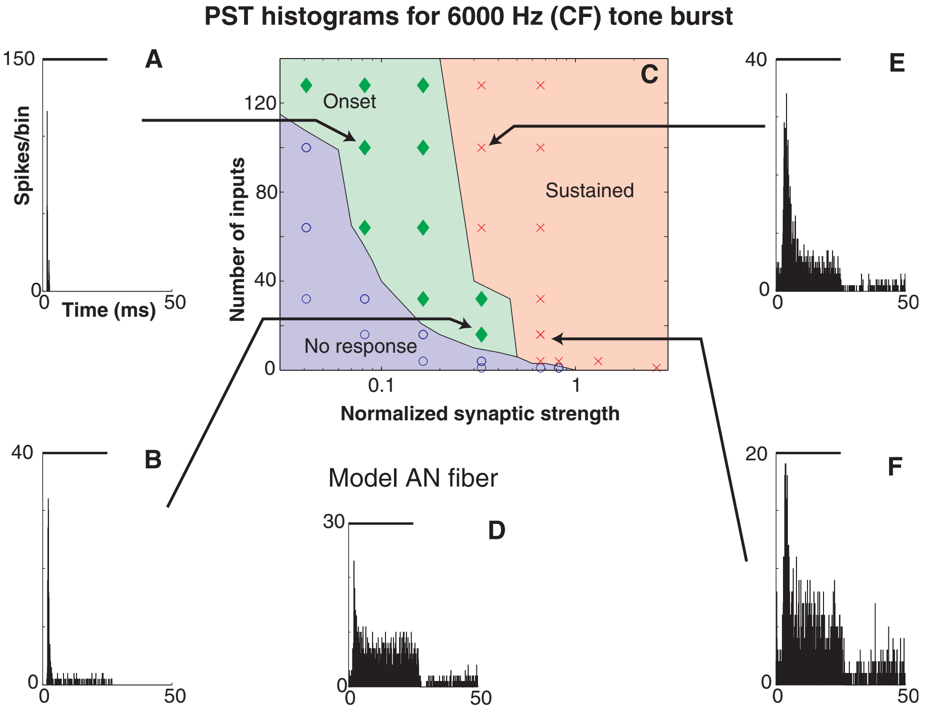


Figure 3 PST histograms of high-frequency, CF tone burst responses as a function of synaptic strength and number of independent AN inputs. A, B, E, and F: PST histograms of model responses for a synaptic strength and number of independent AN inputs indicated by arrows. Synaptic strength is indicated by arrows (250 stimuli, bin width = 0.1 ms). D. PST histogram of a 6-kHz (CF) tone burst response (50 dB SPL) of a model AN fiber (250 stimuli, bin width = 0.1 ms). Bars above the histograms indicate duration of the stimulus. C. Variation of the shape of PST histograms of high-frequency CF tone responses with a variable number of inputs and synaptic strength. Criterion for classifying onset PST histograms was the same as used by Winter and Palmer [46] — the ratio of onset rate to steady-state rate must be greater than 10, and the steady-state rate must be less than 50 spikes/s. It is assumed that the cell is unresponsive to CF tonal stimuli when the threshold is greater than 70 dB SPL.

3.2.2 Phase-Locking to Low-Frequency Tones

VCN onset units phase-lock to low-frequency tones (< 1 kHz) with greater precision than most other VCN units; their synchronization to such signals is also greater than that of AN fibers [13][35]. These units also entrain to low-frequency tones below 1 kHz (i.e., they discharge on every cycle of the tone). Entrainment is a singular property that is rarely (if ever) observed in other types of VCN units. Our analysis shows how entrainment to a 600-Hz tone burst (presented at 90 dB SPL) depends on synaptic strength and the number of independent AN inputs associated with a model neuron (Figure 4).

For reference, an interspike-interval histogram and a period histogram of an AN fiber are shown in Figure 4B. These histograms are similar to those observed in actual AN fibers. Figure 4A shows interval and period histograms for the model when it has 100 inputs and the synapses are weak. The period histogram shows that the response is highly synchronized to the stimulus. The single peak in the interval histogram at a duration associated with a single stimulus period (1.7 ms) indicates that the response is entrained to the stimulus. There is a

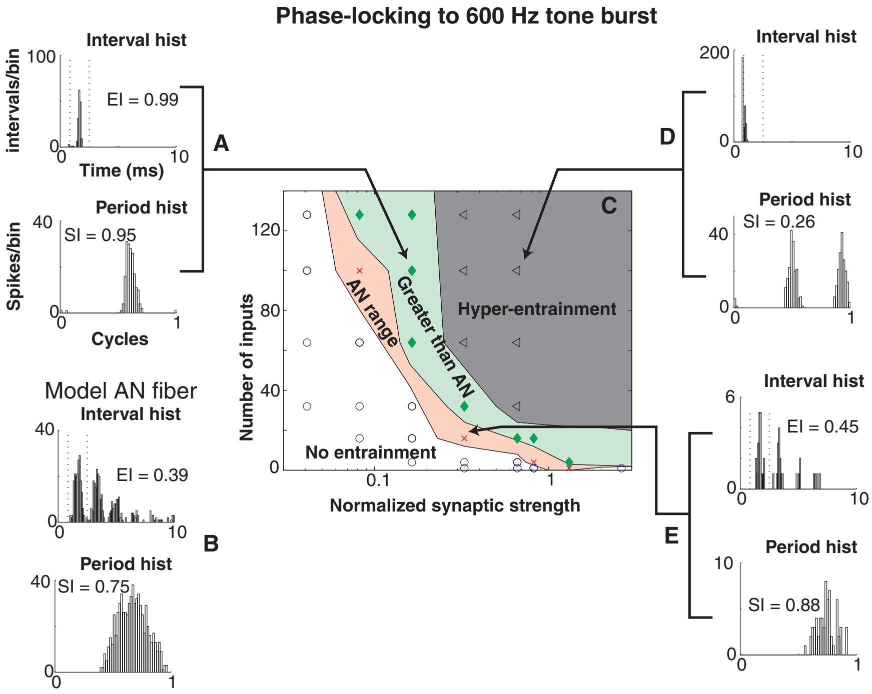


Figure 4 Entrainment of the model neuron (CF = 6 kHz) to a 90 dB SPL, 600-Hz tone as a function of synaptic strength and number of independent AN inputs. A, D, and E: Interval histograms (20 stimuli, bin width = 0.1 ms) and period histograms (64 bins) of the model with the number of inputs and synaptic strength indicated by arrows. B. Interval histogram (200 stimuli, bin width = 0.1 ms) and period histogram (64 bins) for the response of a model AN fiber (CF = 6 kHz). C. Variation of entrainment index (EI) with synaptic strength and number of inputs. EI is the ratio of the number of intervals with a duration equal to a stimulus period divided by the total number of intervals. The different symbols and shading delineate four separate regions according to qualitative differences in entrainment: i) No entrainment (EI = 0), ii) range for model AN inputs (0 < EI < 0.78), iii) greater than model AN inputs (EI > 0.78), and iv) hyper-entrainment (more than one spike per cycle).

qualitative change in response characteristics when the synaptic strength is increased (Figure 4D). The presence of two separate peaks in the period histogram indicates that the response has multiple spikes in a stimulus period (i.e., hyper-entrainment). When both the synaptic strength and the number of inputs are low, entrainment in the model is similar to entrainment observed in the AN inputs (Figure 4E).

Figure 4C illustrates how entrainment varies in the model as a function of the synaptic strength and the number of independent AN inputs. The parameter space is divided into four regions according to how well the model response entrains to the tone (quantified by the entrainment index [EI], defined in the caption of Figure 4):

(1) no entrainment,

(2) entrainment in the range of the AN model (0 < EI < 0.78),

(3) entrainment greater than the AN model, which is the range we expect to be appropriate for onset units (EI > 0.78), and

(4) hyper-entrainment (multiple spikes per cycle of the stimulus).

Entrainment occurs within the onset range when there are either few inputs with a high synaptic strength or there are many weak AN inputs. Synchronization to low-frequency

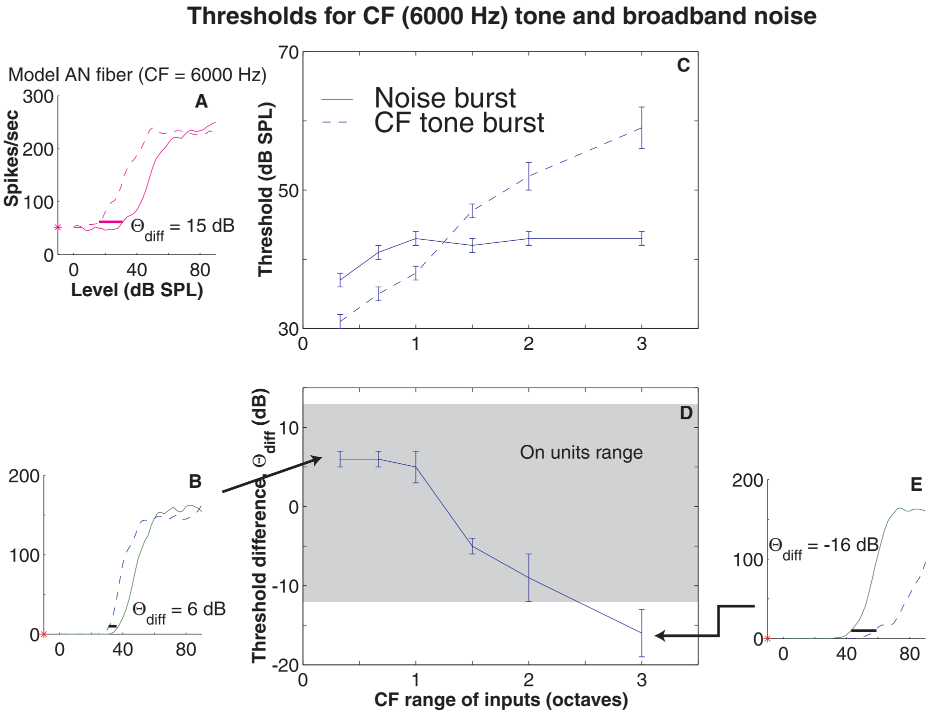


Figure 5 Broadband noise (bandwidth = 20 kHz) threshold and 6-kHz (CF) tone threshold as a function of CF range of inputs. A. Rate versus level for broadband noise (solid) and CF tone (dashed) for a model AN fiber. Horizontal bars show difference between broadband noise threshold and CF-tone threshold. B and E. Same as in the model for the CF range of inputs indicated by arrow. C. Broadband noise threshold and CF tone threshold as a function of CF range of inputs. D. Difference between broadband noise threshold and CF tone threshold (Θ_{diff}) as a function of CF range of inputs. Shaded area is the range for Θ_{diff} in onset units with CFs near 6 kHz [46].

tones and the standard deviation of first-spike latency in high-frequency tone-burst responses also show similar, but more weakly constrained dependence on synaptic strength and the number of inputs.

Taken together, the results of Figures 3 and 4 show that the model must have relatively weak synapses and many independent AN inputs (> 32) in order to exhibit both onset PST histograms for high-frequency tones and entrainment to low-frequency tones. In this parameter range, the model also has a spontaneous rate, a standard deviation of first-spike latency and synchronization indices characteristic of onset units. In the following figures, the number of inputs is 100 and the synaptic strength is 0.16. These values correspond to the region of the parameter space (Figure 3C and 4C) where the model exhibits temporal response characteristics typical of onset units.

3.3 CF Distribution of AN Inputs

Winter and Palmer [46] have found that the threshold for broadband noise (in dB SPL) exceeds the CF-tone threshold by more than 15 dB in AN fibers and VCN chopper units but is less than 15 dB among onset units. They propose that this particular response property can be accounted for by assuming that onset units integrate their inputs across a broader frequency range than do other VCN units or AN fibers. Their proposal is based on the difference

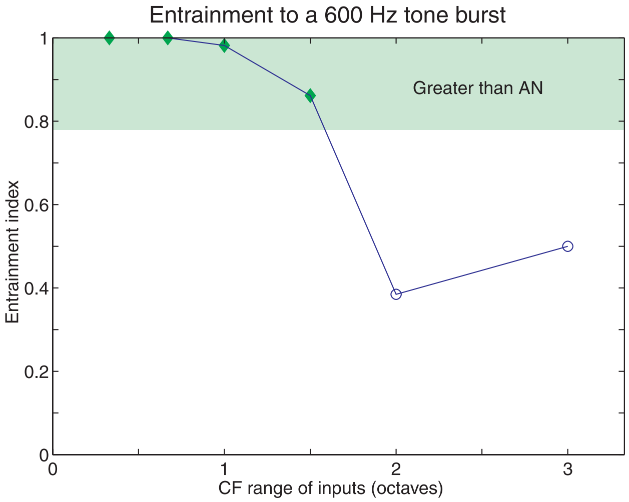


Figure 6 Entrainment to a 600-Hz tone as a function of the CF range of AN inputs for a model cell with a CF of 6 kHz. The EI decreases with an increasing CF range of inputs. Shaded region (EI > 0.78) indicates an entrainment greater than occurs in model AN fibers and corresponds to the similarly labeled region in Figure 4.

between broadband noise energy and CF-tone energy integrated by a linear broadband filter being less than the energy difference for a linear narrow band filter. We examined whether a broad CF range of AN inputs to the model is required in order to obtain a difference in threshold between broadband noise and CF tones in the range observed in onset units.

Figure 5 shows how the CF range of AN inputs in the model affects thresholds for broadband noise and CF tones. Serving as a reference point, Figure 5A shows rate versus level curves for broadband noise and CF tones associated with an AN input from the model. The rate-level curve for noise is shifted to higher stimulus levels relative to the rate-level curve for tones. The difference in threshold between the tone and noise is 15 dB. Figure 5B shows rate-level curves for the model when the CF range of AN inputs is 1/3 octave. Although the curve for noise is displaced toward higher stimulus levels relative to the curve for tones (as occurs in the AN model), the threshold difference is only 6 dB. Figure 5E shows rate–level curves for the model when the CF range of AN inputs is 3 octaves. In this case the rate-level curve for noise is shifted toward lower stimulus levels relative to the rate-level curve for tones. The threshold for noise is actually 16 dB less than the threshold for tones.

Figure 5D shows that the difference between the broadband noise threshold and the CF-tone threshold decreases monotonically as the CF range of the AN inputs increases from 1/3 octave to 3 octaves. Figure 5C shows the noise thresholds and the tone thresholds used to compute the threshold differences. The threshold difference decreases as the CF range of AN inputs increases because the noise threshold varies less than the tone threshold. These observations follow from the properties associated with the fast membrane dynamics as well as from the large number of AN inputs acting via weak synapses that make the model response dependent on coincident spikes on the inputs. The model responds more readily when there are spikes associated with several inputs as compared to when there are only a few inputs. The noise threshold does not vary much because the number of AN inputs responding to the stimulus stays relatively constant as the CF range of inputs increases from 1/3 octave to 3 octaves. On the other hand, the tone threshold varies substantially because the number of inputs that respond to the stimulus varies greatly as the CF range of inputs increases. When the CF range of AN inputs is small, a tone evokes a large response in many of the model inputs, thus leading to a low threshold. When the CF range of AN inputs is broad, only a few model inputs respond to the tone, thus leading to a high threshold.

For all of the CF ranges of AN inputs examined, the model produced onset PST histograms and low spontaneous rates. However, the fine-time structure of the discharge pattern associated with low-frequency (entrainment and synchronization) and high-frequency tone bursts (standard deviation of first-spike latency) varied as a function of the CF range of AN inputs. Of these response properties, entrainment to low-frequency tones was the most sensitive to the CF range of AN inputs.

Figure 6 shows that entrainment to a 90-dB SPL, 600-Hz tone decreases as the CF range of AN inputs increases in the model. A major reason why entrainment decreases is that AN fibers with different CFs have different response latencies introduced by the cochlear travelling wave. Therefore, there is an increasing degree of desynchronization of spikes across the tonotopically organized array of AN fibers as the CF range of inputs increases. Because the model is highly sensitive to temporal coincidence of spikes from the array of AN fibers, this desynchronization causes entrainment in the model to diminish.

These results show that the fine-time structure of discharge patterns, as well as the thresholds associated with broadband noise and CF tones, constrain the CF range of AN inputs to the model. The shaded regions of Figure 5D and Figure 6 indicate the range of threshold difference and entrainment, respectively, for VCN onset units. In the model unit, the CF range of AN inputs needs to be less than 1.5 octaves for both the threshold difference and the entrainment to be within the range observed for onset units.

4. Discussion

In this chapter, we have identified a minimum set of properties required for obtaining onset discharge patterns in response to acoustic stimulation. Specifically, we have determined constraints on the nature of the AN inputs and associated synapses such that the model exhibits responses characteristic of VCN onset units. Although several parameters affect each of the onset response properties, some are especially important. A weak synaptic strength is the most important determinant of an onset PST histogram for high-frequency tone bursts. This finding is consistent with results from previous models of VCN bushy cells and octopus cells [23][24][38]. We also have found that entrainment to low-frequency tones is jointly determined by synaptic strength, number of inputs and the range of CFs spanned by the inputs. In order to simultaneously produce onset PST histograms for high-frequency tone bursts and entrainment to low-frequency tones, the model must have weak synapses and many independent AN inputs (> 32) whose CFs span less than 1.5 octaves. The difference between broadband noise threshold and CF tone threshold also depends greatly on the CF range of AN inputs. Together, the threshold differential and entrainment constrain the CF range of AN inputs to be at most 1.5 octaves in our model.

Model parameters that produce realistic onset responses are generally consistent with the anatomical data. Specifically, the constraint of a large number of AN inputs giving rise to weak synapses is consistent with anatomical observations from octopus cells and other labeled onset responders that these cells have a large number of synapses, all of which are small relative to the size of the cell [15][27][31][41]. On the other hand, the model predicts that the CF range of inputs needs to be limited to 1.5 octaves, a property in apparent conflict with the view that onset units receive AN inputs from a very wide CF range. This view is based on the observation that dendrites of octopus cells and other labeled onset units are oriented perpendicularly to iso-frequency bands of incoming AN fibers and therefore must span a wide range of CFs [15][22][41]. However, for a CF of 6 kHz, our 1.5 octave limit corresponds to a substantial length (20%) of the basilar membrane according to the Liberman cochlear frequency map for the cat [26]. Since the CF distribution of onset units is biased toward high frequencies [46], this 20% of the basilar membrane length might actually represent an even more substantial proportion of the relevant array of AN inputs to the onset cell population. A more rigorous test of our model prediction would require quantification of the length of the basilar membrane innervated by AN inputs to labeled onset neurons.

Previous model-based investigations have focused on onset responses to sinusoidal signals. Synaptic strength was found to be the principal factor determining PST histogram shape in response to high-frequency tone bursts in models of octopus and bushy cells [23][24][38][39]. Rothman and Young [39] have further shown that convergence of many independent AN inputs is required for their model of VCN bushy cells to exhibit the exquisite synchronization to low-frequency tones observed in onset units. Our findings are consistent with these observations but show that entrainment imposes an even more powerful constraint on the model. Evans [9] has shown, using a phenomenological model, that fast membrane excitability and threshold accommodation together provide more faithful onset PST histograms and a better match to the threshold differential between tonal and noise stimuli than either property alone. Our own model results are in accord with those of Evans and further show that the CF range of inputs is an important factor as well. Our results extend the findings of previous studies by considering the model properties required to simultaneously explain several onset-response properties. Specifically, we find that onset PST histograms for high-frequency tone bursts, entrainment to low-frequency tones and the threshold difference between broadband noise and CF tones are particularly informative measures for constraining model parameters when considered together as a group.

In this chapter, we have shown how model responses depend on properties of AN inputs and associated synapses for fixed membrane properties obtained by fitting the membrane model to octopus cell responses to current injections. It is known that membrane electrical properties differ among cells in the VCN [11][15][30][47]. Electrical properties may vary across the heterogeneous onset population of neurons as well. Certain properties of the model response patterns (e.g., the detailed features of PST histograms) vary with changes in the cell's membrane characteristics. However, the key response properties examined in this chapter do not change significantly as long as parameters of the membrane model remain within certain limits. Specifically, our conclusions regarding inputs and synapses remain valid as long as the time constant of membrane excitability is small (< 1 ms) [20][21]. Furthermore, in order to produce both onset PST histograms in response to high-frequency tone bursts and entrainment to low-frequency tones, the model must have a strong and relatively rapid accommodative threshold.

Although accommodation is a phenomenological concept, it could be instantiated by incorporating voltage-gated ion channels into the model. Specifically, at least two kinds of channels may be necessary to include. The transient maximum in membrane voltage observed upon stimulation by a depolarizing, sub-threshold step current can be implemented using an outward-rectifying ion channel that is active at voltages slightly above the resting potential. The low-threshold potassium channel that is blocked by 4-aminopyridine is such a channel. It is found in VCN bushy and octopus cells [12][15][28]. Accommodation, in the form used in our model, also causes a transient decrease in membrane voltage at the onset of negative current steps and a transient peak in the membrane voltage at their offsets (as well as spikes for sufficiently negative-going current injections). These aspects of the response to negative current steps are also observed in octopus cells. Channels that rectify inward at voltage levels below the resting potential can be used for implementing these membrane properties. The hyperpolarization-activated channel that is blocked by Cs^+ (a.k.a. I_h) and found in principal cells of the medial nucleus of the trapezoid body and octopus cells of the VCN [3][15] possesses such characteristics. A model for the octopus cell that includes the sorts of channels described has been implemented; its responses to step-current injections are qualitatively similar to data recorded from octopus cells [7].

Accommodation might also be implemented using other mechanisms, such as desensitization of synaptic receptors [44]. Although receptor desensitization may not be required to account for octopus-cell-response properties (because membrane voltage shows accommodation in the absence of synaptic inputs [12][15]), it may play a role in other onset responding classes of neurons.

When onset units were first studied, recurrent inhibition was proposed as a possible mechanism for producing the onset-discharge pattern of response to tone bursts [13]. This particular hypothesis lost much of its appeal as a consequence of intracellular recordings from onset units failing to manifest sustained hyperpolarization in response to tonal stimulation [37]. Recent work with iontophoretic injection of inhibitory transmitter antagonists [10] has provided evidence that inhibition plays a role in shaping the response properties of a subclass of onset units. Despite this recent evidence we have not included inhibitory inputs in our model because too little is known about the origin of these inputs to develop a quantitative model and because our primary focus is on response properties common to all types of onset units.

Identifying which neuronal characteristics underlie responses to sound is part of the more general problem of correlating cell structure with function. Understanding such relationships in one class of neurons may help understand its relation in other cell classes. For example we have observed a trade-off between the CF range (and therefore latency spread) of AN inputs and the ability of onset units to entrain to low-frequency tones. Recent evidence suggests that cortical pyramidal cells are similar to VCN onset units in that they act as coincidence detectors [1][42]. Thus, our trade-off may be an instance of a general constraint (with potential counterparts among the pyramidal cells of the cortex) on the ability of neurons to precisely follow successive transients in the stimulus when their inputs are desynchronized.

In this study we have used a phenomenological model, rather than a detailed biophysical model, to identify the characteristics underlying response patterns of onset units. In the context of this book, which examines computational methods for studying audition, it is worth noting our example of a phenomenological model of a neuron that yields useful results pertaining to the mechanisms underlying neuronal signal processing.

5. Conclusions

Using a simple functional model of a VCN cell, we have determined the characteristics required for obtaining onset-response patterns to acoustic stimulation. We find that no single neuronal property confers onset-response characteristics on a VCN neuron. Instead, a combination of properties must be simultaneously present for the model to manifest all response characteristics of onset units. Many independent AN inputs (> 32), weak synapses, fast membrane dynamics, and a high-pass filtering process (such as an accommodative threshold) are all necessary for simulating onset response properties. The CF range of AN inputs further affects the fine temporal structure of discharge as well as the threshold for tones and noise. Together, these effects strongly constrain the entire set of model parameters. Our results suggest three sets of data for characterizing the underlying neuronal features of an onset unit:

(1) PST histograms of high-frequency tone-burst responses,
(2) entrainment to low-frequency tones, and
(3) the differential response threshold of broadband noise and CF tones.

Acknowledgments

Supported by research grant DC02258 and training grant DC00038 from the National Institute on Deafness and Other Communicative Disorders, National Institutes of Health. We thank Stephanie Hequembourg and Craig Atencio for comments on the manuscript.

Note

1. In the original formulation of Hill [16], accommodation was implemented as an increase in the time-varying threshold. Because intracellular recordings from octopus cells show an accommodation of membrane voltage in response to sustained current injections [12][14], we prefer to model accommodation as a change in voltage, *v(t),* rather than as a change in threshold. Therefore a fixed threshold was used rather than one of a time-varying nature. Because the difference between threshold and membrane voltage determines spiking, the two formulations are mathematically equivalent, but the current formulation is better suited for comparison with intracellular data.

References

[1] Abeles, M. “Role of the cortical neuron: Integrator or coincidence detector.” *Israel J. Med. Sci.*, 18: 83–92, 1982.

[2] Arle, J. and Kim, D. “Neural modeling of intrinsic and spike discharge properties of cochlear nucleus neurons.” *Biol. Cybern.*, 64: 273–283, 1991.

[3] Banks, M., Pearce, R., and Smith, P. “Hyperpolarization-activated cation current (I_h) in neurons of the medial nucleus of the trapezoid body: Voltage-clamp analysis and enhancement by norepinephrine and cAMP suggest a modulatory mechanism in the auditory brain stem.” *J. Neurophysiol.*, 70: 1420–1432, 1993.

[4] Blackburn, C. and Sachs, M. “Classification of unit types in the anteroventral cochlear nucleus: PST histograms and regularity analysis.” *J. Neurophysiol.*, 62: 1303–1329, 1989.

[5] Bourk, T. *Electrical Responses of Neural Units in the Anteroventral Cochlear Nucleus of the Cat.* Ph. D. Thesis, Massachusetts Institute of Technology, 1976.

[6] Bregman, A. S. *Auditory Scene Analysis.* Cambridge, MA: MIT Press, 1990.

[7] Cai, Y., Walsh, E. J., and McGee, J. “Mechanisms of onset responses in octopus cells of the cochlear nucleus: Implications of a model.” *J. Neurophysiol.*, 78: 872–883, 1997.

[8] Carney, L. “A model for the responses of low-frequency auditory-nerve fibers in cat.” *J. Acoust. Soc. Am.*, 93: 401–417, 1993.

[9] Evans, E. F. “Modeling characteristics of Onset-I cells in guinea pig cochlear nucleus.” *Proc. NATO Advanced Study Institute on Computational Hearing*, S. Greenberg and M. Slaney (eds.), pp. 1–6, 1998.

[10] Evans, E. F. and Zhao, W. “Periodicity coding of the fundamental frequency of harmonic complexes: physiological and pharmacological study of onset units in the ventral cochlear nucleus.” In *Psychophysical and Physiological Advances in Hearing*, A. R. Palmer, A. Rees, A. Q. Summerfield and R. Meddis, (eds.), London: Whurr Publishers, pp. 186–192, 1998.

[11] Feng, J., Kuwada, S., Ostapoff, E.-M., Batra, R. and Morest, D. “A physiological and structural study of neuron types in the cochlear nucleus. I. Intracellular responses to acoustic stimulation and current injection.” *J. Comp. Neurol.*, 346:1–18, 1994.

[12] Ferragamo, M. J. and Oertel, D. “Shaping of synaptic responses and action potentials in octopus cells.” 21st Midwinter Meeting Assn. Res. Otolaryngol., St. Petersburg Beach, FL, 1998.

[13] Godfrey, D., Kiang, N., and Norris, B. “Single unit activity in the posteroventral cochlear nucleus.” *J. Comp. Neurol.*, 162: 247–268, 1975.

[14] Golding, N. L., Ferragamo, M. J., and Oertel, D. “Personal communication.” 1998.

[15] Golding, N. L., Robertson, D., and Oertel, D. “Recordings from slices indicate that octopus cells of the cochlear nucleus detect coincident firing of auditory-nerve fibers with temporal precision.” *J. Neurosci.*, 15: 3138–3153, 1995.

[16] Hill, A. V. “Excitation and accommodation in nerve.” *Proc. Roy. Soc. (London), Ser. B*, 119: 1936.

[17] Irvine, D. R. F. *The Auditory Brainstem*. Berlin: Springer-Verlag, 1986.

[18] Johannesma, P. I. M. “The pre-response stimulus ensemble of neurons in the cochlear nucleus.” *Proc. Symp. Hearing Theory.* Eindhoven: IPO, pp. 58–69, 1972.

[19] Johnson, D. and Swami, A. “The transmission of signals by auditory-nerve fiber discharge patterns.” *J. Acoust. Soc. Am.*, 74: 493–501, 1983.

[20] Kalluri, S. and Delgutte, B. “A general model of spiking neurons applied to onset responders in the cochlear nucleus.” Presented at *Computational Neurosciences Conference*, Cambridge, MA, 1996.

[21] Kalluri, S. and Delgutte, B. “An electrical circuit model for cochlear nucleus onset responders.” Presented at *20th Midwinter Meeting of Assoc. Res. Otolaryngol.*, St. Petersburg Beach, FL, 1997.

[22] Kane, E. C. “Octopus cells in the cochlear nucleus of the cat: Heterotypic synapses upon homeotypic neurons.” *Intern. J. Neurosci.*, 5: 251–279, 1973.

[23] Kipke, D. R. and Levy, K. L. “Sensitivity of the cochlear nucleus octopus cell to synaptic and membrane properties: A modeling study.” *J. Acoust. Soc. Am.*, 102: 403–412, 1997.

[24] Levy, K. and Kipke, D. “Mechanisms of the cochlear nucleus octopus cell’s onset response: synaptic effectiveness and threshold.” *J. Acoust. Soc. Am.*, 103: 1940–1950, 1998.

[25] Levy, K. L. and Kipke, D. R. “A computational model of cochlear nucleus octopus cells.” *J. Acoust. Soc. Am.*, 102: 391–402, 1997.

[26] Liberman, M. C. “The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency.” *J. Acoust. Soc. Am.*, 72: 1441–1449, 1982.

[27] Liberman, M. C. “Central projections of auditory-nerve fibers of differing spontaneous rate. I. Posteroventral and dorsal cochlear nucleus.” *J. Comp. Neurol.*, 327: 17–36, 1993.

[28] Manis, P. and Marx, S. “Outward currents in isolated ventral cochlear nucleus neurons.” *J. Neurosci.*, 11: 2865–2880, 1991.

[29] Monnier, A. *L’Excitation Electrique des Tissus*. Hermann: Paris, 1934.

[30] Oertel, D., Wu, S. H., Garb, M., and Dizack, C. “Morphology and physiology of cells in slice preparations of the posteroventral cochlear nucleus of mice.” *J. Comp. Neurol.*, 295: 136–154, 1990.

[31] Ostapoff, E.-M., Feng, J., and Morest, D. “A physiological and structural study of neuron types in the cochlear nucleus. II. Neuron types and their structural correlation with response properties.” *J. Comp. Neurology*, 346: 19–42, 1994.

[32] Pfeiffer, R. “Classification of response patterns of spike discharges for units in the cochlear nucleus: Tone-burst stimulation.” *Exp. Brain Res.*, 1: 220–235, 1966.

[33] Rashevsky, N. “Outline of a physico-mathematical theory of excitation and inhibition.” *Protoplasma*, 20: 42–56, 1933.

[34] Rhode, W., Oertel, D., and Smith, P. “Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus.” *J. Comp. Neurol.*, 213: 448–463, 1983.

[35] Rhode, W. and Smith, P. “Encoding of timing and intensity in the ventral cochlear nucleus of the cat,” *J. Neurophysiol.*, 56: 261–286, 1986.

[36] Ritz, L. and Brownell, W. “Single unit analysis of the posteroventral cochlear nucleus of the decerebrate cat.” *Neuroscience*, 7: 1995–2010, 1982.

[37] Romand, R. “Survey of intracellular recording in the cochlear nucleus of the cat.” *Brain Res.*, 148: 43–65, 1978.

[38] Rothman, J., Young, E., and Manis, P. “Convergence of auditory nerve fibers onto bushy cells in the ventral cochlear nucleus: Implications of a computational model.” *J. Neurophysiol.*, 70: 2562–2583, 1993.

[39] Rothman, J. S. and Young, E. D. “Enhancement of neural synchronization in computational models of ventral cochlear nucleus bushy cells.” *Aud. Neurosci.*, 2: 47–62, 1996.

[40] Rouiller, E. and Ryugo, D. “Intracellular marking of physiologically characterized cells in the ventral cochlear nucleus of the cat.” *J. Comp. Neurol.*, 225: 167–186, 1984.

[41] Smith, P. H. and Rhode, W. S. "Structural and functional properties distinguish two types of multipolar cells in the ventral cochlear nucleus." *J. Comp. Neurol.*, 282: 595–616, 1989.

[42] Softky, W. R. and Koch, C. "The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs." *J. Neurosci.*, 13: 334–350, 1993.

[43] Stevens, K. N. and Blumstein, S. E. "Invariant cues for place of articulation in stop consonants." *J. Acoust. Soc. Am.*, 64: 1358–1368, 1978.

[44] Trussell, L. O., Zhang, S., and Raman, I. M. "Desensitization of AMPA receptors upon multi-quantal neurotransmitter release." *Neuron*, 10: 1185–1196, 1993.

[45] Westerman, L. and Smith, R. "A diffusion model of the transient response of the cochlear inner hair cell synapse." *J. Acoust. Soc. Am.*, 83: 2266–2276, 1988.

[46] Winter, I. and Palmer, A. "Level dependence of cochlear nucleus onset unit responses and facilitation by second tones or broadband noise." *J. Neurophysiol.*, 73: 141–159, 1995.

[47] Wu, S. and Oertel, D. "Intracellular injection with horseradish peroxidase of physiologically characterized stellate and bushy cells in slices of mouse anteroventral cochlear nucleus." *J. Neurosci.*, 4, 1577–1588,1984.

[48] Zurek, P. M. "The precedence effect." In *Directional Hearing*, W. A. Yost and G. Gourevitch (eds.). New York: Springer–Verlag, 1987.

MODELING THE COCHLEA

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

MODELING THE COCHLEA

Malcolm Slaney

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120, USA

We use computational modeling to build quantitative simulations that describes the auditory system. Throughout this book, authors describe how models are built and are compared to real auditory systems. Using these models, we gain insight into how the modeled auditory system works, and thus are able to make predictions about how that system will perceive a new sound.

In the auditory field, the most fervent computational modeling has been directed at the periphery—often the cochlea. These cochlear models are important, not only because they help to inform a debate about how the cochlea works, but also because they permit work that uses the output of the cochlea.

One of the most difficult problems in modeling is deciding what is the optimal level of detail. Given a correct model, including more detail usually increases the accuracy of model predictions; but also increases the computational burden. Furthermore, the effort needed to model these details might be wasted if the user of the model does not care about these details. Thus, a model of pitch perception based on inner-hair cell firing rates probably will not care about the exact timing of each spike.

Simplifying our models helps us to understand how the model will behave in new situations. Yet these simplifications can be misleading. For example, a simple model of the cochlea based on Fourier analysis might omit the fact that the peak response for any one frequency moves along the basilar membrane as the sound level changes. This simplification might not affect the outcome of many masking simulations, but in other cases it could lead a modeler to the wrong conclusion.

Cochlear modelers work in three different areas: deciphering the cochlea, replicating the cochlea's behavior, and predicting how the periphery will respond to a prosthetic device's outputs.

Scientists use the most sophisticated mathematics to create detailed mechanical models of the tissues in the cochlea. This task is especially difficult because many tissues are connected to each other and it is hard to separate their behaviors. How does the tectorial membrane connect to the basilar membrane? Why does the bandwidth of a cochlear filter change so dramatically with increasing sound level? Scientists have built many cochlear models to answer these questions [1].

These detailed mechanical models are computationally intensive and provide more details then we need to investigate in many auditory tasks. For much psychoacoustics research the front-end of the auditory system is modeled with critical-band filters. Critical bands are based on psychoacoustic measurements of an auditory filter's characteristics. Listeners hear, for example, a tone embedded in noise. The noise exists everywhere except at frequencies near the tone's frequency. The model assumes that people can hear the tone when the energy of the noise within any one auditory filter is less than a fixed fraction of the energy in the tone. The width of the noise notch is a measurement of the auditory channel's bandwith, often expressed as the equivalent rectangular bandwidth.

The first chapter in this section describes the gammachirp, an extension to the gammatone filter. The gammachirp was derived as an optimal time-frequency analyzer for audio signals. Fitting nonlinear gammachirp filters that match critical-band data over a range of sound levels is easy. Irino and Unoki describe a simple implementation of the gammachirp. They combine a normal gammatone filter with an asymmetric low-pass filter, which changes with level. The pair of filters closely approximates the desired gammachirp response, and thus is a good model of the auditory periphery.

A third use for peripheral auditory models is to predict how humans will hear with prosthetic devices that compensate for hearing loss. In most cases, simple amplification—a hearing aid—is sufficient. With profound hearing loss, electrodes threaded into the cochlea are stimulated electrically, causing the auditory nerve to fire. Unlike that of a hearing aid, the interaction of a cochlear implant with the wearer's auditory system is hard to predict.

Bruce and his colleagues refine a simple model of neural firing that predicts the sounds perceived by a cochlear implant user. The simplest model predicts that the neuron always fires when the intensity of the stimulating pulse is above a fixed threshold. Bruce's model predicts that the neuron's probability of discharge grows slowly around the threshold. The authors demonstrate that a probabilistic model of neural excitation is a more accurate predictor of perceived loudness, both at the threshold of hearing and at the level at which loudness becomes uncomfortable.

Computational models give us insights and predictions about the function of the auditory system. These two chapters provide just two examples: one to simulate the cochlea and to provide a tool for researchers investigating higher-level functions, the other to predict how a prosthetic device will behave in a human patient.

Reference

[1] Lewis, E. R., Long, G. R., Lyon, R. F., Narins, P. M., Steele, C. R., Hecht-Poinar, E. (eds.). *Diversity in Auditory Mechanics*. Singapore: World Scientific, 1997.

AN ANALYSIS/SYNTHESIS AUDITORY FILTERBANK BASED ON AN IIR GAMMACHIRP FILTER

Toshio Irino[1] and Masashi Unoki[1,2]

[1] *ATR Human Information Processing Research Labs*
2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-02, Japan
[2] *Japan Advanced Institute of Science and Technology*
1-1 Asahidai Tatsunokuchi Nomi Ishikawa, 923-1292, Japan

1. Introduction

A number of auditory models have been developed for telecommunications systems that incorporate human auditory characteristics. Recent attempts do a good job, based on physiological comparisons, simulating the peripheral auditory system (for a review, see [4]). But, unfortunately, none of these models have had as much success in speech recognition systems as linear predictive analysis and the Fourier transforms. There are a number of reasons for this dilemma. An obvious problem is that realistic, non-linear auditory models require complex calculation that preclude real-time processing. However, this problem should be resolved in the near future by fast digital signal processors. Another factor is that these models do not facilitate proper signal resynthesis, which is straightforward for both linear predictive analysis and Fourier analysis because of their linearity.

Linear auditory filterbanks or wavelet transforms have been used for signal resynthesis [2][24], but they are unable to account for the dynamic characteristics of basilar-membrane motion. Iterative procedures for reconstructing signals from cochleagrams (i.e., short-time averaged amplitude responses of basilar membrane motion without phase information) [6][23] are applicable to such non-linear filterbanks, but can not guarantee the precision of the resynthesis because of local minima. Thus it is desirable to develop an adaptive auditory filterbank that also provides a sound resynthesis procedure resulting in no perceptual distortion. This paper shows that such an adaptive, analysis/synthesis filterbank is possible through the implementation of a new "gammachirp" auditory filter [9].

The gammachirp function was analytically derived to have minimal uncertainty in a joint time-scale representation [1][7][8]. The gammachirp auditory filter is an extension of the popular gammatone filter (for a review, see [17]); it has an additional frequency modulation term to produce an asymmetric amplitude spectrum. When the degree of asymmetry is associated with the stimulus level, the gammachirp filter provides an excellent fit to 12 sets of notched-noise masking data from three different studies [9]. The gammachirp has a much simpler impulse response than recent physiological models on cochlear mechanics [4], which do provide a good fit to human masking data. Moreover, the chirp term in the gammachirp is consistent with physiological observations on frequency-modulations or frequency "glides" in measurements of the mechanical responses of the basilar membrane [3][15][20].

The gammachirp filter has been implemented as a finite impulse response (FIR) filter because the gammachirp is defined as a time-domain function. Including this filter in an auditory filterbank, however, poses problems. For simulations of the dynamic characteristics of the cochlea the filter coefficients have to be recalculated and then convolved with the signal on a moment-by-moment basis. Unfortunately, the large number of FIR coefficients, especially at low frequencies, precludes fast implementations. Moreover, the simulation becomes unrealistic if the filter output is not calculated simultaneously with the update of the filter coefficients. The calculation of the filter output and the update of the filter coefficients need to be performed simultaneously. Therefore, the gammachirp filter should be implemented with a small number of filter coefficients using an infinite impulse response (IIR) filter [10][11].

IIR implementations of modified gammatone filters have been developed to introduce asymmetry into auditory filter shapes, i.e., the All-Pole Gammatone Filter (APGF) or One-Zero Gammatone Filter (OZGF) [13][18][22]. The shapes of these filters, however, depend on the sampling rate of the system [10] and have not been directly fitted to psychoacoustic masking data. Moreover, it seems difficult to resynthesize signals from their output representations without uncontrollable errors since they did not provide a well-defined synthesis scheme. These issues are the main topics of this paper.

Section 2 describes an IIR implementation of the gammachirp. Section 2.1 shows the definition and the Fourier transform of the gammachirp decomposed into a gammatone and an asymmetric function. Section 2.2 explains the characteristics of the asymmetric function. Section 2.3 shows that the asymmetric function can be implemented by an IIR “asymmetric compensation filter.” Section 2.4 shows the approximation error in the amplitude spectrum between the original gammachirp filter and the combination of a gammatone and an IIR asymmetric compensation filter. Section 2.5 shows the stability of the inverse filter of the IIR filter, which enables signal resynthesis using the procedure described in Section 3.3. Section 3 shows an implementation of the gammachirp filterbank. Section 3.1 shows an example of an adaptive analysis filterbank controlled by the sound pressure level estimated at the output of the filterbank. Section 3.2 shows another example of a filterbank based on physiological constraints. Finally, Section 3.3 describes an adaptive, analysis/synthesis auditory filterbank that has never been accomplished by conventional auditory models simulating basilar membrane motion.

2. Implementation of the Gammachirp Filters

2.1 Definition and Fourier Transform of the Gammachirp

The complex impulse response of the gammachirp [7][8][9] is given as

$$g_c(t)= at^{n-1}\exp(-2\pi b\mathrm{ERB}(f_r)t)\exp(j2\pi f_r t + jc\ln t + j\phi), \qquad (1)$$

where time $t > 0$, a is the amplitude, n and b are parameters defining the envelope of the gamma distribution, f_r is the asymptotic frequency, c is a parameter for the frequency modulation or the chirp rate, ϕ is the initial phase, ln t is a natural logarithm of time, and ERB (f_r) is the equivalent rectangular bandwidth of the auditory filter at f_r. At moderate levels, ERB(f_r)=24.7 + 0.108f_r in Hz [5]. When $c = 0$, the chirp term, $c$ln t, vanishes and this equation represents the complex impulse response of the gammatone that has the envelope of a gamma distribution function and its carrier is a sinusoid at frequency, f_r [17]. Accordingly, the gammachirp is an extension of the gammatone with a frequency-modulation term.

The Fourier transform of the gammachirp in Equation (1) is derived as follows.

$$\begin{aligned} G_C(f) &= \frac{a\Gamma(n+jc)e^{j\phi}}{\{2\pi bERB(f_r) + j2\pi(f-f_r)\}^{n+jc}} \\ &= \frac{\bar{a}}{\left\{2\pi\sqrt{\bar{b}^2+(f-f_r)^2}\cdot e^{j\theta}\right\}^{n+jc}} \\ &= \bar{a}\cdot\frac{1}{2\pi\left\{\sqrt{\bar{b}^2+(f-f_r)^2}\right\}^{n}\cdot e^{jn\theta}}\cdot\frac{1}{\left\{2\pi\sqrt{\bar{b}^2+(f-f_r)^2}\right\}^{jc}\cdot e^{-c\theta}}\quad, \end{aligned} \tag{2}$$

$$\theta = \arctan\frac{f-f_r}{\bar{b}} \tag{3}$$

where $\bar{a} = a\Gamma(n+jc)e^{j\phi}$ and $\bar{b} = b\mathrm{ERB}(f_r)$. The first term $\bar{a}$ is a constant. The second term is known as the Fourier spectrum of the gammatone,$G_T(f)$. The third term represents an asymmetric function, $H_A(f)$, that is described in more detail in the next Section. If we normalize the amplitude, the frequency response of the gammachirp can be represented as

$$G_C(f) = G_T(f)\cdot H_A(f). \tag{4}$$

The amplitude spectrum is

$$|G_C(f)| = |G_T(f)|\cdot|H_A(f)| = \frac{\bar{a}}{\left\{2\pi\sqrt{\bar{b}^2+(f-f_r)^2}\right\}^{n}}\cdot e^{c\theta}. \tag{5}$$

Obviously, when *c=0*, $|H_A(f)|$ ($=e^{c\theta}$) becomes unity and Equation (5) represents the amplitude spectrum of the gammatone, $|G_T(f)|$. Figure 1 shows the amplitude spectra of (a) a gammachirp filter $|G_C(f)|$, (b) a gammatone filter $|G_T(f)|$, and (c) an asymmetric function $|H_A(f)|$ with the chirp parameter *c*=-2. The amplitude of $|H_A(f)|$ is biased by about -4 dB to normalize the peak of $|G_C(f)|$ to 0 dB. Since the amplitude spectrum of the gammatone filter$|G_T(f)|$ is almost symmetric on a linear-frequency axis, the asymmetric function $|H_A(f)|$ introduces spectral asymmetry and a peak frequency shift into the gammachirp $|G_C(f)|$.

The peak frequency f_p in the amplitude spectrum is obtained analytically by setting the derivative of Equation (4) to zero and solving the equation for the peak frequency. The result is

$$f_p = f_r + \frac{c\cdot\bar{b}}{n} = f_r + \frac{c\cdot b\mathrm{ERB}(f_r)}{n}. \tag{6}$$

Therefore, the size of the peak shift is proportional to the chirp parameter, *c,* and the ratio of the envelope parameter, *b* ERB(f_r), to *n*.

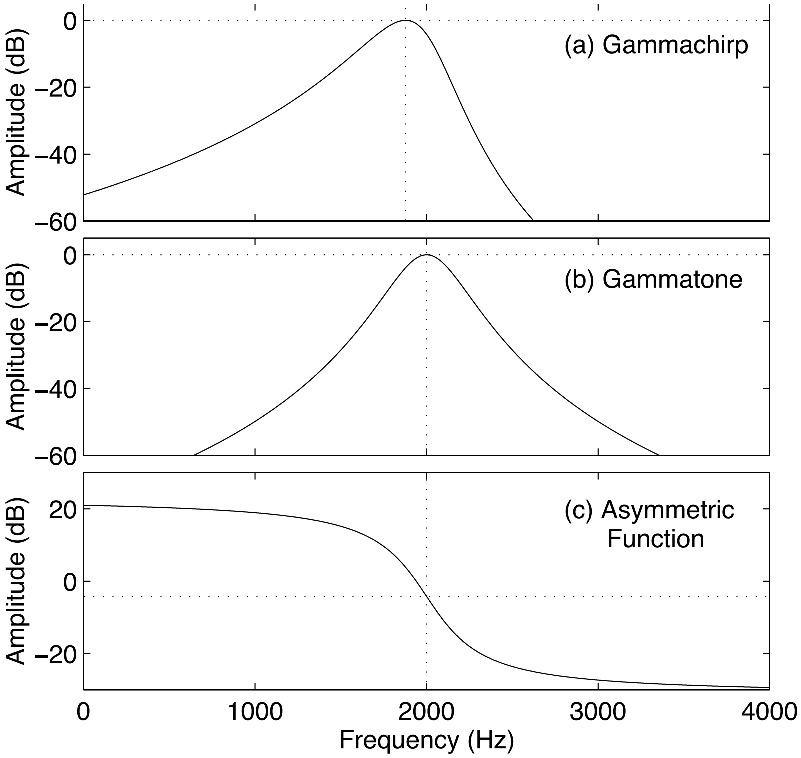


Figure 1 Amplitude spectra of (a) a gammachirp filter $|G_C(f)|$, (b) a gammatone filter $|G_T(f)|$, and (c) an asymmetric function $|H_A(f)|$, where n = 4, b = 1.019, c = -2, and f_r = 2000 Hz.

2.2 Characteristics of the Gammachirp and the Asymmetric Function

To precisely describe the spectral characteristics of the gammachirp and the asymmetric function Equation (4) is rewritten in a form that explicitly uses the relevant parameters, that is,

$$G_C(f;n,b,c,f_r) = G_T(f;n,b,f_r) \cdot H_A(f;b,c,f_r)\,. \tag{7}$$

The asymmetric function uses parameters b, c, and f_r whereas the gammatone uses parameters n, b, and f_r.

Figure 2 shows the amplitude spectra of (a) the gammachirp $G_C(f;n,b,c,f_r)$ and (b) the asymmetric function $H_A(f;b,c,f_r)$ when the values of the chirp parameter c are integers between -3 and 3. Several characteristics are derived from this figure and the equations described above.

(a) Figure 2(a) shows that the filter slope of a gammachirp below the peak frequency is shallower than the slope above it when the parameter c is negative. The situation is the reverse when the parameter c is positive. The filter shape is symmetric when c is zero because the chirp term is removed and the resulting function is identical to the standard gammatone function.

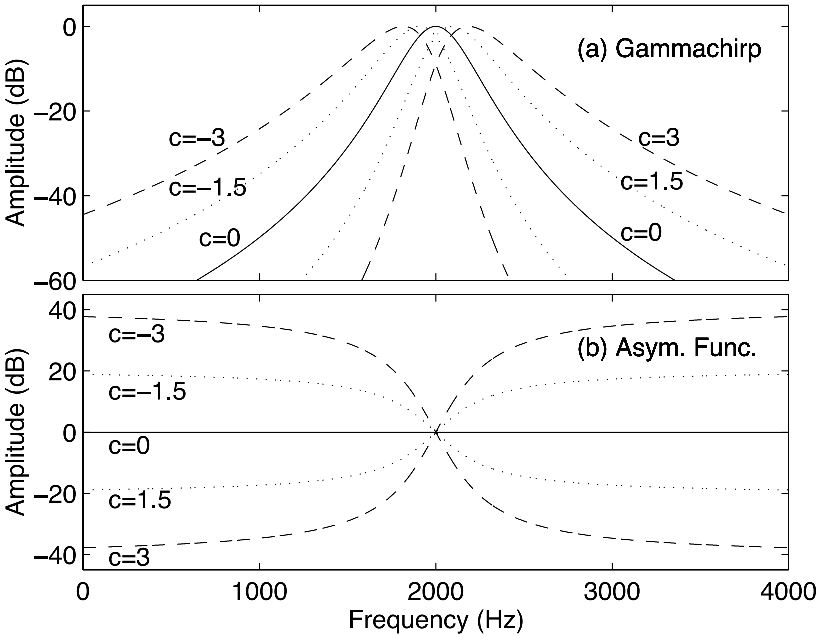


Figure 2 Amplitude spectra of (a) a gammachirp filter $|G_C(f)|$ and (b) an asymmetric compensation filter $|H_A(f)|$ as a function of the chirp parameter c where $n = 4$, $b = 1.019$, and f_r=2000 Hz. The amplitude is normalized to 0 dB at the peak frequency in panel (a) and at f_r in panel (b).

(b) The asymmetric function $H_A(f;b, c, f_r)$ in Figure 2(b) is an all-pass filter when $c = 0$. Using Equation (2),

$$H_A(f;b, 0, f_r) = 1\ . \tag{8}$$

$H_A(f;b, c, f_r)$ is a high-pass filter when $c>0$, and a low-pass filter when $c<0$. The slope and the range of the amplitude increase when the absolute value of c increases. The filter shapes of the gammachirp in Figure 2(a) reflect these characteristics.

(c) $H_A(f;b, c, f_r)$ changes monotonically in frequency. Neither a peak nor a dip ever occurs in this function.

(d) For an arbitrary frequency f_l, the asymmetric function is restricted by

$$|H_A(f_r - f_l;b, c, f_r)| = |H_A(f_r + f_l;b, c, f_r)|^{-1}\ . \tag{9}$$

(e) With Equation (2), the asymmetric function satisfies:

$$H_A(f;b, c, f_r) = H_A(f;b, -c, f_r)^{-1}\ . \tag{10}$$

(f) For arbitrary chirp parameters c_1 and c_2, the asymmetric function is multiplicative with respect to c:

$$H_A(f;b, c_1 + c_2, f_r) = H_A(f;b, c_1, f_r) \cdot H_A(f;b, c_2, f_r)\ . \tag{11}$$

(g) Using Equations (7), (10), and (11),

$$\begin{aligned} G_C(f;n,b,c,f_r) &= G_T(f;n,b,f_r)\cdot H_A(f;b,c,f_r) \\ &= G_T(f;n,b,f_r)\cdot H_A(f;b,c_1+c_2,f_r)\cdot H_A(f;b,-c_2,f_r) \\ &= G_C(f;n,b,c_1+c_2,f_r)\cdot H_A(f;b,-c_2,f_r) \qquad (12) \end{aligned}$$

Equation (12) states that a gammachirp with an arbitrary chirp parameter c_1 is a product of a gammachirp with a different chirp parameter c_1+c_2, and an asymmetric function with the difference between them of c_2. This is because the asymmetric function $H_A(f;b,c,f_r)$ is an exponential function of the parameter c.

These characteristics are necessary conditions for designing the approximation filter in the next section, and they act as a guide for establishing an analysis/synthesis filterbank in Section 3.

2.3 Asymmetric Compensation Filter

As shown by Equation (4), a gammachirp filter can be implemented by cascading a gammatone filter and an asymmetric filter. Since efficient implementations of the gammatone are already known [17][22], this section concentrates on an approximation filter for the asymmetric function described in the previous section. It is necessary to design a filter satisfying the conditions (a) through (g) in the previous section. As a first step, a filter satisfying condition (d) is considered because its characteristic seem the most relevant for filter design purposes.

FIR filters cannot satisfy Equation (9) in the strict sense since they only have zeros and no poles. They can, however, satisfy Equation (9) approximately if a linear-phase FIR filter designed with the Remez algorithm is employed. Unfortunately, this is ineffective since the number of coefficients is comparable to that of the original FIR gammachirp and, moreover, the coefficients seem to require a table indexed with parameters *b, c,* and f_r. The well-known IIR Butterworth and Chebyshev filters cannot satisfy Equation (9) either. Consequently, a new IIR filter has to be designed that is an explicit function of these parameters to satisfy this condition.

IIR filters satisfying Equation (9) have the same numbers of poles and zeros symmetrically located at $f_r+\Delta f$ and $f_r-\Delta f$ for a design frequency Δf. This makes the magnitude, *r*, of the corresponding poles and zeros equal. In addition, these magnitudes must be less than unity for the IIR filters to be stable; this is known as the minimum-phase condition [16]. Since the bandwidth gets narrower when *r* gets closer to unity, *r* is negatively correlated with the bandwidth parameter $b\mathrm{ERB}(f_r)$. Condition (b) implies that Δf is proportional to *c* and is positively correlated with $b\mathrm{ERB}(f_r)$. A cascaded second-order digital filter satisfying these properties is

$$H_C(z) = \prod_{k=1}^{N} H_{Ck}(z) \qquad (13)$$

$$H_{Ck}(z) = \frac{(1-r_k e^{j\varphi_k}z^{-1})(1-r_k e^{-j\varphi_k}z^{-1})}{(1-r_k e^{j\phi_k}z^{-1})(1-r_k e^{-j\phi_k}z^{-1})}, \qquad (14)$$

$$r_k = \exp\{-k\cdot p_1\cdot 2\pi b\mathrm{ERB}(f_r)/f_s\} \qquad (15)$$

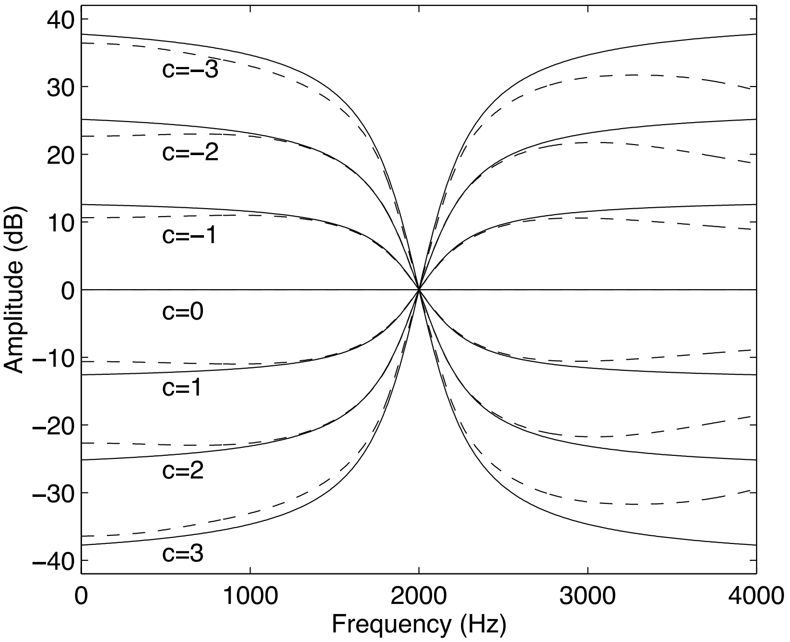


Figure 3 Amplitude spectra of asymmetric functions $|H_A(f)|$ (solid lines) and asymmetric compensation filters $|H_C(f)|$ (dashed lines) where $n = 4$, $b = 1.019$, c is an integer between –3 and 3, and f_r= 2000 Hz. The amplitude is normalized to 0 dB at f_r.

$$\phi_k = 2\pi\{f_r + p_0^{k-1} \cdot p_2 \cdot c \cdot 2\pi b\mathrm{ERB}(f_r)\}/f_s \tag{16}$$

$$\varphi_k = 2\pi\{f_r - p_0^{k-1} \cdot p_2 \cdot c \cdot 2\pi b\mathrm{ERB}(f_r)\}/f_s \tag{17}$$

where p_0, p_1, and p_2 are positive coefficients and f_s is the sampling rate. The reason for cascading filters with gradually located poles and zeros is to satisfy condition (c) approximately. This filter is referred to as an “asymmetric compensation” (AC) filter.

Figure 3 shows the amplitude spectra of this digital filter $H_C(f)$ (dashed lines) and the asymmetric function $H_A(f)$ (solid lines) in Equation (5) as a function of the chirp parameter, c. There were four cascaded filters. The amplitude was normalized at frequency f_r, and the values of p_0, p_1, and p_2 were set, as described in the next section. The dashed lines are very close to the solid lines when the frequency is less than 3000 Hz. Above 3000 Hz the disparity gets larger. However, this does not cause serious errors, because the asymmetric compensation filter is always accompanied by the gammatone filter, which is a band-pass filter.

The results will show that four cascaded second-order filters provide a reasonable fit when the parameter b is equal to or greater than unity and the chirp parameter c is between –3 and 1. In this case, there are a total of 16 poles and zeros. Although it is possible to improve the fit by increasing the number of cascaded filters, a reasonable number can be determined by considering the trade-off between the number of coefficients and the degree of fitting.

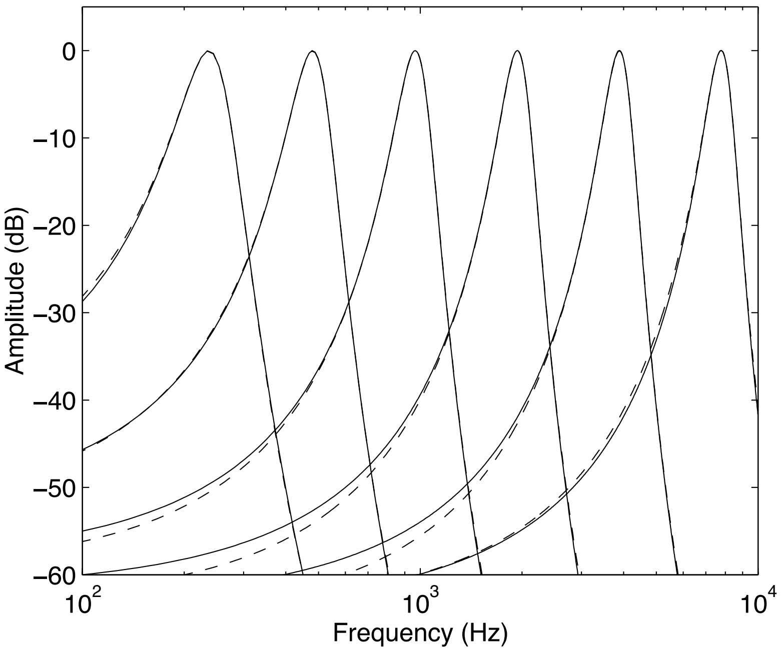


Figure 4 Amplitude spectra of original FIR gammachirp filters $|G_c(f)|$ (solid lines) and asymmetric compensation (AC) gammachirp filters $|G_{CAC}(f)|$ (dashed lines) where n=4, b=1.019, c=-1, and the values for f_r are 250, 500, 1000, 2000, 4000, and 8000 Hz.

2.4 Asymmetric Compensation Gammachirp

The asymmetric compensation filter cascaded with the gammatone filter approximates the gammachirp filter. The amplitude spectrum of this filter is found by replacing $H_A(f)$ with $H_C(f)$ in Equation (6),

$$|G_{CAC}(f)| = |G_T(f)| \cdot |H_C(f)|. \qquad (18)$$

This filter $|G_{CAC}(f)|$ is referred to as an “Asymmetric Compensation-gammachirp” or “AC-gammachirp” filter until the end of Section 2, so as to distinguish it from the original gammachirp defined by Equation (1).

2.4.1 Comparison in the Amplitude Spectrum

Figure 4 shows the amplitude spectra of the gammachirp $|G_C(f)|$ in Equation (5) (solid lines), the AC-gammachirp $|G_{CAC}(f)|$ in Equation (14) (dashed lines), and the gammatone $|G_T(f)|$. The amplitude $|G_{CAC}(f)|$ has been normalized properly to improve the fit. The frequency for normalizing the amplitude of each second-order filter is closely related to the peak shift in Equation (6) and is set with a coefficient, p_s, for the k-th filter,

$$f = f_r + k \cdot p_3 \cdot c \cdot b\mathrm{ERB}(f_r)/n. \qquad (19)$$

The coefficients p_0, p_1, p_2, and p_3 are set heuristically as

$$p_0 = 2, \tag{20}$$

$$p_1 = 1.35 - 0.19\ |c|, \tag{21}$$

$$p_2 = 0.29 - 0.0040\ |c|, \tag{22}$$

$$p_3 = 0.23 + 0.0072\ |c|. \tag{23}$$

The root-mean-squared (rms) error between the original gammachirp filter and the AC-gammachirp filter is less than 0.41 in Figure 4 over the range where $|G_C(f)| > -50\,\text{dB}$. The average rms error is only 0.63 dB for 90 sets of parameter combinations $\{n = 4$; $b = 1.0$, 1.35, and 1.7; $c = 1, 0, -1, -2$, and -3; $f_r = 250, 500, 1000, 2000, 4000$, and 8000 (Hz)$\}$. The rms error exceeds 2 dB only for three sets when $f_r = 8000$ Hz and $c = -3$.

The fit improved only slightly when the coefficients in Equations (21), (22), and (23) were optimized using an iterative least squared-error method. It is possible to improve the fit by changing the locations of the poles and zeros defined in Equations (15), (16), and (17), but this is beyond the scope of this paper.

2.4.2 Comparison of the Impulse Response and the Phase Spectrum

Figure 5(a) shows an example of the impulse response of the gammachirp defined in Equation (1) (solid line) and the AC-gammachirp obtained from Equation (18) (dashed line). The difference in the impulse responses between the original gammachirp and the AC-gammachirp is about –50 dB in rms amplitude and is therefore almost negligible. Their phase spectra, shown in Figure 5(b), are very close to each other. Therefore, the AC-gammachirp provides an excellent approximation to the original gammachirp in terms of its phase characteristics, i.e.,

$$G_C(f) \cong G_{CAC}(f) \ = \ G_T(f) \cdot H_C(f) \tag{24}$$

and also in the time domain,

$$g_c(t) \cong g_{cAC}(t) \ = \ g_T(t) * h_c(t), \tag{25}$$

where * denotes convolution.

2.4.3 Similarity to the Asymmetric Function

The asymmetric compensation filter, $H_C(z)$, defined in Equations (13) and (14) can strictly satisfy Equations (8), (9), and (10) and condition (b), and approximately satisfy Equation (11) and condition (c). The reasons are as follows. For conditions (b), (c) and Equation (11), the correspondence is obvious from Figure 2. For Equation (8), when c is 0, the locations of the poles and zeros of Equations (16) and (17) are the same, and then Equations (13) and (14) become unity. For Equation (9), since ERB (f_r) is a linear function of f_r, changing f_r+f_l to f_r–f_l simply replaces the poles and zeros in Equations (16) and (17). For Equation (10), changing the sign of c replaces the poles and zeros in Equations (16) and (17) and it is possible to derive a stable inverse filter since the asymmetric compensation filter

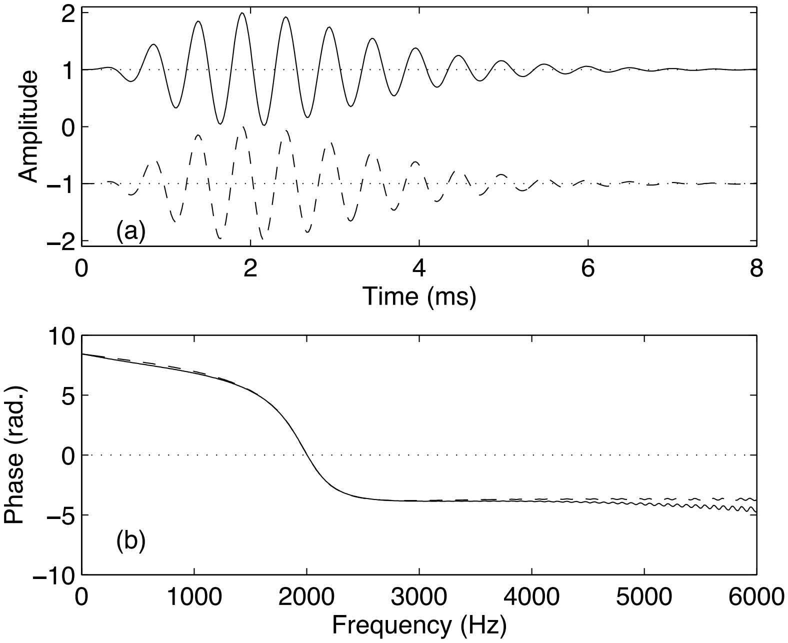


Figure 5 (a) Impulse responses and (b) phase spectra of an FIR gammachirp filter (Equation (1)) (solid lines) and an asymmetric compensation (AC) gammachirp filter (dashed lines). The parameters are $n = 4$, $b = 1.019$, c=-1, and f_{r}. = 2000 Hz.

satisfies the minimum phase condition. The inverse filter is always stable even if the parameter values are time varying. Accordingly, it is possible to cancel the forward filter with the inverse filter. Then, the total response of the combination is a unit impulse. This feature leads to an analysis/synthesis filterbank (described in Section 3.3).

Since the IIR asymmetric compensation filter has few coefficients, fast level-dependent, adaptive auditory filtering can be performed by a combination of the compensation filter with a fast implementation of the gammatone [17][22].

3. Gammachirp Filterbank

This section describes an adaptive, analysis/synthesis gammachirp filterbank. Since the auditory filter shape is level-dependent [5][9][14], it is necessary to estimate the sound pressure level of incoming signals. Section 3.1 shows an example of the analysis filterbank with a level estimation mechanism. Section 3.2 shows another type of filterbank structure using physiological constraints. Although no specific structure or parameter set has been determined yet, these examples are sufficient for presenting the most important issue in this paper. Section 3.3 shows the general structure of an adaptive, analysis/synthesis gammachirp filterbank. This chapter shows that sound resynthesis is always possible independent of the method of parameter control. The analysis/synthesis errors are shown to be time-invariant and small enough to avoid perceptual distortions, even when listening to synthetic sounds.

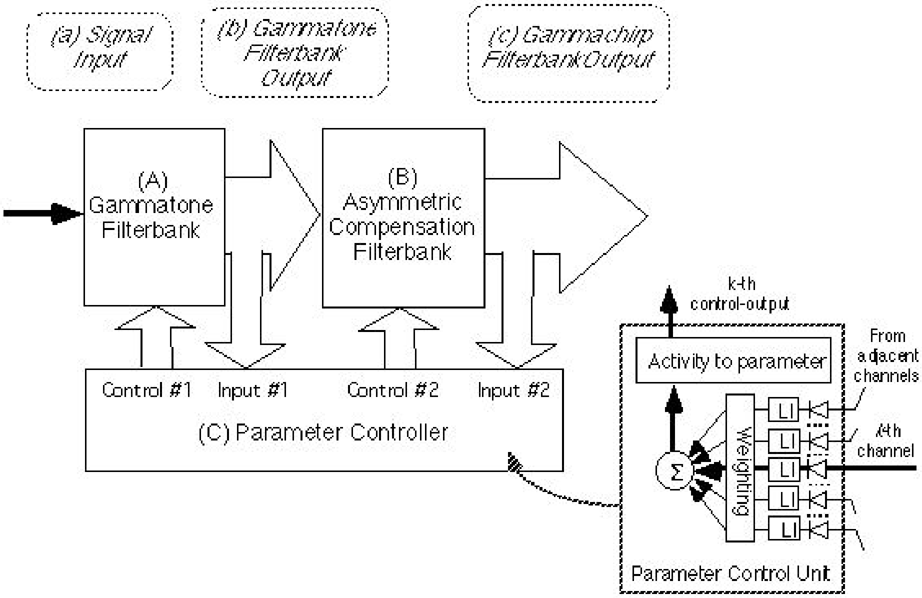


Figure 6 Block diagram of a level-dependent gammachirp filterbank.

3.1 Implementation with level estimation

Figure 6 shows an example of a gammachirp filterbank that consists of a gammatone filterbank, a bank of asymmetric compensation filters, and a parameter controller. The sound pressure level of incoming signals is estimated in the parameter controller using the output of the gammatone filterbank and the asymmetric compensation filterbank. The parameter controller is a bank of parameter control units as shown in the right-bottom block. When considering the k-th channel, the input signal to this block is first rectified, and is then put into a leaky integrator (LI) for smoothing. Any value for the time constant is possible in the following simulation as long as the feedback system is stable; 30 ms was used for the error estimation in Section 3.3. A weighting function (i.e., a Hamming window of 3 ERB width across the filter channels), is applied to the LI output of the k-th and adjacent channels, which are summed together to obtain the activity, a_{ak}, for the k-th channel. The estimated sound pressure level, P_s in decibels, is calculated using a straightforward equation,

$$P_s = 20\log(q \cdot a_{ak}) \tag{26}$$

where q is a constant. The estimated sound pressure level controls the parameters of the gammachirp filterbank.

It has been demonstrated that the constant, q, can be determined using psychoacoustic masking data [10][11]. It does not, however, sufficiently describe the procedure and the results in this chapter since they largely depend on the filterbank structure and the parameter controller. In those simulations, however, the individual filters were all set on the basis of the auditory filter shape at a probe frequency of 2000 Hz [9]. Obviously, it is necessary to use the outputs of several adjacent auditory filters. The formulation with the leaky integrator and

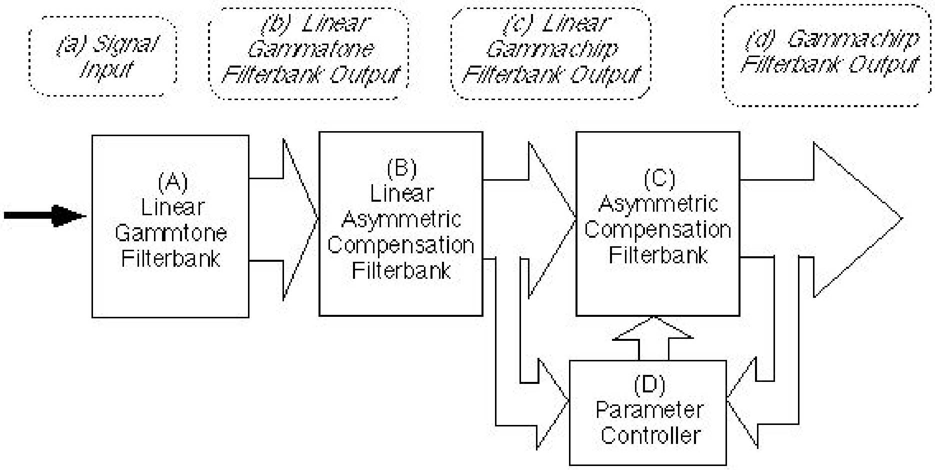

(a) Signal Input
(b) Linear Gammatone Filterbank Output
(c) Linear Gammachirp Filterbank Output
(d) Gammachirp Filterbank Output
(A) Linear Gammtone Filterbank
(B) Linear Asymmetric Compensation Filterbank
(C) Asymmetric Compensation Filterbank
(D) Parameter Controller

Figure 7 Block diagram of a gammachirp filterbank based on physiological constraints.

Equation (26) is restricted to an initial approximation of the parameter control since it does not include fast compression [21] and the compression function is physiologically realistic [4].

Instead of determining the parameter values for this model, we use a basic structure to establish a synthesis procedure (described in Section 3.3) that is independent of the method of parameter control. For the filterbank in Figure 6, it is possible to perform signal resynthesis, provided only the chirp parameter, *c,* varies with the level in the level-dependent auditory filters. Then, control #1 (Figure 6 (c)) is unnecessary because the gammatone filter is not a function of the chirp parameter *c*. The next section shows another filterbank structure based on physiological observations.

3.2 Another filterbank structure

Let us introduce physiological knowledge into the filterbank structure. When the sound pressure level is sufficiently high, the cochlear filter has a broad bandwidth and behaves like a passive and linear filter. As the signal level decreases, the filter gain increases and the bandwidth becomes narrower because of the active processes [19]. This suggests a physiologically plausible auditory filter is a combination of a linear, broadband filter and a nonlinear, level-dependent filter that sharpens the filter shape. Recent observations have shown that the frequency modulation or “glide” persists even post-mortem or after high sound pressure levels [20]. Accordingly, the linear filter can be simulated with a broadband gammachirp filter. As shown in Equation (12), a gammachirp filter with an arbitrary chirp parameter *c* can be produced with a combination of another gammachirp filter and an asymmetric function. Therefore, the second filter can be simulated by a level-dependent asymmetric compensation filter as long as the total filter response can be simulated with the gammachirp.

Accordingly, a candidate filterbank structure is proposed in Figure 7. It consists of a linear gammatone filterbank, a linear asymmetric compensation filterbank, and an adaptive asymmetric compensation filterbank controlled by a parameter controller. The output of the linear asymmetric compensation filterbank is equivalent to the output of a linear gammachirp filterbank (c). This output is fed into the asymmetric compensation filterbank to obtain the total output (d). The parameter controller is similar to that described in Section

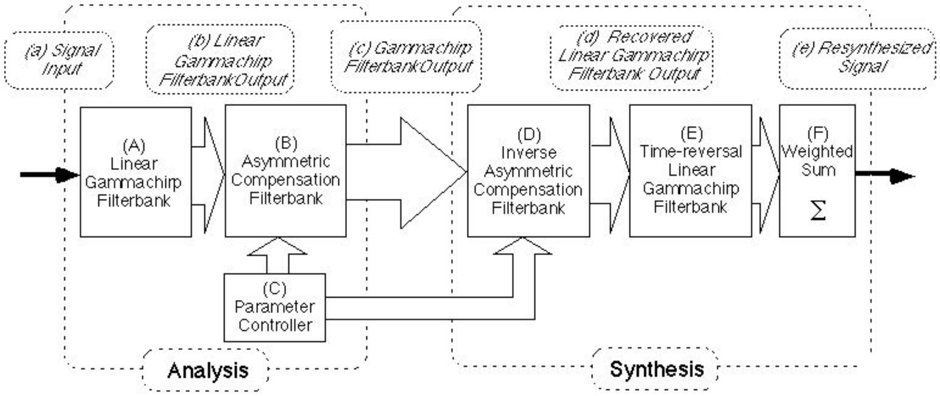

(a) Signal Input
(b) Linear Gammachirp FilterbankOutput
(c) Gammachirp FilterbankOutput
(d) Recovered Linear Gammachirp Filterbank Output
(e) Resynthesized Signal
(A) Linear Gammachirp Filterbank
(B) Asymmetric Compensation Filterbank
(C) Parameter Controller
(D) Inverse Asymmetric Compensation Filterbank
(E) Time-reversal Linear Gammachirp Filterbank
(F) Weighted Sum Σ
Analysis
Synthesis

Figure 8 Block diagram of an adaptive, analysis/synthesis gammachirp filterbank.

3.1. However, before determining the structure and the parameters it is necessary to wait for results fitting the gammachirp to psychoacoustic masking data across the full range of center frequencies. The structure is based on a combination of a linear filterbank with bandpass filters and a non-linear asymmetric compensation filterbank. For signal processing applications, this filterbank structure has a very important feature that has never been accomplished by conventional auditory filterbanks as described in the next section.

3.3 Analysis/synthesis Filterbank

One of the most important features of the gammachirp filterbank is its ability to establish an analysis/synthesis system as shown in Figure 8. Moreover, this feature is valid for any kind of parameter controller. Initially, a signal (a) is filtered by a linear, passive gammachirp filterbank (A). When the chirp parameter c is set to zero for all channels, this is a gammatone filterbank. The output of the linear filterbank (b) is converted into the output of the gammachirp filterbank (c) using a bank of active asymmetric compensation filters (B). Section 2.4 shows how to make a bank of inverse asymmetric compensation filters (D). The output of the adaptive gammachirp filterbank (c) is then converted into a representation (d) that is strictly the same as the output of the linear gammachirp filterbank (b) when using the same parameter set produced by the parameter controller (C) at each moment in time. The filterbank output is then equalized in phase using the time-reversal gammachirp filterbank (E); this is identical to the linear filterbank (A) except that the impulse response of each filter is reversed in time. Finally, the output after this phase equalization is summed with a weighting function to reproduce the signal.

A combination of the linear analysis filterbank (A), the linear synthesis filterbank (E), and the weighted sum (F) is almost equivalent to a linear, wavelet, analysis/synthesis procedure [2]. Since the combination of the asymmetric compensation filterbank (B) and its inverse filterbank (D) produces unit impulses for all channels, the error between the original and synthetic signals is strictly determined by this linear analysis/synthesis filterbank.

Figure 9 shows an example of analysis/synthesis frequency characteristics for an adaptive gammachirp filterbank with equally spaced filters for ERB rates between 100 and 6000 Hz using a gammatone filterbank in (A) and (E) (i.e., the gammachirp filterbank when $c = 0$ for all channels). Figure 9(b) shows the same graph with a magnified ordinate scale. The maximum error is less than 0.01 dB with 100 channels and is only about 0.03 dB even with

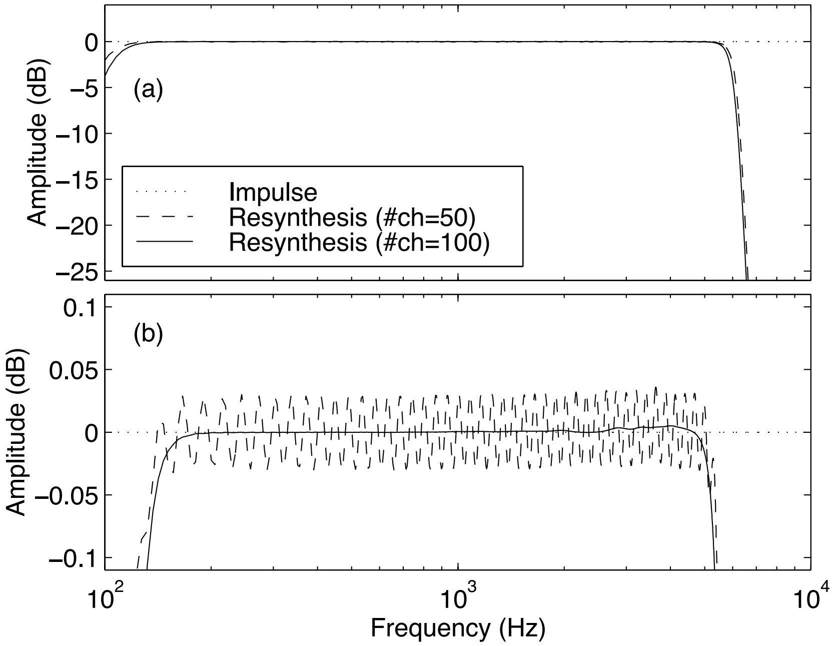


Figure 9 Frequency responses of the analysis/synthesis gammachirp filterbank shown in Figure 8 when the frequency range of the filterbank is between 100 and 6000 Hz and the number of channels is 50 (dashed lines) or 100 (solid lines). Panel (b) is the magnified ordinate of panel (a).

50 channels. It appears that about 100 channels are sufficient to minimize the errors. Moreover, the errors are completely independent of parameter control. Consequently, the gammachirp filterbank is able to perform signal resynthesis without producing any undesirable distortion.

The discussion above guarantees the minimum distortion of the analysis/synthesis filterbank system. This filterbank is applicable to various applications when inserting a modification block between the asymmetric compensation filterbank (B) and its inverse filterbank (C). For example, it is possible to construct a noise-suppression filterbank that does not produce any musical noise (which would be perceptually undesirable) [12].

4. Summary

This paper presents an adaptive, analysis/synthesis auditory filterbank using the gammachirp. Initially, the gammachirp function is analyzed to find characteristics for effective digital filter simulation. The gammachirp filter is shown to be well approximated by the combination of a gammatone filter and an IIR asymmetric compensation filter. The new implementation reduces the computational cost for adaptive filtering because both filters can be implemented with only a few filter coefficients. The inverse filter of the asymmetric compensation filter is shown to be stable. Then two examples of gammachirp filterbanks are presented, each is a combination of a linear gammachirp filterbank and a bank of adaptive, nonlinear asymmetric compensation filters, controlled by the signal-level estimation mechanism. A synthesis procedure for such analysis filterbanks is proposed to accomplish signal resynthesis with a guaranteed precision and no undesirable distortion. This feature has never been accomplished with conventional auditory filterbanks. The adaptive, analysis/synthesis gammachirp filterbank is usable in various signal processing applications requiring the modeling of human auditory filtering.

Acknowledgements

The authors wish to thank Roy D. Patterson of Cambridge Univ., for his continuous advice, Minoru Tsuzaki and Hani Yehia of ATR-HIP and Malcolm Slaney of IBM for their valuable comments. A part of this work was performed while the second author was a visiting student at ATR-HIP. The authors also wish to thank Masato Akagi of JAIST, and Yoh'ichi Tohkura, Hideki Kawahara, and Shigeru Katagiri of ATR-HIP for the arrangements. This work was partially supported by CREST (Core Research for Evolutional Science and Technology) of the Japan Science and Technology Corporation (JST).

References

[1] Cohen, L. “The scale transform.” *IEEE Trans. Signal Processing*, 41: 3275–3292, 1993.

[2] Combes, J. M., Grossmann, A. and Tchamitchian, Ph. (eds.). *Wavelets*. Berlin: Springer-Verlag, 1989.

[3] de Boer, E. and Nuttall, A. L. “The mechanical waveform of the basilar membrane. I. Frequency modulations (“glides”) in impulse responses and cross-correlation functions.” *J. Acoust. Soc. Am.*, 101: 3583–3592, 1997.

[4] Giguère, C. and Woodland, P. C. “A computational model of the auditory periphery for speech and hearing research. I. Ascending path.” *J. Acoust. Soc. Am.*, 95: 331–342, 1994.

[5] Glasberg, B. R. and Moore, B. C. J. “Derivation of auditory filter shapes from notched-noise data.” *Hearing Res.*, 47: 103–138, 1990.

[6] Irino, T. and Kawahara, H. “Signal reconstruction from modified auditory wavelet transform.” *IEEE Trans. Signal Processing*, 41: 3549–3554, 1993.

[7] Irino, T. “An optimal auditory filter.” In *IEEE Signal Processing Society Workshop Applic. Sig. Proc. Audio Acoust.*, 1995.

[8] Irino, T. “A ‘gammachirp’ function as an optimal auditory filter with the Mellin transform.” *IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP-98)*,pp. 981–984, 1996.

[9] Irino, T. and Patterson, R.D. “A time-domain, level-dependent auditory filter: The gammachirp.” *J. Acoust. Soc. Am.*, 101: 412–419, 1997.

[10] Irino, T. and Unoki, M. “An efficient implementation of the gammachirp filter and its filterbank design.” *ATR Technical Report*, TR-H-225, 1997.

[11] Irino, T. and Unoki, M. “A time-varying, analysis/synthesis auditory filterbank using the gammachirp.” *IEEE Int. Conf. Acoust., Speech Sig. Proc. (ICASSP-98)*, 1998.

[12] Irino, T. “Noise suppression using a time-varying, analysis/synthesis gammachirp filterbank.” *IEEE Int. Conf. Acoust., Speech Sig. Proc. (ICASSP-99)*,1999.

[13] Lyon, R. F. “The all-pole gammatone filter and auditory models.” In *Forum Acusticum '96*, Antwerp, 1996.

[14] Lutfi, R. A. and Patterson, R. D. “On the growth of masking asymmetry with stimulus intensity.” *J. Acoust. Soc. Am.* 76, 739–745, 1984.

[15] Møller, A. R. and Nilsson, H. G. “Inner ear impulse response and basilar membrane modelling.” *Acustica, 41*: 258–262, 1979.

[16] Oppenheim, A. V. and Schafer, R. W. *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[17] Patterson, R. D., Allerhand, M. and Giguère, C. “Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform.” *J. Acoust. Soc. Am.*, 98: 1890–1894, 1995.

[18] Pflueger, M., Hoeldrich, R. and Reidler, W. ”Nonlinear all-pole and one-zero gammatone filters.” *Acta Acoustica*, 84: 513–519, 1998.

[19] Pickles, J. O. *An Introduction to the Physiology of Hearing*. London: Academic Press, 1988.

[20] Recio, A. R., Rich, N. C., Narayan, S. S. and Ruggero, M. A. ”Basilar-membrane response to clicks at the base of the chinchilla cochlea.” *J. Acoust. Soc. Am.*, 103: 1972–1989, 1998.

[21] Robles, L., Rhode, W. S. and Geisler, C. D. "Transient response of the basilar membrane measured in squirrel monkeys using the Moessbauer effect." *J. Acoust. Soc. Am.*, 59, 926–939, 1976.

[22] Slaney, M. "An efficient implementation of the Patterson–Holdsworth auditory filter bank." *Apple Computer Technical Report #35*, 1993.

[23] Slaney, M. "Pattern playback from 1950 to 1995." *IEEE Conf. Syst. Man., Cybernetics,* Vancouver, Canada, 1995.

[24] Yang, X, Wang, K. and Shamma, S. A. "Auditory representations of acoustic signals." *IEEE Trans. Information Theory*, 38: 824–839, 1992.

Computational Models of Auditory Function
Greenberg and M. Slaney (eds.)
IOS Press, 2001

ADVANCES IN COMPUTATIONAL MODELING OF COCHLEAR IMPLANT PHYSIOLOGY AND PERCEPTION

I. C. Bruce[1,2] M. W. White[3], L. S. Irlicht[2,4], S. J. O'Leary[2] and G. M. Clark[2]

[1]*Department of Biomedical Engineering, Johns Hopkins University*
505 Traylor Building, 720 Rutland Ave, Baltimore MD 21205, USA

[2]*Department of Otolaryngology, The University of Melbourne*
384-388 Albert Street, East Melbourne VIC 3002, Australia

[3]*Department of Electrical and Computer Engineering*
North Carolina State University, Raleigh NC 27695, USA

[4]*Rothschild Australia Asset Management, Level 10,*
One Collins Street, Melbourne VIC 3000, Australia

1. Introduction

Models of cochlear implant physiology and perception have historically utilized deterministic descriptions of auditory-nerve (AN) responses to electrical stimulation, which ignore stochastic activity present in the response. Physiological models of AN responses have been developed that do incorporate stochastic activity [8][13][14][27][38][39], but the consequences of stochastic activity for the perception of cochlear implant stimulation have not been investigated until recently [3].

Such an investigation is prompted by inaccuracies in predicting cochlear implant perception by deterministic models. For example, studies of single-fiber responses, where only an arbitrary deterministic measure of threshold is recorded, do not accurately predict perceptual threshold versus phase duration (strength-duration) curves for sinusoidal stimulation [24] or for pulsatile stimulation [25][26]. Furthermore, strength-duration curves of cochlear implant users are not well predicted by deterministic Hodgkin–Huxley type models [25] [30].

However, the complexity of previous stochastic physiological models has made the computation of responses for large numbers of fibers both laborious and time-consuming. Furthermore, the parameters of these models are often not easily matched to the fiber characteristics of the auditory nerve in humans or other mammals. This has prompted us to develop a simpler and more computationally efficient model of electrical stimulation of the auditory nerve [1][2][4] which is capable of direct and rapid prediction of perceptual data [3].

2. Computational Modeling of Cochlear Implant Physiology

In [1] and [4] we have described a model of the AN response to electrical stimulation, following the conceptual approach used in [35], [38] and [39]. This model can be represented by the electrical circuit diagram shown in Figure 1. Based on the Hill threshold model [12], our model includes a number of significant components of action potential generation, including membrane noise, as recorded by Verveen and colleagues [35], which has a Gaussian amplitude distribution and a 1/f frequency spectrum. Threshold models are much simpler conceptually and are more computationally efficient than Hodgkin–Huxley models. They

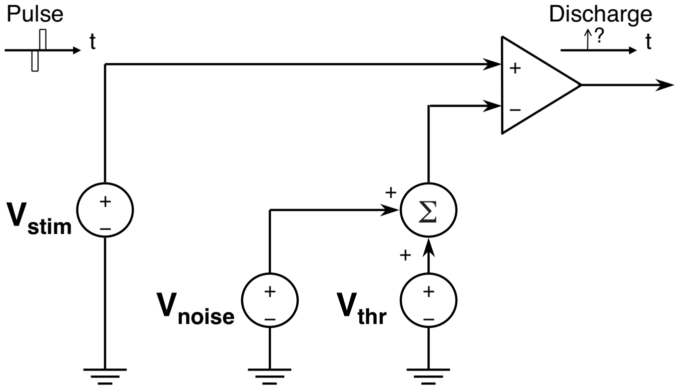


Figure 1 Stochastic model of single-pulse response. Reprinted from Fig. 2 of [4] © 1999 IEEE.

have also been shown to provide a good approximation to more complex models [16]. Additionally, our model can be fitted easily to the statistics of AN parameters collected from anatomical and physiological studies.

We are able to derive an analytical description of this model, because it is in effect a Bernoulli process, where a discharge in response to a pulse is considered to have a value of 1 and no discharge has a value of 0. The probability of discharge, p(n), in response to a single pulse approximated by the equation [1] [4]

$$p(n) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{V_{\mathrm{stim}}(n) - V_{\mathrm{thr}}(n)}{\sqrt{2}\sigma}\right)\right) \tag{1}$$

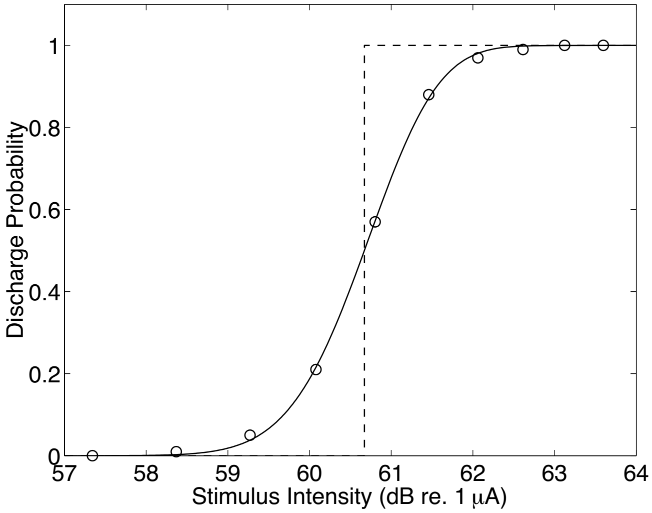


Figure 2 Stochastic (solid line) and deterministic (dashed line) model fits to discharge probability data (circles) from Neuron 2-22 of [15].

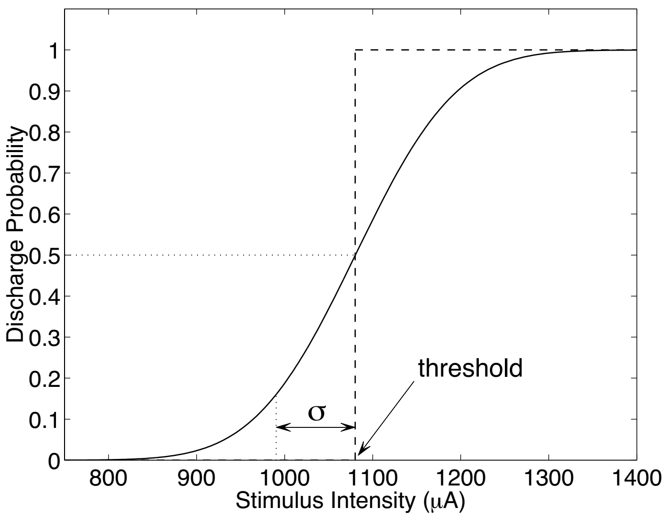


Figure 3 Stochastic (solid line) and deterministic (dashed line) model I/O functions showing how threshold and the standard deviation of the Gaussian noise, σ, are defined to determine RS (= σ/threshold).

where σ is the standard deviation of the noise potential, $V_{noise}(n)$. The equivalent deterministic model can be simulated by setting the noise to zero (σ = 0), producing a step function at $V_{thr}(n)$ [1][4].

Deterministic and stochastic model fits to physiological data from a single cat AN fiber are plotted in Figure 2. The stochastic model provides a much better fit to the data ($r^2 = 1.0$) than does the deterministic model ($r^2 = 0.92$).

Following Verveen et al.'s convention, we characterize the input/output (I/O) functions by defining threshold as the intensity corresponding to a discharge probability of 0.5 and

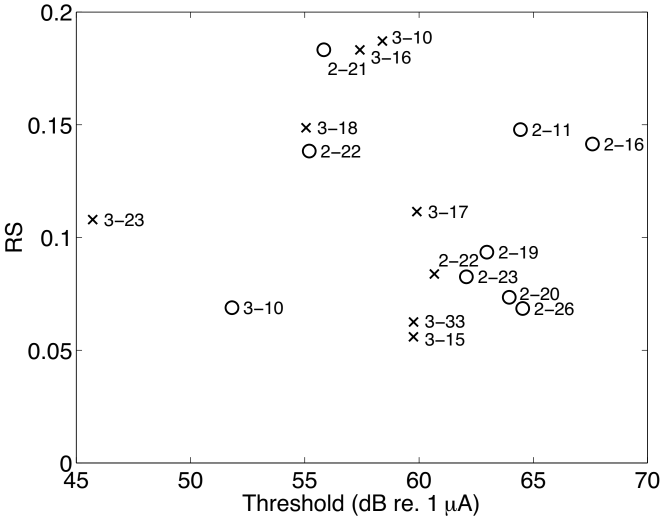


Figure 4 Relative spread versus threshold for neurons from [15] as labeled, in response to a single biphasic pulse of duration 200 μs/phase (x) or 400 μs/phase (o). Reprinted from Fig. 5 of [4] © 1999 IEEE.

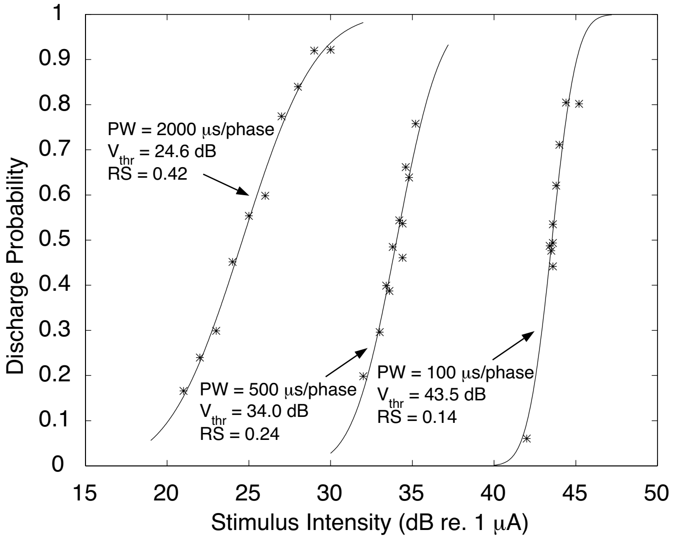


Figure 5 Discharge probability vs. stimulus intensity for single symmetric biphasic anodic/cathodic pulses of durations 100, 500 and 2,000 μs/phase, from Cat 76: Unit 2 in the Dynes data set [8]. Plotted are individual measures (asterisks) and stochastic model fits (solid lines). Reprinted from Fig. 6 of [4] © 1999 IEEE.

Relative Spread (RS) as the standard deviation of the Gaussian noise divided by threshold. The deterministic model is characterized by threshold alone (σ=0), as illustrated in Figure 3.

In order to model the response of a population of AN fibers we need to determine the model parameters for each neuron in the population, as well as the intensity of the excitatory current at the initial site of action potential generation in each neuron.

Figure 4 shows a plot of RS versus threshold for 15 neurons from the Javel et al. data set for a 200-μs/phase (x) or a 400-μs/phase (o) biphasic pulse. It can be seen that both thresholds and RSs cover a broad range of values. Combining these data with similar published data [19][34], we are able to estimate the distribution of I/O function parameters (threshold and RS) for a local population of fibers in the AN. In our “total AN” (large-scale population) model we simulate this distribution of I/O functions [4].

Although higher rate pulsatile stimuli are typically used in modern cochlear implants, necessitating short phase durations (i.e., pulse widths), there is a relatively large body of psychophysical data available in which long phase duration stimuli were used. Furthermore, these psychophysical data show large discrepancies with deterministic model predictions at long phase durations [24][25][26][30]. However, the method used for suppressing the stimulus artifact in the Javel et al. experiments did not allow for much data collection at phase durations longer than 400 μs/phase and none longer 600 μs/phase. Thus we also conducted a *post hoc* analysis of previously unpublished data collected by Dynes from single AN fibers of cats [8], where a pair of closely spaced micropipettes were used in differential-like recording to produce a high signal-to-artifact ratio even at long phase durations.

In Figure 5 discharge probability is plotted versus stimulus intensity curves for three different phase durations from Cat 76: Unit 2 in the Dynes data set. As the phase duration increases, the slope of the curve becomes shallower, indicating a greater dynamic range. Computing the RSs of these curves shows that RS increases as the phase duration of the

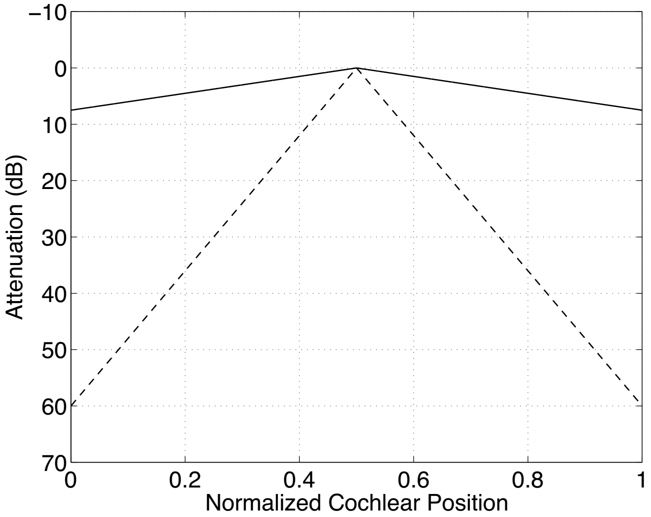


Figure 6 Attenuation of the stimulus across the cochlea for monopolar (solid line) and bipolar (dashed line) electrode configurations. Reprinted from Fig. 9 of [4] © 1999 IEEE.

anodic/cathodic biphasic stimulus increases. This effect is seen in all fibers of this data set. For every fiber, the RS increases as the duration per phase of the stimulus increases. To incorporate this behavior in our model, we fit discharge-probability functions (Eq. 1) to the complete data set and calculated the mean threshold and RS at phase durations of 100, 500, 2000 and 5000 μs/phase. We then fit appropriate functions to threshold and RS versus phase duration plots. These functions are used in the model to interpolate values of threshold and RS at phase durations other than those used in the Dynes experiments [8].

The two electrode configurations that we investigate in these studies are commonly known as monopolar and bipolar. In the case of monopolar stimulation, the active electrode is one of the electrodes on the array within the cochlea and the return electrode is an electrode external to the cochlea. In the case of bipolar stimulation both the active electrode and the return electrode are on the electrode array within the cochlea.

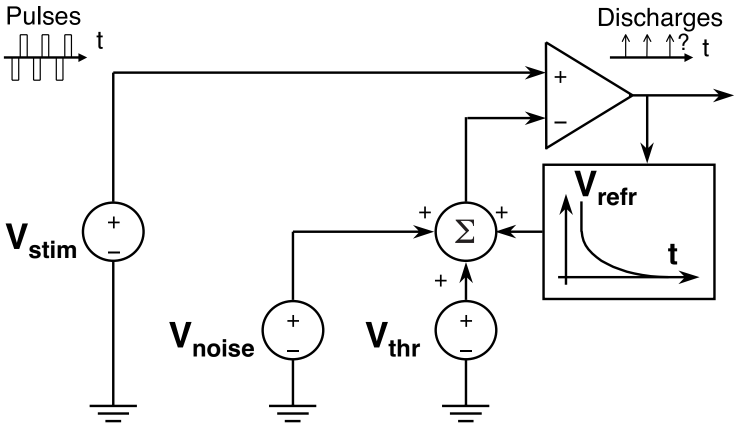


Figure 7 Stochastic model of pulse-train response.

Following [21], we approximate the electrode array by a point source of current at the active electrode and the AN tissue by a homogeneous resistive medium consisting of a uniform density of single AN fibers. To calculate the stimulus intensity at each AN fiber, we assume that the stimulus is attenuated at the rate of 0.5 dB/mm for monopolar stimulation [18] and 4 dB/mm for bipolar stimulation—the latter value is appropriate for both radial-bipolar pairs [18] and closely spaced longitudinal-bipolar pairs [20]. Modeling an electrode placed 15 mm inside a cochlea 30-mm long produces attenuation curves as shown in Figure 6.

In [1] and [2] we go on to extend this model to describe responses to pulse-train stimuli, by introducing a phenomenological refractory mechanism. To the single-pulse model of [1] and [4] we add a refractory potential as shown in Figure 7. Following an action potential, the threshold with which the stimulus potential is compared is raised over the refractory period by some chosen function, typically an exponential [1][2][8][22]. We also derive analytical expressions to approximate the pulse-train model, which although more complex than the single-pulse model, are still computationally efficient and can be fitted easily to the statistics of AN parameters collected from physiological studies.

3. Computational Modeling of Cochlear Implant Perception

In [3] we investigate whether inaccuracies in predictions of loudness perception could be due to ignoring the stochastic response of the AN to electrical stimulation. In order to avoid the complication of inter-pulse interactions and to enable the use of the simpler and computationally faster single-pulse model as shown in Figure 1, we restrict our investigation to single biphasic pulses and low-rate (< 200 pulses per second) pulse trains. We derive a model of loudness based on the single-pulse model of neural excitation developed in [1] and [4] and compare the deterministic and stochastic model predictions. We develop the psychophysical (perceptual) section of the model in such a way that signal detection theory can be applied to predict directly how behavioral threshold, dynamic range and intensity difference limen change with stimulus parameters and nerve survival. The resulting model is shown in Figure 8.

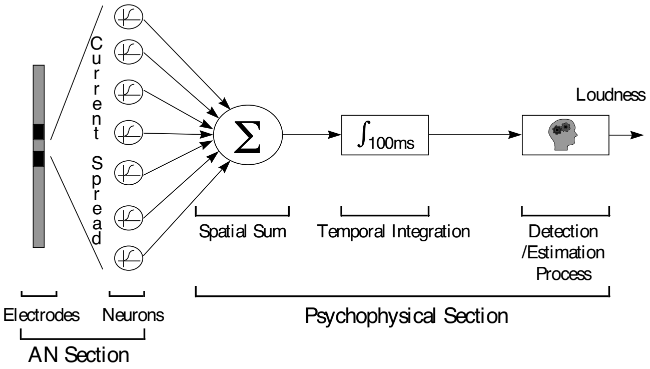


Figure 8 Composite computational physiological and perceptual model. Reprinted from Fig. 1 of [3] © 1999 IEEE.

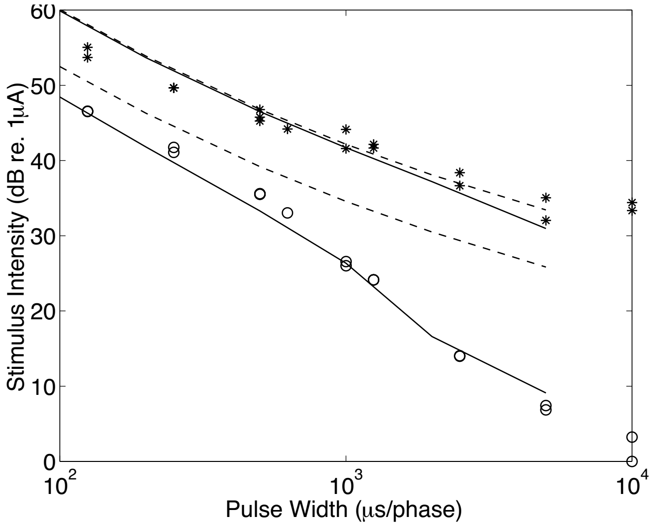


Figure 9 Perceptual data for uncomfortable loudness (*) and threshold (o) are plotted against phase duration, along with the deterministic (dashed lines) and stochastic (solid lines) model predictions.

In all the cases examined in this set of studies, the stochastic model predicts perceptual data better than does the deterministic model. For example, plotted in Figure 9 are perceptual data from a cochlear implant user showing uncomfortable loudness and threshold versus phase duration. While the stochastic and deterministic models predict similar uncomfortable loudness levels, the deterministic model overestimates the threshold data, particularly for longer pulse durations. In contrast, the stochastic model, consistent with the physiological data, predicts (i) absolute values of threshold that are significantly lower than those predicted by the deterministic model, and (ii) slopes that begin to steepen with phase durations greater than 500 μs/phase and slopes that are steeper than –6 dB/doubling in the region from 1000 to 2000 μs/phase. This is more than would be expected if it were assumed that threshold corresponds to a certain level of charge delivered by an implant.

Our study [3] also shows that the stochastic model better predicts perceptual data for:

- *threshold* versus *phase duration* as a function of *electrode configuration* (bipolar or monopolar),
- the ratio of *bipolar dynamic range* versus *monopolar dynamic range*,
- *threshold* versus *number of pulses* (temporal-integration), and
- *intensity difference limen* as a function of intensity (Weber functions).

The physiological model is based on data from the cat AN, but the resulting perceptual model gives good qualitative predictions of data from implanted humans, monkeys, guinea pigs and cats. This suggests that stochastic activity in the AN is perceptually significant across a wide range of measures of loudness perception and regardless of the species, although anatomical, physiological and cognitive differences may have small quantitative effects.

Ferguson et al. [9] have implemented a model similar to ours and have compared its predictions of threshold as a function of pulse duration for monopolar and bipolar stimulation modes with experimental data. Analysis of data from three species indicated that the variance of perceptual thresholds is also a function of phase duration, and that these results are corroborated by the predictions of the stochastic version of the model. These results are not predicted by the deterministic model, indicating that the importance of stochastic activity in the AN extends beyond the perceptual data investigated in our own studies.

4. Future Directions

In these studies we derive a model of loudness in cochlear implants users based on physiological data and use this model to investigate a number of different perceptual phenomena. In all the cases examined so far, the model predicts the perceptual performance of cochlear implant users significantly better when stochastic activity is included in the neural section of the model.

However, extensions or revisions of this AN model may further improve predictions and our understanding of the functional significance of the physiology—specific suggestions follow.

The neural section of our model is derived from physiological data collected in cats. Further physiological data may be collected from humans using cochlear implant telemetry and non-invasive electrophysiology, which should prove useful in refining our simple model of current spread and neural response. A model of current spread in the human cochlea constructed from human cochlear sections [6] may also help to this end.

Another extension to the model would be to allow for other sources of noise. For instance, the survival of inner hair cells in some subjects could result in some residual synapse-driven spontaneous activity in the AN. This would affect the amount of noise present in the total AN response. Other sources of noise may also be present in more central sections of the auditory pathways. The effects of both of these potential noise sources can be included in our perceptual model if their behavior is known.

By changing parameters of the model to reduce the amount of stochastic activity we may also account for such data which lie somewhere between the deterministic model and the stochastic model predictions. For instance, particularly focused current fields or extremely low neural survival may cause higher probabilities of firing at stimulus intensities within the behavioral operating range. Because neural responses at high discharge probabilities exhibit relatively little variability, stochastic and deterministic model predictions are similar under such conditions.

The physiological data on which our model is based are from acutely implanted animals. This model does not take into account the effects which etiology, prolonged deafness and implantation have on the response of AN fibers to electrical stimulation [31] [37]. An extension to these studies could be to model the effects of various etiologies on single-fiber I/O functions and current spread.

Only responses to stimulation from a single electrode have been investigated in these studies. In order to model responses to stimulation from multiple electrodes, even at moderate pulse rates, refractory effects should be incorporated [1][2] when the electrodes are stimulating overlapping populations of fibers. Also, loudness summation effects may need to be considered when the neural populations excited do not overlap [17][29][33][40].

In these studies we have also limited our investigation to low pulse-rate stimuli. With the pulse-train model developed in [1] and [2] and shown in Figure 7, we may now have an appropriate tool for extending this investigation to the prediction of perceptual data for moderate stimulation rates (200–1,000 pps). However, to develop the model for high pulse rate (> 1,000 pps) stimulation, neurophysiological data must be collected for a range of discharge probabilities (possibly as low as 0.01 or lower) at such high pulse rates. Physiological data

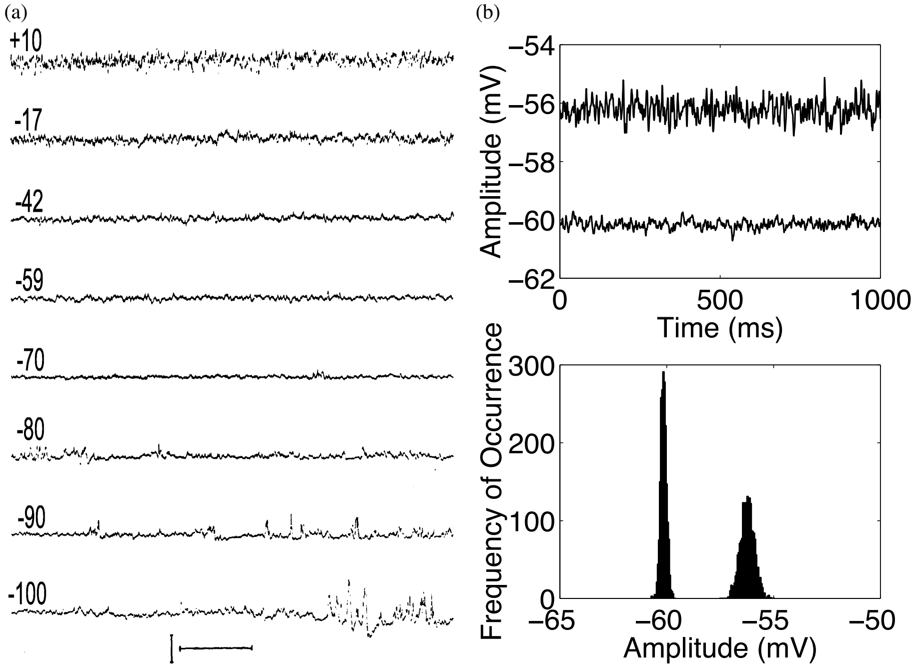


Figure 10 (a) Membrane-potential traces at different levels of membrane potential (given in mV above each trace). Horizontal scale: 1 s; vertical scale: 5 mV. (Reprinted with permission from Fig. 19 of [35] © 1968 IEEE). (b) Model predictions of membrane potential fluctuations (top pane) and their distributions (bottom pane): at resting potential (-60 mV) and at a sustained depolarized potential.

and modeling results [8][19][28] reveal inter-pulse interactions occurring at high pulse rates which can significantly increase or decrease the level of stochastic activity in a fiber.

To further investigate such nonlinearities we are now developing a more computationally efficient stochastic Hodgkin–Huxley type model than those such as Rubinstein's [27][28]. This is achieved by applying Chua's [5] reformulation of the Hodgkin–Huxley model to Fox and Lu's [10] stochastic version of the model. Chua's reformulation permits efficient simulation of complex biological neurons using a standard circuit analysis program such as SPICE [23][32]. Initial simulation results show that such a model can accurately and efficiently predict a number of properties of the random fluctuations in the membrane potential as characterized by Verveen et al. [35]. Plotted in Figure 10 are membrane-potential traces from [35] showing fluctuations in nerve-fiber transmembrane potentials at the nodes of Ranvier and model predictions of these fluctuations. Both Verveen's recordings and the model predictions exhibit a Gaussian amplitude distribution and an increase in membrane noise variance with depolarization. The model also predicts the 1/f frequency spectrum observed by Verveen et al., which tends towards a white (flat) spectrum at higher frequencies.

However, there is some preliminary evidence that the Fox and Lu approximation may become inaccurate when the model neuron is spiking [Jay Rubinstein, pers. comm.]. One possibility is that the approximation deals incorrectly with the noise distribution when the membrane is hyperpolarized or depolarized, which Verveen and Derksen observed to be highly non-Gaussian [36]. If this can be corrected, then with Chua's reformulation we will be able to investigate simply and efficiently how stochastic versions of both single-node models and anatomically correct multi-node models [7][11] predict physiological data for pulsatile cochlear implant stimulation.

5. Summary and Conclusions

A number of aspects of cochlear implant physiology and perception are better predicted by a stochastic model than by the equivalent deterministic model. These results show that loudness perception in implant subjects is highly dependent on the statistics of AN response, not just on some form of absolute threshold. This may imply that loudness perception of acoustic stimuli in normal hearing and hearing-impaired subjects is also dependent on the statistics of AN response, not just on absolute thresholds. High-rate electrical stimulation may produce significant inter-pulse interactions related to changes in levels of membrane noise, which cannot be predicted by deterministic models.

Acknowledgments

The authors would like to thank Scott Dynes and Eric Javel for their contribution of physiological data, and Chris van den Honert and Charlie Finley for their assistance given to Mark White in initial investigations of stochastic activity in the auditory nerve. This work was funded by the Human Communication Research Centre, The University of Melbourne, the Bionic Ear Institute, and the CRC for Cochlear Implants, Speech and Hearing Research.

References

[1] Bruce, I. C., Irlicht, L. S., White, M. W., O'Leary, S. J. and Clark, G. M. "Renewal-process approximation of a stochastic threshold model for electrical neural stimulation." *J. Comput. Neurosci.,* 9(2): 119–132, 2000

[2] Bruce, I. C., Irlicht, L. S., White, M. W., O'Leary, S. J., Dynes, S., Javel, E. and Clark, G. M. "A stochastic model of the electrically stimulated auditory nerve: Pulse-train response." *IEEE Trans. Biomed. Eng.,* 46(6): 630–637, 1999.

[3] Bruce, I. C., White, M. W., Irlicht, L. S., O'Leary, S. J. and Clark, G. M. "The effects of stochastic neural activity in a model predicting intensity perception with cochlear implants: Low-rate stimulation." *IEEE Trans. Biomed. Eng.,* 46(12): 1393–1404, 1999.

[4] Bruce, I. C., White, M. W., Irlicht, L. S., O'Leary, S. J., Dynes, S., Javel, E. and Clark, G. M. "A stochastic model of the electrically stimulated auditory nerve: Single-pulse response." *IEEE Trans. Biomed. Eng.,* 46(6): 617–629, 1999.

[5] Chua, L. O. "Device modeling via basic nonlinear circuit elements." *IEEE Trans. Circuits Syst.*, CAS-27(11): 1014–1044, 1980.

[6] Cohen, L. T., Saunders, E., Busby, P. A. and Clark, G. M. "Place information in cochlear implants: An integrated approach employing psychophysics and modeling." In *Program and Abstracts of 1997 Conference on Implantable Auditory Prostheses,* Asilomar Conference Center, Pacific Grove, California, 1997.

[7] Colombo, J. and Parkins, C. W. "A model of electrical excitation of the mammalian auditory-nerve neuron." *Hear. Res.,* 31: 287–312, 1987.

[8] Dynes, S. *Discharge Characteristics of Auditory Nerve Fibers for Pulsatile Electrical Stimuli*. Ph.D. thesis. Massachusetts Institute of Technology, 1996.

[9] Ferguson, W. D., Smith, D. W., Finley, C. C., Pfingst, B. E. and Collins, L. M. "Prediction of the variance in behavioral thresholds using a stochastic model of electrical stimulation." *ARO Midwinter Meeting Abstracts,* 1998 (see http://www.ee.duke.edu/~wdf/ARO/index.htm).

[10] Fox, R. F. and Lu, Y.-N. "Emergent collective behavior in large numbers of globally coupled independently stochastic ion channels." *Phys. Rev. E,* 49(4): 3421–3431, 1994.

[11] Frijns, J. H. M. *Cochlear Implants — A Modelling Approach*. Ph.D. thesis. Leiden University, The Netherlands, 1995.

[12] Hill, A. V. "Excitation and accommodation in nerve." *Proc. R. Soc. B,* 119: 305–355, 1936.

[13] Hochmair-Desoyer, I. J., Hochmair, E. S., Motz, H. and Rattay, F. "A model for the electrostimulation of the nervus acusticus." *Neurosci.,* 13(2): 553–562, 1984.

[14] Irlicht, L. S. and Clark, G. M. "Control strategies for neurons modeled by self-exciting point processes." *J. Acoust. Soc. Am.,* 100: 3237–3247, 1996.

[15] Javel, E., Tong, Y. C., Shepherd, R. K. and Clark, G. M. "Responses of cat auditory nerve fibers to biphasic electrical current pulses." *Ann. Otol. Rhinol. Laryngol.,* 96 (Suppl. 128): 26–30, 1987.

[16] Kistler, W. M., Gerstner, W. and van Hemmen, J. L. "Reduction of the Hodgkin–Huxley equations to a single-variable threshold model." *Neural Comput.,* 9(5): 1015–1045, 1997.

[17] McKay, C. M., McDermott, H. J. and Clark, G. M. "Loudness summation for two channels of stimulation in cochlear implants: Effects of spatial and temporal separation." *Ann. Otol. Rhinol. Laryngol.,* 104 (Suppl. 166): 230–233, 1995.

[18] Merzenich, M. M. and White, M. W. "Cochlear implants: The interface problem." In *Functional Electrical Stimulation: Applications in Neural Prostheses,* Volume. 3, F. T. Hambrecht and J. B. Reswick (eds.), New York: Marcel Dekker, pp. 321–340, 1977.

[19] Miller, C. A., Abbas, P. J., Robinson, B. K., Rubinstein, J. T. and Matsuoka, A. J. "Electrically evoked single-fiber action potentials from cat: Responses to monopolar, monophasic stimulation." *Hearing Res.,* 130: 197–218, 1999.

[20] O'Leary, S. J., Black, R. C. and Clark, G. M. "Current distributions in the cat cochlea." *Hearing Res.,* 18: 273–281, 1985.

[21] O'Leary, S. J., Clark, G. M. and Tong, Y. C. "Model of discharge rate from auditory nerve fibers responding to electrical stimulation of the cochlea: Identification of cues for current and time-interval coding." *Ann. Otol. Rhinol. Laryngol.,* 104 (Suppl. 166): 121–123, 1995.

[22] Parkins, C. W. "Temporal response patterns of auditory nerve fibers to electrical stimulation in deafened squirrel monkeys." *Hearing Res.,* 41: 137–168, 1989.

[23] Parodi, M. and Storace, M. "On a circuit representation of the Hodgkin and Huxley nerve axon membrane equations." *Int. J. Circuit Theory Appl.,* 25: 115–124, 1997.

[24] Pfingst, B. E. "Comparisons of psychophysical and neurophysiological studies of cochlear implants." *Hear. Res.,* 34: 243–252, 1988.

[25] Pfingst, B. E. "Psychophysical constraints on biophysical/neural models of threshold." In *Cochlear Implants — Models of the Electrically Stimulated Ear*, J. M. Miller and F. A. Spelman (eds.), New York: Springer-Verlag, pp. 161–185, 1990.

[26] Pfingst, B. E., Haan, D. R. D. and Holloway, L. A. "Stimulus features affecting psychophysical detection thresholds for electrical stimulation of the cochlea. I: Phase duration and stimulus duration." *J. Acoust. Soc. Am.,* 90: 1857–1866, 1991.

[27] Rubinstein, J. T. "Threshold fluctuations in an N sodium channel model of the node of Ranvier." *Biophysical Journal,* 68: 779–785, 1995.

[28] Rubinstein, J. T., Wilson, B. S., Finley, C. C. and Abbas, P. J. "Pseudo spontaneous activity: Stochastic independence of auditory nerve fibers with electrical stimulation." *Hearing Res.*, 127(1–2): 108–118, 1999.

[29] Shannon, R. V. "Multichannel electrical stimulation of the auditory nerve in man. II. Channel interaction." *Hearing Res.,* 12: 1–16, 1983.

[30] Shannon, R. V. "A model of threshold for pulsatile electrical stimulation of cochlear implants." *Hearing Res.,* 40: 197–204, 1989.

[31] Shepherd, R. K. and Javel, E. "Electrical stimulation of the auditory nerve: I. Correlation of physiological responses with cochlear status." *Hear. Res.,* 108: 112–144, 1997.

[32] Storace, M., Bove, M., Grattarola, M. and Parodi, M. "Simulations of the behavior of synaptically driven neurons via time-invariant circuit models." *IEEE Trans. Biomed. Eng.,* 44(12): 1282–1287, 1997.

[33] Tong, Y. C. and Clark, G. M. "Loudness summation, masking, and temporal interaction for sensations produced by electric stimulation of two sites in the human cochlea." *J. Acoust. Soc. Am.,* 79: 1958–1966, 1986.

[34] van den Honert, C. and Stypulkowski, P. H. "Single fiber mapping of spatial excitation patterns in the electrically stimulated auditory nerve." *Hearing Res.,* 29: 195–206, 1987.

[35] Verveen, A. A. and Derksen, H. E. "Fluctuation phenomena in nerve membrane." *Proc. IEEE,* 56(6): 906–916, 1968.

[36] Verveen, A. A. and Derksen, H. E. “Amplitude distribution of axon membrane noise voltage.” *Acta Physiol. Pharmacol. Neerl.,* 15: 353–379, 1969.

[37] Viemeister, N. F., Javel, E. and Ganesan, G. K. “Auditory nerve correlates of intensity discrimination for electrical stimuli.” *ARO Midwinter Meeting Abstracts,* 1997.

[38] White, M. W. *Design Considerations of a Prosthesis for the Profoundly Deaf,* Ph.D. thesis, University of California, Berkeley, 1978.

[39] White, M. W., Finley, C. C. and Wilson, B. S. “Electrical stimulation model of the auditory nerve: Stochastic response characteristics.” *Proc. Ninth Ann. Conf. IEEE Eng. Med. Biol. Soc.,* Boston, pp. 1906–1907, 1987.

[40] White, M. W., Merzenich, M. M. and Gardi, J. N. “Multi-channel cochlear implants: Channel interactions and processor design.” *Arch. Otolaryngol.,* 110: 493–501, 1984.

SOUND LOCALIZATION AND BINAURAL MECHANISMS

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

SOUND LOCALIZATION AND BINAURAL MECHANISMS

Jens Blauert

Institute of Communication Acoustics
Ruhr-Universität Bochum
D- 44780 Bochum, Germany

There is no doubt that humans can hear reasonably well with only one ear (monaural hearing). Nevertheless, hearing with two properly functioning ears (binaural hearing) is superior to monaural hearing in many ways. This is due to the fact that there is additional information available to the auditory system when listening through its two ears in contrast to when listening through only one of them. This additional information is encoded in the differences of the input signals to the two ears. The auditory system is capable of decoding part of this information and can make use of it when forming auditory percepts.

The advantage of binaural over monaural listening can be observed with regard to the following auditory functions (among others) [2]:
(1) Separation of different sound sources
(2) Sound localization
(3) Suppression of coloration and reverberance
(4) Suppression of noise when concentrating on a desired signal

As to the separation of sound sources, signals from concurrent sound sources become more distinguishable — at least when the sources are at different locations in space (i). The spatial coincidence of sound sources and auditory events becomes better and localization blur decreases considerably (ii). When reflected sound is present — as in enclosed spaces with reflecting walls — coloration and reverberance are less pronounced perceptually (iii). As sources are better separated and the auditory events more precisely localized, it becomes easier for the listeners to concentrate on the desired signals and to disregard those which they perceive as noise in a given context (iv).

When analyzing binaural hearing in more detail it is useful to make a distinction among its physical, psychophysical and psychological components [1]. As to the physics of binaural hearing, the following statement has for a long time been considered common wisdom in the field – since air is an (approximately) acoustically linear medium, the difference of the sound signals at the two ears can be considered an LTI system, and as such, can be described by means of an "interaural transfer function."

A transfer function can be separated into a magnitude function and a phase function, whereas the latter can also be plotted as a phase-delay and/or as a group-delay function. Consequently, methods were developed to measure individual interaural transfer functions for all possible angles of sound incidence with high accuracy. These functions were then analyzed, especially with regard to interaural-level differences and interaural arrival-time differences. The data collected in this way enable researchers to better understand the differences of the input signals to the ears and, further, to generate binaural input to the two ears for experimental purposes — with either natural differences or other, deliberately chosen differences.

Interaural transfer functions do not only play a role in the auditory systems' capability of forming the direction of the auditory events, but also in forming their perceptual distance. The article by Brungart (in this volume) elaborates on this relationship.

Although there in no doubt that the approach depicted above has gained fruitful results, it is basically restricted to the static case where the sound sources and the listeners have a fixed position in an environment where everything else is spatially fixed as well (e.g., nobody moving around, no doors being opened or closed). This is, indeed, a very constraining and highly unnatural assumption. In a realistic environment the system is neither time invariant nor linear. Listeners move their heads around all the time and sound sources are often moving about. Present research has just begun to touch the physical problems of binaural hearing in spatially variant environments (e.g., by considering the directionality of moving sources and listeners or Doppler shifts).

As to the psychoacoustics of binaural hearing, many models of signal processing in the subcortical auditory system are able to accurately simulate a substantial amount of binaural perceptual phenomena, such as sound localization in a free field with a limited number of concurrent sound sources, or enhancement of the signals from one sound source with respect to those from other ones (source separation). The paper by Ito & Akagi (this volume) is a good example along these lines of thinking. One of the highlights of this kind of binaural research is the availability of so-called "cocktail-party processors." These are programs which are able to separate one talker from concurrent ones in a multi-talker (cocktail-party) situation. The enhancement of the desired talker can be up to 20 dB (expressed as S/N-ratio improvement).

Yet current cocktail-party processors, based on interaural signal differences only, show a significant deficiency: They dramatically decrease in performance as soon as reflected sound is present (e.g. in reverberant spaces). The reason for this behavior is that they mainly evaluate interaural-arrival-time differences. Due to the superposition of direct and reflected sounds the interaural phase differences are seriously corrupted in reverberant environments and consequently, the interaural time-differences as well. The interesting fact is that the human auditory system can cope quite well with reverberant environment and still achieve a significant enhancement of the desired talker.

Two approaches to solve this problem can be observed at this time: First there is physiological research into the subcortical auditory system with the aim of a better understanding of how nature evaluates interaural signal differences (see Hartung & Sterbing, this volume). Second, models are being designed which take into account further acoustic cues in addition to interaural differences (e.g., harmonic structures, attack times, co-modulation). This is clearly an approach which reaches beyond the psychoacoustics of binaural hearing. As these systems often include pattern-recognition procedures they are, among other things, based on prior knowledge (cognition).

The role of cognition in binaural hearing is in fact one of the topical fields of research at this time. To name a few problems in this context — human listeners, when moving their heads around, know the direction and amount of these movements and interpret their effect on the signals at the two ears. It is well accepted that these dynamic cues are dominant relative to static cues in directional hearing, yet, it is hardly known how the auditory system actually processes them. In fact, in humans this problem is still simple as compared to animals which can move their pinnae around deliberately. Further, when forming auditory percepts, information from other senses, such as visual or tactile information, is taken into account by the central nervous system. As to the cocktail-party problem — when the listener knows the acoustic characteristics of a desired talker's voice, or even knows what he/she is talking about, speech enhancement is much better. Again, cognition is involved.

References

[1] Blauert, J. "An introduction to binaural technology." In *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. Anderson (eds.), Hilldale, NJ: Lawrence Erlbaum, pp. 593–609, 1996.

[2] Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization* (2nd ed.). Cambridge, MA: MIT Press, 1996.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

PRELIMINARY MODEL OF AUDITORY DISTANCE PERCEPTION FOR NEARBY SOURCES

Douglas S. Brungart

Air Force Research Laboratory Human Effectiveness Directorate
2610 Seventh Street, WPAFB
Dayton, OH 45433, USA

1. Introduction

After almost a century of research, the mechanisms that allow humans to determine the direction of a sound source are well documented. Auditory localization in the horizontal plane is known to depend primarily on differences in the time and intensity of the sound reaching the ears of a listener [14]. The sound arriving at the more distant ear is delayed and attenuated relative to the sound at the closer ear due to the longer propagation path as well as the diffraction effects of the head and torso. The resulting interaural time delay (ITD) and interaural intensity difference (IID) can be used to determine the lateral position of the source. Additional information about the location of the sound source is provided by the spectral shaping effects of the outer ear, or pinna [13]. Pinna-based spectral cues allow listeners to determine the elevation of a sound source and to distinguish between sounds in the front and rear hemispheres.

In a free-field listening environment these directional cues are effectively independent of distance when the sound source is located at least 1 m away from the listener. Virtually all auditory localization research has focused on this far-field region. However, when a sound source approaches within one meter of the head the interaural difference cues become highly dependent on the distance of the source. Several researchers have examined the distance-dependence of auditory localization cues for sources closer than a meter. Stewart [15] was the first to calculate the IID and ITD for a head modeled by a rigid sphere. This rigid-sphere model was extended by Hartley and Frey [9] and was recently revisited by Brungart and Rabinowitz [2]. Acoustic measurements of the IID and ITD as a function of distance have been made on a "bowling ball" head [5] and on anthropomorphic manikin heads [1] [7]. Both the calculations and the measurements have shown that the IID increases dramatically when a lateral sound source approaches the head while the ITD is roughly independent of distance, even when the source is close.

Two factors contribute to this increase in IID for nearby sources. The first is related to the scattering of sound by the head and torso. When a sound source is located outside the median plane, the direct path from the source to the more distant ear is blocked by the head. The sound must diffract around the head and torso to reach the further ear and is attenuated in the process. Thus, the further ear is said to lie in the acoustic "shadow" of the head. The magnitude of this head-shadowing effect depends on both the frequency and distance of the sound source. The frequency dependence is determined by the size of the head relative to the wavelength of sound. Thus, the head-shadowing effects are negligible at low frequencies, where the head is small relative to the wavelength of the sound, and increase systematically with increasing frequency. The dependence on distance is determined by the size of the head relative to the distance of the source. When the source distance is large relative to the width of the head, the attenuation due to head-shadowing approaches an asymptotic value, and is roughly independent of distance. At closer distances, however, the effects of head-shadowing increase systematically with decreasing distance. This systematic increase contributes to the enlarged IIDs found for close sources. Head-shadowing is not the only factor contributing to these enlarged IIDs. When the source is close to the head, the inverse relationship between the intensity of a sound and the square of its distance from the source plays a major role in determining the IID. As a sound source approaches the head, the intensity of the sound increases more rapidly at the location of the closer ear than at the further ear, resulting in an increased IID. This can be illustrated by the simple example of a sound source along the interaural axis of a listener with ears separated by 20 cm. When the sound source is located 1 m from the center of the head, the source is located 110 cm away from the further ear and 90 cm away from the closer ear. Ignoring the effects of diffraction by the head, the IID will be 1.7 dB. If the distance of the source is decreased from 1 m to 15 cm, the source will be located 25 cm from the further ear but only 5 cm from the closer ear, resulting in an interaural intensity difference of 14 dB. Thus, a substantial increase in IID is expected for close sources, even if the effects of diffraction by the head are ignored. Note that, in contrast to the effects of head-shadowing, these proximity effects are independent of the frequency of the sound source. For this reason source-proximity effects can produce substantial low-frequency IIDs for nearby sources that would never occur at more distant locations. For example, the IID at 500 Hz for a source along the interaural axis is approximately 20 dB at a distance of 12 cm, but never exceeds 6 dB when the source distance is greater than 1 m [1].

The combined effects of head-shadowing and source proximity can produce dramatic increases in the IID as a sound source approaches the head. In contrast, the interaural time delay, which depends primarily on the difference in path lengths from the source to the left and right ears, is roughly independent of distance even for nearby sources. This contrast is especially dramatic when perceptual issues are considered. Although the ITD has been shown to increase slightly as the source approaches the head, both in calculations of the ITD for a rigid sphere model of the head [2] [4] [5] and measurements of the ITD with a KEMAR acoustic manikin [1], most of this increase occurs at lateral locations where the ITD is greater than 400 ms and the auditory system is known to be relatively insensitive to changes in ITD. In terms of earlier measurements of the smallest perceptible change (just-noticeable-difference or JND) in ITD for a 500-Hz tone [10], the ITD never increases by more than 2-3 JND units as a source at a fixed direction approaches the head. In contrast, the IID for a lateral source will increase by 15-20 or more JND units as distance decreases from 1 m to 12 cm.

The combination of distance-varying IIDs and distance-invariant ITDs makes it theoretically possible for a listener to determine the distance of a nearby sound source by first determining its lateral position from the ITD and then estimating its distance based on the magnitude of the IID. This model of near-field distance perception, which was first proposed by Harley and Frey in 1921 [9], is supported by recent measurements of auditory localization that have shown that distance perception is relatively good for nearby lateral sources and poor for nearby medial sources [1]. This chapter discusses two models of auditory distance perception based on binaural difference cues. The first model, described in the next section, was first proposed by Hirsch (and later expanded by Molino) and is based on a geometrical calculation of the path lengths from a source distant to the left and right ears. The second model, described in the remaining portion of the chapter, is a new model of binaural

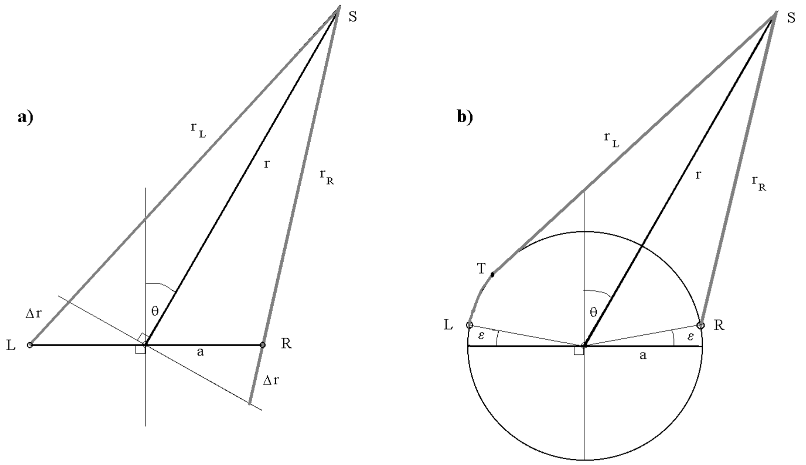


Figure 1 Models of binaural distance perception. In these figures, a sound produced by source, *S*, at distance, *r,* from the center of the head and at angle, *θ,* from the median plane is received at the left and right ears (*L* and *R*) of a listener with head radius, *a*. In Hirsch's model (a), diffraction effects of the head are ignored and the sound follows the direct paths SL and SR to the ears. In Molino's model (b), the sound travels a direct path to the closer ear (*SR*) but must wrap around the surface of the head to reach the farther ear (*L*). This is accomplished by following the tangent line from the source to the surface of the sphere (*ST*) and then curving around the circumference of the head from the tangent point to the ear (*TL*), thus following the total path *STL*. Note that in Molino's model the ears are placed forward of the frontal plane by angle *ε*. See text for details.

distance perception based on measurements of the IID for a 500-Hz tone on a KEMAR manikin and the JND for a 500-Hz tone measured by Hershkowitz and Durlach [10].

This new model is used to calculate the minimum perceptible decrease in the distance of a sound source as a function of distance and direction, as well as to simulate performance in an auditory-distance-identification experiment. The results of this simulation are compared to data from a psychoacoustic localization experiment, and the advantages and disadvantages of the model are discussed.

2. Hirsch's Model of Binaural Distance Perception

In 1968, Hirsch [11] proposed a model of binaural distance perception that allowed a listener to determine the distance of a sound source directly from the IID and ITD. In Hirsch's model, the effects of sound diffraction by the head are ignored and the ears are represented by point receivers in free space separated by the diameter of the head, 2*a* (left panel of Figure 1). Hirsch showed that, if the distance from the center of the head to source, *S*, is sufficiently large, the path lengths to the left and right ears can be approximated by:

$$r_L \approx r + \Delta r \tag{1}$$

$$r_R \approx r - \Delta r \tag{2}$$

where

$$\Delta r = a \sin(\theta), \tag{3}$$

r is the distance from the source to the center of the head, θ is the angle from the median sagittal plane to the sound source, and a is the radius of the head.

The interaural time delay $\Delta t(\theta)$ is simply the additional propagation time required for the sound to reach the further ear after it reaches the closer ear, so

$$\Delta t(\theta) = \frac{r_L - r_R}{c} = \frac{2\Delta r}{c} = \frac{2a}{c}\sin(\theta) \tag{4}$$

where c is the speed of sound. The intensity of the sound is inversely proportional to the square of its distance from the source, so the interaural intensity difference, i (expressed as the ratio of the increase in amplitude from the left ear to the right ear relative to the amplitude at the left ear), can be defined as:

$$\frac{i_r - i_l}{i_l} = \frac{\dfrac{I_0}{(r - \Delta r)^2} - \dfrac{I_0}{(r + \Delta r)^2}}{\dfrac{I_0}{(r + \Delta r)^2}} \approx \frac{4\Delta r}{r} \tag{5}$$

where i_r is the intensity at the right ear, i_l is the intensity at the left ear, and I_0 is the intensity of the sound source at a distance of 1 m in the free-field. Solving equation 5 for r and substituting from equations 3 and 4 yields

$$r \approx \frac{4\Delta r}{i} \approx \frac{2c\Delta t}{i} \tag{6}$$

Thus, under the free-field assumptions of this model, the range of a sufficiently distant sound source can be determined directly from the ratio of the interaural time delay to the interaural intensity difference without any knowledge about the intensity of the source.

Hirsch's model is appealing because it reduces the determination of distance based on binaural cues to a simple ratio of the interaural time and intensity differences. The model does, however, rely on the unrealistic assumption that diffraction by the head is negligible. Furthermore, the model is restricted to relatively distant source locations, where the interaural difference cues are virtually independent of distance. Consequently, a listener using Hirsh's model for distance perception would perform poorly. Using measured discrimination thresholds of 28 *μs* for Δt and 0.4 *dB* for i, Greene [8] showed that the smallest detectable relative change in distance predicted by Hirsch's model for a source at $\theta = 90^\circ$, $r = 2$ m, and $a = 10$ cm can be estimated as follows:

$$\frac{\delta r}{r} = \frac{\delta(\Delta t)}{\Delta t} + \frac{\delta i}{i} = \frac{28\mu\text{s}}{600\mu\text{s}} + \frac{0.092}{0.200} = 0.507 \tag{7}$$

Thus, even for a lateral source producing the maximum possible binaural difference cues, a listener making distance judgments based on Hirsch's model will only be able to detect changes in distance on the order of 50%.

Molino [12] has suggested an augmented version of Hirsch's model that eliminates the free-field assumption and the requirement of a distant source. Molino's model approximates the head as a rigid sphere and estimates the path length from the source to the contralateral ear by assuming the sound first follows the path from, *S,* tangent to the sphere at point *T*, and then wraps around the circumference of the sphere to the left ear, *L* (line *STL* in Figure 1). The sound reaching the closer ear travels a direct path from the source to the ear (line *SR*). Molino also assumes that the ears are separated by 165° rather than 180°, placing the ears slightly forward on the head (at angle, ε, from the frontal plane). These assumptions result in a significantly more complicated estimate of the path-length difference than that of Hirsch's model (Equation 3):

$$STL\text{-}SR = (r^2 - a^2)^{0.5} + a(\sin^{-1}(a/r) + \theta - \varepsilon) - (r^2 - 2ar\sin(\theta + \varepsilon) + a^2)^{0.5} \qquad (8)$$

Molino's model demonstrates that Hirsch's prediction — the difference in path length between the left and right ears changes as a function of distance holds even when significantly more realistic assumptions about source distance and head diffraction are used. However, it eliminates the appealing mathematical relationship between the distance of the source and the interaural time and intensity differences provided by Hirsch's model, and substantially increases the complexity of the calculation of ITD and IID (for which no equation is provided). Furthermore, Molino's augmented model requires the subject to have *a priori* knowledge about the direction of the source in order to unambiguously calculate source distance. In Hirsch's model, the equation for calculating *r* is based solely on the ratio of ITD to IID and is independent of θ (Equation 6), so no directional information is required to determine the distance of the source. In Molino's model, however, certain combinations of *r* and θ will produce identical IID and ITD pairs. Therefore, it is necessary to know the direction of the sound source *independent of the ITD* in order to calculate its distance using Molino's model. Consequently, it is substantially harder to predict human localization performance with Molino's extended model than with Hirsch's original model.

3. Assumptions of the Model

Although they are conceptually interesting, neither Hirsch's model nor Molino's extension of that model are capable of predicting human localization performance for nearby sound sources. The rest of this chapter outlines a new model of auditory-distance perception for nearby sources based on acoustic measurements of the IID from a KEMAR manikin and measurements of the JND in IID for a 500-Hz tone. This model is based on three major assumptions.

3.1 Accurate Knowledge of Lateral Position

The first assumption of the model is that the listener is able to use the ITD, which is largely independent of the distance of the source, to determine the lateral position of the sound. The model assumes that the listener is able to determine the lateral position accurately, so any errors caused by misperceptions about the direction of the sound are ignored. The model does, however, account for front–back confusions about the location of the sound, where a sound is perceived to be at the mirror image of its true location across the frontal plane [16].

3.2 Distance Judgments Based on IID

The second assumption of the model is that the listener makes his or her judgments about the distance of the sound source based solely on the perceived IID of the stimulus. In this model, which is loosely based on the Durlach and Braida model of intensity perception [6], each exposure to the stimulus generates a Gaussian-distributed random perceptual variable

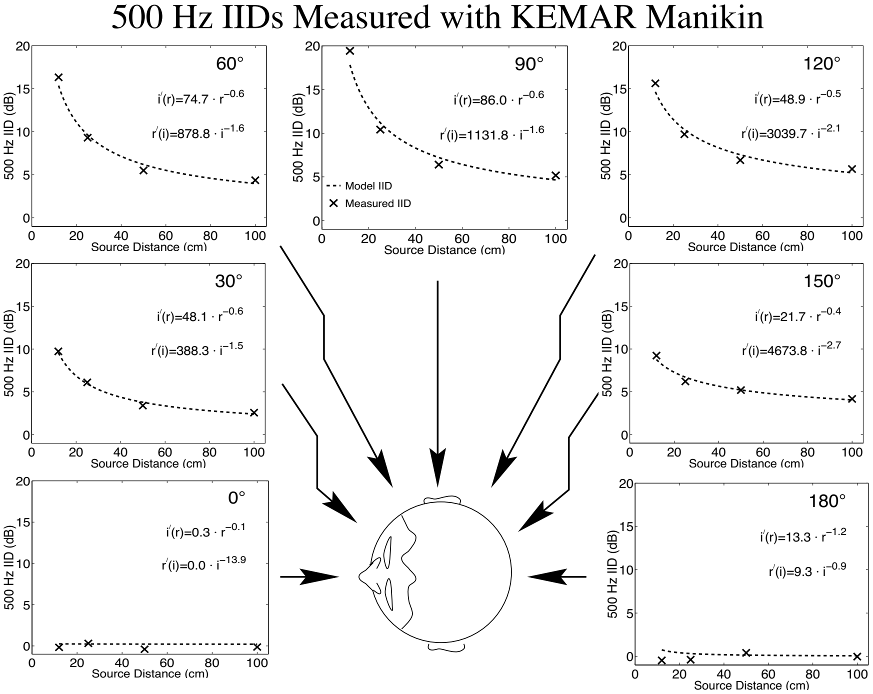


Figure 2 Interaural intensity difference for a 500 Hz tone as a function of source distance. The x's represent measurements made with a KEMAR manikin. The dashed lines represent the IID model based on the best linear fit of the log of the IID (in dB) to the log of the source distance, of the form $i' = \alpha r^{\beta}$. The coefficients of the model, as well as the inverse function relating distance to IID, are shown at the upper right of each panel.

with a mean value equal to the actual IID and a standard deviation of 0.8 dB. The standard deviation is derived from the measurements of Hershkowitz and Durlach [10], who found that the 75% correct JND in IID for a 500-Hz tone was approximately 0.8 dB across a wide range of reference IID values. The 75% criterion corresponds to the probability of correctly distinguishing between two identically distributed Gaussian random variables (the perceived IID during the reference and reference+Δ presentations of the stimulus) with their means separated by one standard deviation. Thus, the 75% correct JND can be considered approximately equivalent to the standard deviation of the underlying perceptual variable associated with the IID.

3.3 Perfect Mapping from IID to Distance

The third assumption of the model is that the source distance, *r,* for a given location in azimuth is related to the IID, *i*, (expressed in dB) by an equation of the form:

$$r = \alpha i^{\beta} \tag{9}$$

and that the listener is able to map the perceived IID to the perceived distance of the source according to this equation without error. Note that the listener is required to have access to a different map of this type at each azimuth location.

4. Calculation of Model Parameters

The parameters of the model were derived from measurements of the IID at 500 Hz made with a KEMAR manikin and a compact, acoustic-point source [1]. The measurements, which were taken at distances of 12 cm, 25 cm, 50 cm, and 100 cm from the center of the manikin head, are shown by the x's in each panel of Figure 2. The dashed line represents the linear least-squared-error estimate of the (linear) IID as a function of the log of the source distance. The dB value of the IID can be expressed in the form:

$$i'(r) = Ar^{B} \tag{10}$$

where i' is the interaural intensity difference (in dB) and r is the distance of the source (in cm). The equation for the curve is shown in this form at the upper right of each panel.

The inverse of this function can be used to calculate the source distance, given the IID in dB. The inverse curve has the form

$$r'(i) = \alpha i^{\beta} \tag{11}$$

where

$$\alpha = A^{-B^{-1}} \tag{12}$$

and

$$\beta = 1/\mathrm{B} \tag{13}$$

The equation for the inverse curve is also shown on each panel.

Note that Figure 2 shows the IID curves for the entire right hemisphere. Although the differences between the IID curves at symmetric locations in the front and rear hemispheres are relatively minor (the curve is generally slightly flatter in the rear hemisphere), the parameters of the inverse curve equation are substantially different. Thus, based on the assumptions of this model, it is reasonable to anticipate that the performance would be degraded in a front–back confusion, where the sound in the front hemisphere is perceived in the rear hemisphere (or vice-versa). The treatment of front–back confusions by the model is discussed in Section 7.

5. Model Estimate of Percent JND Decrease in Distance

It is relatively easy to determine the smallest perceptible percentage decrease in the distance of a sound source (% JND) with this model. Figure 3 shows %JND in the distance of a sound source predicted by the model at five source locations ranging between 30° and 150° in azimuth. The threshold decrease in distance is the decrease necessary to increase the IID by one JND, or 0.8 dB. Thus, for a given distance, *r*, the % JND in distance (for 75% correct detection) is equivalent to:

$$\frac{r - r'(i'(r) + 0.8)}{r} \cdot 100 \tag{14}$$

As seen in Figure 3, the % JND in distance decreases as the source distance decreases.

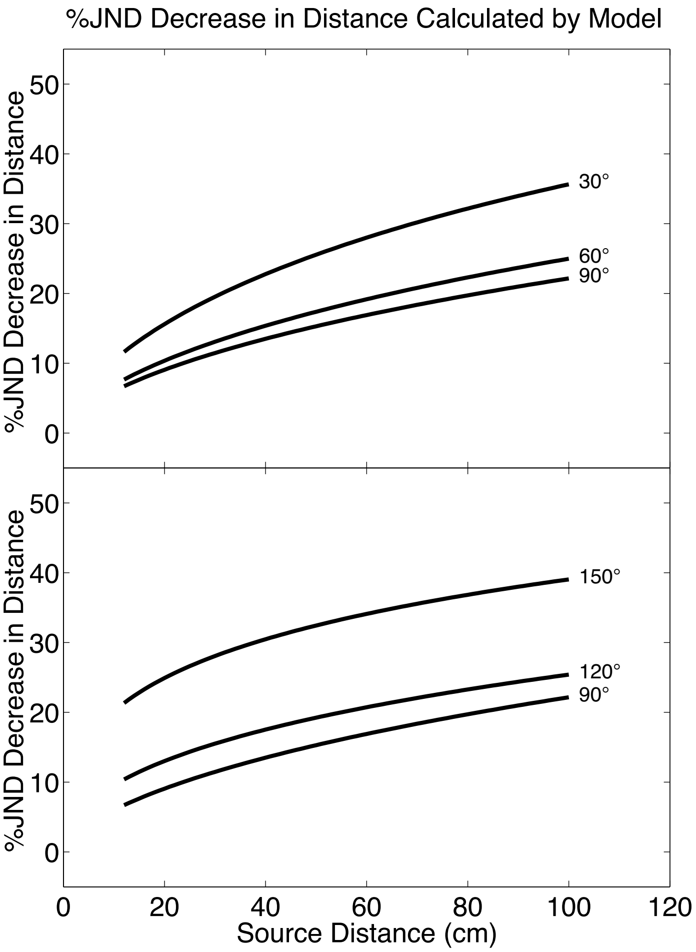


Figure 3 Predictions of %JND decrease in distance as a function of source distance and direction. Locations in the front and rear hemispheres are shown in the top and bottom panels, respectively. No data are shown at 0° or 180° where the IID is negligible at all distances and no IID-based distance judgments are possible.

Note that this implies that very small changes in distance will be detectable when the source is located near the head. For a source 12 cm from the center of the head at 90° in azimuth, a decrease in distance of 8%, or 1 cm, will produce a detectable change in IID. In contrast, a decrease in distance of more than 20 cm is required to produce a detectable change in IID when the source is located 1 m from the listener.

The %JND is only slightly dependent on azimuth between 60° and 120°, but increases substantially at the more medial locations (30° and 150°) where the IID is less distance-dependent. In the median plane the IID is essentially 0 dB at all distances, and no changes in distance can be detected based on the IID.

These predictions of the % JND in distance should be robust in the sense that they provide an upper bound on the minimum detectable change in the distance of a sound source near the head. This level of performance could be obtained simply by listening for changes in the IID, and does not require accurate knowledge about the direction of the source. The next section describes a more advanced adaptation of the model that predicts behavior in a distance-identification experiment, where listeners are required to determine the location of the sound from a single presentation of the stimulus.

6. Simulation of Distance Identification Performance

Distance identification is fundamentally more difficult than distance discrimination. In discrimination, the listener is required only to determine whether there is a difference in IID between two consecutive presentations of a stimulus. In identification, however, the listener is required to make an absolute judgment about the distance of the sound from a single presentation of the stimulus. In this model of distance perception, the listener must determine the lateral position of the source, choose the correct map between IID and distance for that lateral position, and use that map to derive the distance of the source from the perceived IID. Under the assumption that lateral position was correctly determined by the listener, the perceived distance of a single stimulus presentation at distance, $r_{stimulus}$, and angle, θ, is simulated by the following equation:

$$r_{response} = (1\text{-}V)(r'_{\theta}\,(i'(r_{stimulus}) + n_{iid})) + (V)(r'_{180\text{-}\theta}\,(i'(r_{stimulus}) + n_{iid})) + n_{response} \quad (15)$$

- $r_{\theta}'()$ is the function relating IID to distance at location, θ (Equation 11);
- $r'_{180-\theta}()$ is the function relating IID to distance at the reversed location of θ;
- $i'()$ is the inverse of $r'_{\theta}()$ (Equation 10);
- n_{iid} is a zero mean Gaussian random variable with $\sigma = 0.8$ dB representing internal noise associated with the perception of the IID;
- $n_{response}$ is a zero-mean Gaussian random variable with $\sigma = 8$ cm representing noise in the response location indicated by the subject;
- V is a Boolean random variable equal to 1 when a front–back reversal occurs ($p = 0.14$).

The value of σ for n_{iid} was based on the measured JND in IID. The probability distributions of the other variables were chosen to best match the data from a psychoacoustic distance localization experiment described in the next section [1]. The probability of V was chosen to match the percentage of front–back reversals that occurred in the psychoacoustic experiment, and the value of $n_{response}$ was chosen to best match the errors found in the psychoacoustic data.

The results of a simulation of distance identification performance using this model are shown in the left panels of Figure 4. Each panel represents 1000 simulated trials with randomly distributed stimulus distances from 12 cm to 1 m. Response locations were restricted to the range 12 cm to 150 cm, and at locations in the median plane (0° and 180°) the IID varies little with distance and the majority of the simulated responses are at the maximum or minimum value.

The number at the upper right of each panel shows the correlation coefficient between the log of the stimulus distance and the log of the response distance at that source location. In the median plane, the stimulus and response locations are essentially uncorrelated. As the source position moves more lateral to the head, the correlation increases systematically until at 90° the correlation reaches its maximum value of 0.89.Comparison of Simulation Data to Psychoacoustic Data

The right panels of Figure 4 show the results of an auditory localization experiment conducted by Brungart [1]. In this experiment, listeners were asked to attend to a series of broadband noise bursts produced by a compact acoustic point source at a random location in

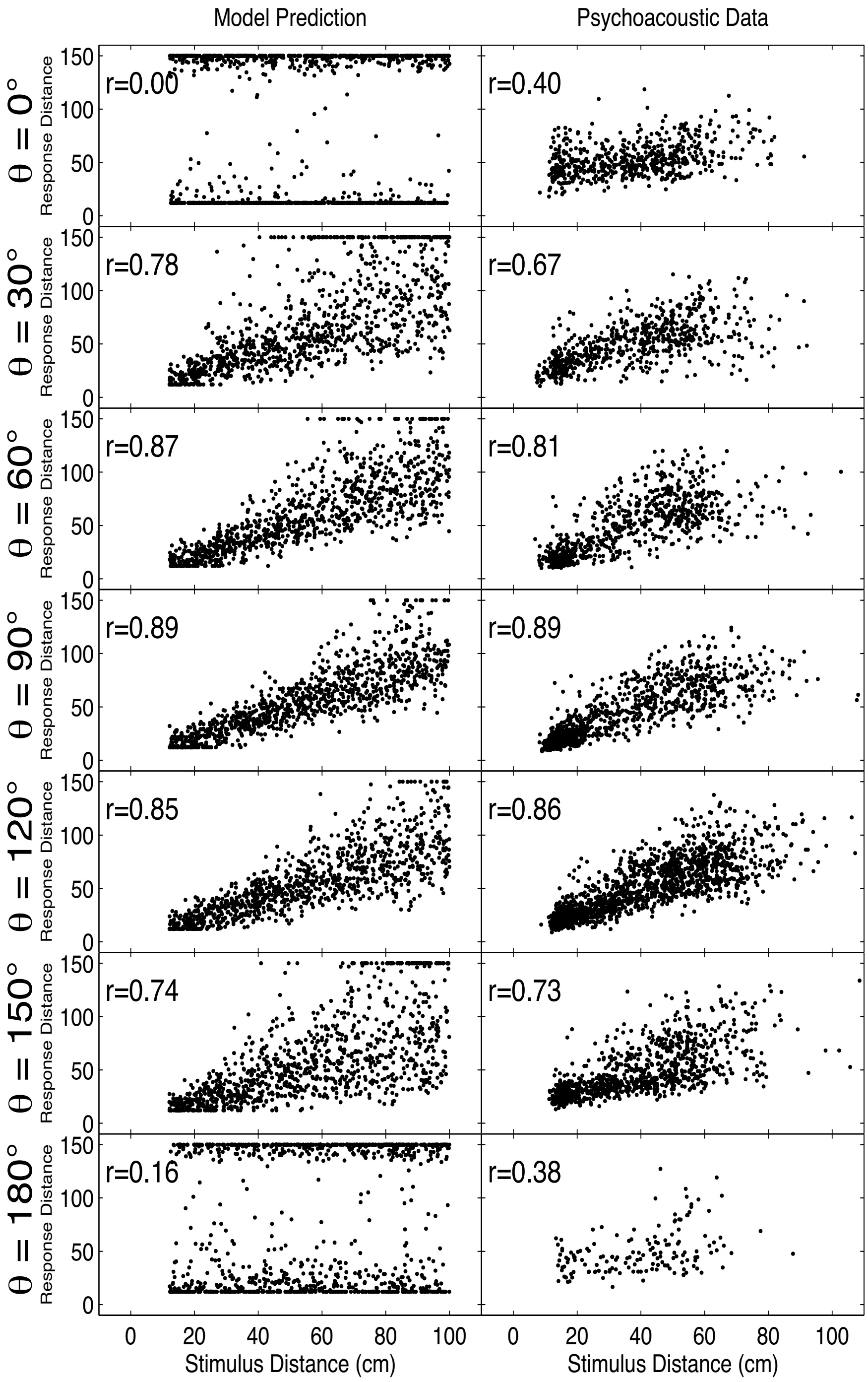


Figure 4 Comparison of localization performance predicted by the model with psychoacoustic localization data at seven azimuthal locations. The correlation coefficient between the log of the stimulus distrance and the log of the response distance is shown in the upper left of each panel.

the right hemisphere, and to identify the location of the source by moving a response pointer to the perceived location of the stimulus. The amplitude of the stimulus was randomized to eliminate intensity-based distance cues. A total of 2000 trials were collected for each of four subjects in the experiment. The results, shown in the right panels of Figure 4, are limited to trials at elevations between –30° and 30°. They have been sorted into seven azimuthal bins, each containing all trials within the 30° range of azimuths centered at the value of θ shown at the left side of the figure. As in the simulation data, the correlation coefficient between the log of the stimulus distance and the log of the response distance is shown at the upper left of each panel.

The noise parameters of the simulation were chosen to generate similar overall performance for azimuth locations near 90°. Thus, the correlation coefficient for the simulated and actual data is similar for sources from 60° to 120°. As in the simulation, performance in the psychoacoustic experiment diminishes as the source position moves toward the median plane. The correlation does not, however, decrease to zero in the median plane. The superior performance in the psychoacoustic experiment probably occurs in part because the psychoacoustic data include locations ranging from –15° to 15° in azimuth, where there is some variation in the IID with distance, while the simulation data include only sources in the median plane. There may also be some monaural spectral distance cues that contribute to localization performance in the median plane and are not covered by this preliminary model [1]. A visual inspection of the raw response data indicates that the primary difference between the model data and the psychoacoustic data is the large number of responses in the simulation which occurred at the limiting distance of 150 cm. Clearly, the model does not adequately represent the type of responses that occur when the perceived IID is low. Either a distribution of n_{iid} with shorter tails than provided by the Gaussian distribution is required, or the responses below a certain IID value should be randomly distributed over a range of possible responses (from 100 to 150 cm, for example) to more accurately simulate the type of responses exhibited by the subjects.

Another interesting difference between the simulation data and the psychoacoustic data is that performance in the simulation is slightly, but significantly, better at $\theta = 30^{\circ}$ than at $\theta = 150^{\circ}$ ($p < 0.05$, from the Fisher transform of correlation coefficients [3]), while performance in the psychoacoustic experiment was slightly, but significantly, better at $\theta = 150^{\circ}$ than at $\theta = 30^{\circ}$ ($p < 0.05$, Fisher transform). The KEMAR HRTF measurements in Figure 2 indicate that the IID is slightly more sensitive to distance at 30°, which is reflected in the behavior of the model. It is unclear why the responses in the psychoacoustic experiment were slightly more accurate in the rear hemisphere.

7. Conclusions

This chapter has examined a preliminary model of near-field auditory depth perception based on distance-dependent variations in the IID for a nearby source. The important attributes of this model can be summarized as follows:

- The model predicts that the distance discrimination threshold for nearby sources, expressed as the minimum audible percent decrease in distance, will decrease as the source moves lateral to the head and as the source approaches the head. When the source is located just outside the ear ($\theta = 90^{\circ}$ and $r = 12$ cm) the model suggests that changes in source distance as small as 1 cm will be detectable based solely on changes in the IID. Although no psychoacoustic data are currently available to verify these predictions, a series of experiments measuring the JND in distance for sources within 1 m of the head is planned.

- The model predicts that performance in a distance identification experiment will systematically improve as the source moves lateral to the head, and worsen as the source moves toward the median plane. These predictions are verified by data from a psychoacoustic localization experiment. However, the psychoacoustic data indicate that performance does not degrade completely in the median plane, as predicted by the IID-based distance perception model. It is likely that monaural spectral cues are providing additional distance information in the median plane, which is not reflected by the model. This type of spectral information could be included in more advanced versions of the model.
- The model predicts somewhat more accurate distance performance at 30° than at 150°, based on the greater sensitivity of the IID to changes in distance at 30° in the KEMAR HRTFs, but this behavior is not reflected in the psychoacoustic data. This discrepancy may be a result of acoustic differences between the KEMAR manikin, which was used to measure the IIDs used by the model and by the actual subjects. It may also result from the assumption that the precise lateral position of the source is derivable from the ITD. A more advanced version of the model would more accurately account for uncertainty in the azimuthal position of the source.

While this is only a preliminary model, it does accurately portray the general direction-dependent behavior of auditory distance localization in the region within 1 m of the listener's head. The accuracy of its predictions also provides evidence of the importance of the IID to the distance perception of nearby sources. Further work is necessary to establish a more comprehensive model that includes monaural spectral cues and more realistic assumptions about the perception of source direction.

Acknowledgements

Parts of this work were performed at the Massachusetts Institute of Technology's Research Laboratory of Electronics. The author would like to thank Nat Durlach, Bill Rabinowitz, Bill Peake and Steve Colburn for their assistance during this research, and Robert Bolia and Rich McKinley for their comments on the manuscript.

References

[1] Brungart, D. S. *Near-Field Auditory Localization*. Ph.D. Thesis, Massachusetts Institute of Technology, 1998.

[2] Brungart, D. S. and Rabinowitz, W. M. "Auditory localization in the near-field." *Proc. of the Third Intern. Conf. on Aud. Disp.* Santa Fe Institute, 1996.

[3] Devore, J. L. *Probability and Statistics for Engineering and the Sciences*. Pacific Grove, CA: Brooks/Cole, 1991.

[4] Duda, R. O. and Martens, W. L. "Range dependence of the HRTF for a spherical head." *Proc. 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Mohonk, NY, October 1997.

[5] Duda, R. O. and Martens, W. L. "Range dependence of a spherical head model." *J. Acoust. Soc. Am.*, 104: 3048–3058.

[6] Durlach, N. I. and Braida, L. D. "Intensity perception. I. Preliminary theory of intensity perception." *J. Acoust. Soc. Am.*, 46: 372–383, 1969.

[7] Firestone, F. A. "The phase difference and amplitude ratio at the ears due to a source of pure tone." *J. Acoust. Soc. Am.*, 2: 260–270, 1930.

[8] Greene, D. L. "Comments on 'Perception of the range of a sound source of unknown strength'." *J. Acoust. Soc. Am.*, 44: 634, 1968.

[9] Hartley, R. V. and Frey, T. C. "The binaural location of pure tones." *Phys. Rev.*, 18: 431–442, 1921.

[10] Hershkowitz, R. M. and Durlach, N. I. “Interaural time and amplitude jnds for a 500-Hz tone.” *J. Acoust. Soc. Am.*, 46: 1464–1467, 1969.

[11] Hirsch, H. R. “Perception of the range of a sound source of unknown strength.” *J. Acoust. Soc. Am.*, 43: 373–374, 1968.

[12] Molino, J. “Perceiving the range of a sound source when the direction is known.” *J. Acoust. Soc. Am.*, 53: 1301–1304, 1970.

[13] Musicant, A. D. and Butler, A. D. “The influence of pinnae-based spectral cues on sound localization.” *J. Acoust. Soc. Am.*, 75:1195–1200, 1984.

[14] Rayleigh, Lord “On our perception of sound direction.” *Phil. Mag.*, 13: 214–232, 1907.

[15] Stewart, G. W. “The acoustic shadow of a rigid sphere with certain applications in architectural acoustics and audition.” *Phys. Rev.*, 33: 467–479, 1911.

[16] Wallach, H. “The role of head movements and vestibular and visual cues in sound localization.” *J. Exp. Psych.*, 27: 339–368, 1940.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

A COMPUTATIONAL MODEL OF AUDITORY SOUND LOCALIZATION

Kazuhito Ito and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology (JAIST)
1-1 Asahidai, Tatsunokuchi, Ishikawa, 923-12 Japan

1. Introduction

Sound localization based on the interaural time difference (ITD) detects sound source locations using the difference in arrival times of the sound waves at the two ears [11] [12]. To understand the process and represent it computationally, we developed a computational model of auditory sound localization based on the ITD.

Sound waves arriving at the ears are decomposed into their frequency components and are changed into impulse trains by the auditory periphery. The impulse trains accurately represent the time intervals between firings because auditory nerve firings tend to be phase-locked or synchronized to the stimulating waveforms (Figure 1) [11][13]. The difference between the temporal information from the two ears is used for sound localization [2]. It is known that humans can perceive an ITD variation of about 10 μs at 900 Hz, corresponding to a minimum audible angle (MAA) of about 1° [11][10].

A nerve impulse is an electrical excitation called an action potential and its duration is over 1 ms. Additional synaptic transmissions in the auditory system extend the duration of the signal [1]. Thus the duration of an auditory firing is long compared with that of the ITD perceived by humans. Such a long-duration signal is problematic for the minute temporal information that should be transmitted and it may obscure temporal information. Although impulses from the auditory nerves are in synchronization with a particular phase of the stimuli, it is known that impulses fluctuate slightly in time [6]. Again, this may obscure temporal information. Given all these conditions, it is amazing that humans can perceive an ITD variation of about 10 μs.

In this study, the signals in the nervous system such as action potentials and synaptic transmission, were modeled computationally and these models were used to detect ITDs. Impulse trains, with fluctuation in time that simulate spikes in the auditory nerve fibers, were

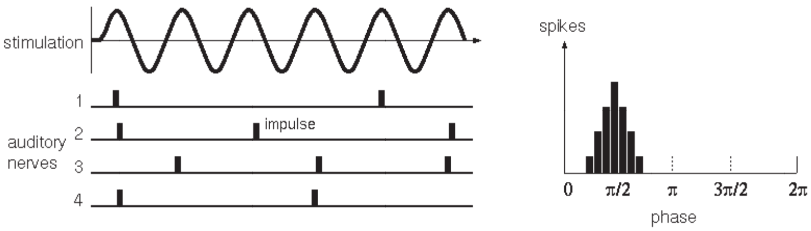


Figure 1 Temporal information and impulse fluctuation.

used as input data to the model. Then processes of sound localization with temporal redundancy and impulse fluctuation were studied through this model.

2. Representation of Signals

To represent the action potential computationally, we use Hodgkin–Huxley type equations [3] [14] [9].

$$C_m \frac{dV(t)}{dt} = -g_{Na}(V(t)-E_{Na}) - g_K(V(t)-E_K) - g_L(V(t)-E_m) \tag{1}$$

Here, $V(t)$ is the membrane potential at time t, C_m is the membrane capacitance, and E_m is the resting potential; E_{Na} and E_{K} are equilibrium potentials for sodium and potassium, respectively; and g_{Na} and g_{K} are the conductances of sodium and potassium. The ion conductances are

$$g_n(t) = a_n(t-t_n)e^{-(t-t_n)/\tau_n} \tag{2}$$

where $g_n(t)$ indicates the conductance for ion n at time t, t_n is the time of the most recent onset of ion n conductance, τ_n is the time constant for the conductance, and a_n is the amplitude constant related to the permeability of ion n. Then, the leakage conductance g_L is represented by

$$g_L(t) = a_L 1 - e^{-\alpha|V(t)-E_m|} \tag{3}$$

where α is the coefficient for the relationship between the leakage conductance and the membrane potential and a_L *is* the maximum conductance.

Synaptic transmission is also represented by the same type of equations. Although applying these equations might not be accurate in this case [1], this is one way to model the temporal redundancy of signals. The equations that describe the behavior of synaptic potentials, (4) and (5), do not model the firing of the post-synaptic cell. These firing thresholds aredefined in equation (6)

$$C_m \frac{dV(t)}{dt} = -G_{Na}(V(t)-E_{Na}) - G_K(V(t)-E_K) - g_L(V(t)-E_m) \tag{4}$$

$$G_n(t) = A_n(t-t_n)e^{-(t-t_n)/T_n} \tag{5}$$

where $V(t)$ is the postsynaptic potential at time t; C_m is the membrane capacitance; E_m is the resting potential; E_{Na} and E_{K} are the equilibrium potentials for sodium and potassium; and G_{Na}, G_{K}, and g_L are the postsynaptic conductances for sodium, potassium, and the leakage, respectively. Again $G_n(t)$ indicates the postsynaptic conductance of ion n at time t. t_n is the time of the most recent onset of ion n conductance. T_n is the time constant for the postsynaptic conductance, and A_n is the amplitude constant related to the permeability of ion n.

In this model — to represent the effects of other conductances such as early potassium channels, voltage-gated calcium channels, and calcium-activated potassium channels — the firing threshold level is varied according to Eq. (6), which translates the magnitude of the grand postsynaptic potentials into the frequency of firing of action potentials [1].

$$V_{threshold}(t) = \beta \cdot e^{-(t-t_r)/(\tau_r+\tau_a)} + E_{threshold} \tag{6}$$

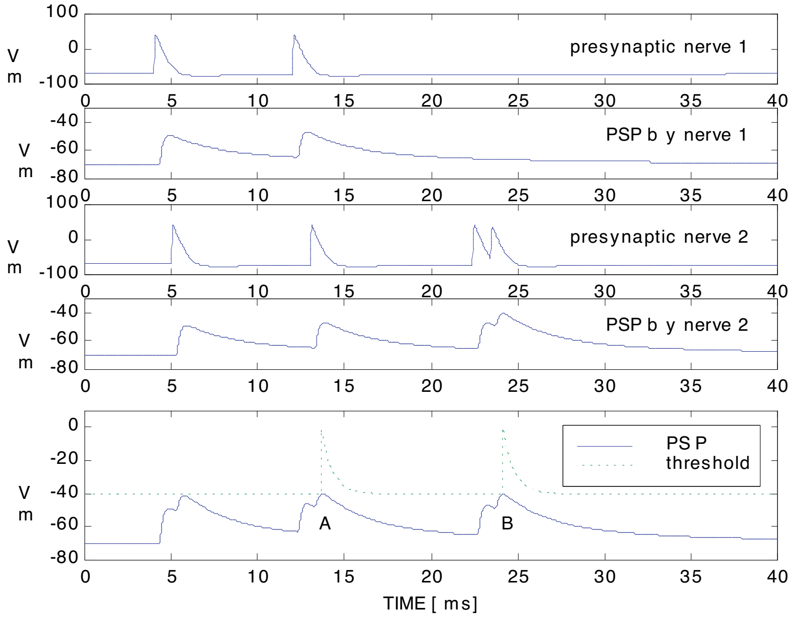


Figure 2 Spatial summation (A) and temporal summation (B)

where $V_{\text{threshold}}(t)$ indicates the threshold level at time t and t_r is the time of the most recent discharge. $E_{\text{threshold}}$ is the basis of threshold level in this function, τ_r is the time constant representing the relative refractory period, and β is the amplitude constant. And to express adaptation to a prolonged stimulation, τ_a is used

$$\tau_a = \tau_{max}\left(1 - \exp\left[-\gamma\int_{t-T}^{t}(V_{threshold}(t) - E_{threshold})dt\right]\right) \tag{7}$$

Adaptation related to the frequency of firing during a certain period of T (hundreds of ms) is represented by extending the length of the relative refractory period τ_a of $V(t)$, and γ is the coefficient for the relationship between the permeability of potassium and the frequency of firing. Hence, if

$$V(t) \geq V_{threshold}(t) \text{ for t-t}_{\text{r}} > \text{absolute refractory period} \tag{8}$$

then the postsynaptic cell fires.

Simulations of these models are illustrated in Figure 2. Presynaptic action potentials are shown in the top panel labeled ‘nerve 1’ and each action potential arriving at a synapse produces a postsynaptic potential (PSP) on the postsynaptic membrane of a cell, as shown in the second panel down from the top. The other pair are shown in the third and fourth panels, labeled ‘nerve.’ When both presynaptic nerves innervate the same cell, PSPs produced by both nerves are summed to produce a larger PSP. While spatial summation combines the effects of signals received at different sites on the membrane (at A in the bottom panel), temporal summation combines the effects of signals received at different times (at B in the bottom panel). The firing threshold level is illustrated by the dotted line in the bottom panel.

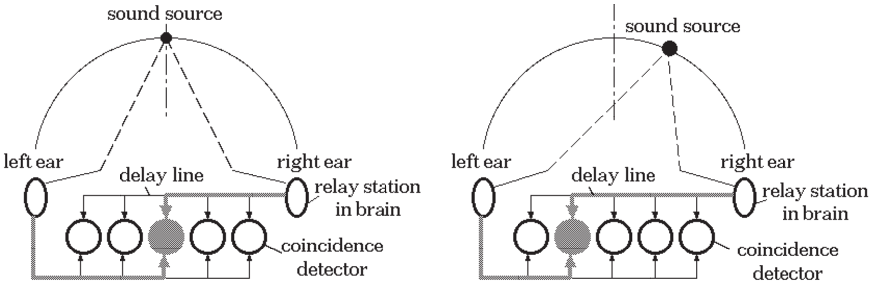


Figure 3 A coincidence detector circuit.

When the summed PSP exceeds a given threshold level, an action potential is generated on the postsynaptic cell. Then, these representations of the signals in the nervous system are used to detect ITDs.

3. Model Circuits For Detecting ITDs

3.1 Cross Correlation

The system for detecting ITDs exists in the medial superior olive (MSO), where the neural pathways from the left and right ears cross for the first time [17] [7]. The Jeffress model is well known as a model for this detection [5]. It is represented as a circuit consisting of an array of coincidence detectors and two nerve fibers from the left and right ears. The coincidence detectors fire most often when impulses from both sides arrive simultaneously. The model calculates ITDs with cross correlation between impulse trains coming from both sides (Figure 3). When a sound source is placed in front of the face, the arrival times from the left and right pathways are the same, because the lengths of time taken by the sound wave to get to the ears and by the impulse trains to get to the circuit are equal. Thus, the center detector in the circuit responds most strongly. The position of the responding detector varies as a sound source moves.

Because auditory nerve firings tend to be phase-locked to the stimulating waveforms, impulse trains contain temporal information. A computational cross-correlation model like the Jeffress model works well to detect ITDs using such impulse trains that synchronize with a particular phase of stimuli. Cross correlation is represented by

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t+\tau)dt \tag{9}$$

In this equation, x and y are impulse trains from the right and left ears and τ is the time difference between the two signals. The cross-correlation model outputs the τ which gives the maximum response. The cross correlation is usually implemented at discrete temporal intervals. If the intervals between adjacent coincidence detectors become smaller, the model will detect ITDs more accurately. Hence, the cross correlation model is a basic system for detecting ITDs [18] [16].

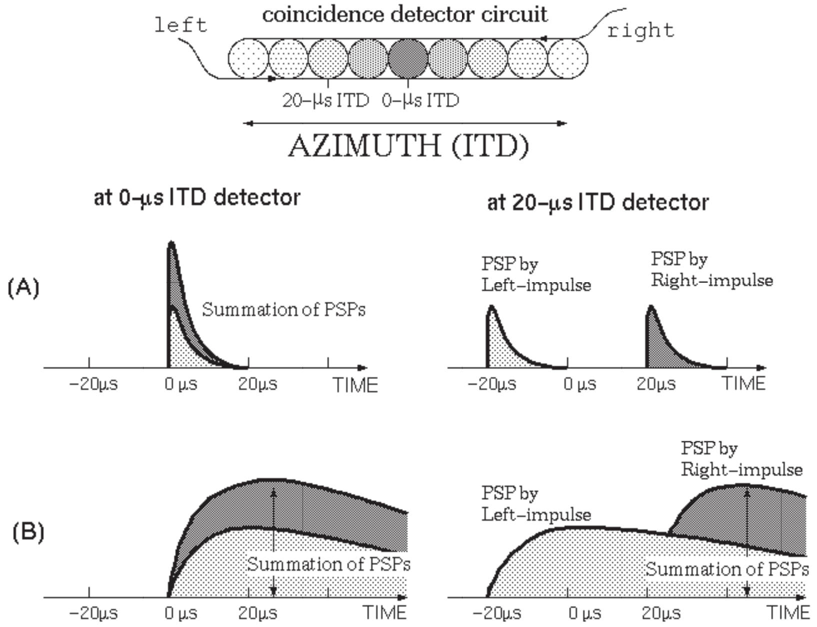


Figure 4 Temporal redundancy for short (A) and normal (B) PSPs.

3.2 Temporal Redundancy

In this study, features of the nervous system were used and the models of nerve impulses and synaptic transmission were applied to a coincidence detector circuit. The duration of the nerve impulse is over 1 millisecond and that of synaptic transmission ranges from several milliseconds to hundreds of milliseconds.

Figure 4 shows two types of temporal transition of PSPs on two coincidence detectors, corresponding to the detection of 0-μs and 20-μs ITDs. On the left side of Figure 4A, an impulse from each side is applied to the circuit without any time difference. For a human to perceive an ITD variation of about 10 μs as the MAA, it is best if the duration of PSP is shorter than 10 μs (Figure 4A). Since impulses arrive at the 0-μs ITD detector at the same time, two PSPs combine together and make a large potential. On the right side of Figure 4B, the arrival times of impulses at the 20-μs ITD detector do not match, PSPs decline without affecting each other. Thus, it is easy to distinguish the difference between the two detectors and determine ITDs.

However, the duration of PSP by synaptic transmission is several milliseconds or more (Figure 4B). At the 0-μs ITD detector, two long PSPs combine and give a large potential. Likewise, at the 20-μs ITD detector, two long PSPs combine and give a large potential, even though the arrival times of the two impulses do not match. This is because the temporal interval between them is much smaller than the duration of PSPs. Thus, the duration of PSP gives an ambiguous result that obscures minute information such as the temporal interval between the arrival times of impulses.

Figure 5 shows the temporal transition of postsynaptic potentials on the coincidence detector cells arranged along the azimuth (the axis of ITDs). An impulse from each side is

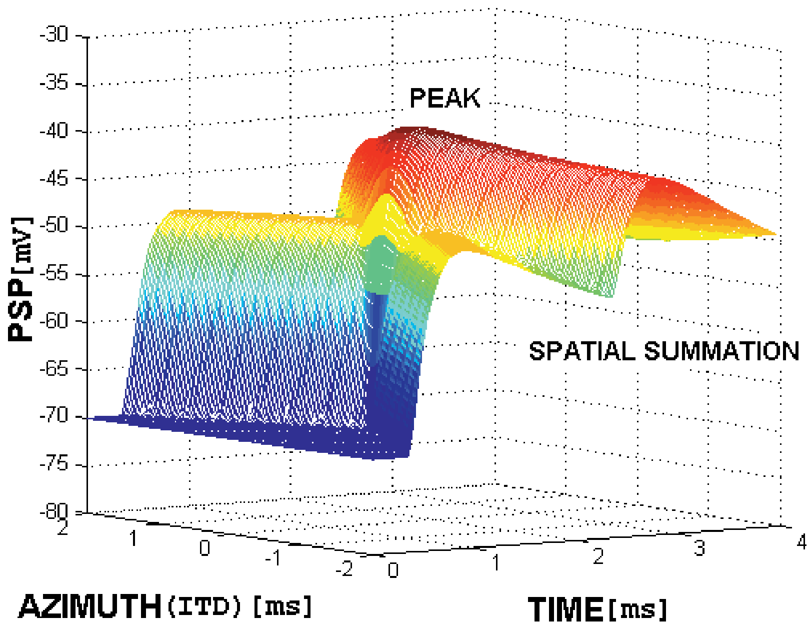


Figure 5 Postsynaptic potentials in a coincidence detector circuit. This graph indicates the temporal transition of PSPs in the model after an impulse is applied to each input of the circuit.

applied to the circuit without any time difference. The impulse from each ear stimulates coincidence detectors sequentially from each side of the circuit and a small postsynaptic potential is generated at every detector. When impulses from both sides meet at the middle detector, the PSPs are summed and a large potential is generated. Then the impulses separate and keep stimulating other detectors on the other side and summations of PSPs are generated on both sides. The envelope with the maximum potential on every detector draws a peak on the axis of ITDs. The peak looks very broad but should indicate the ITD.

Although it is not certain that threshold levels on all detectors in the MSO are the same, we assume that they are the same in this model. When the threshold level is set to the same level as the peak of the potential envelope, its simulation is equivalent to calculating the cross correlation because just one detector fires in this case (Figure 6).

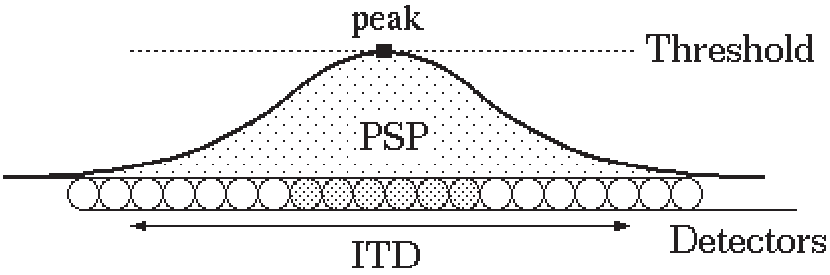


Figure 6 Threshold level at the peak of the potential envelope.

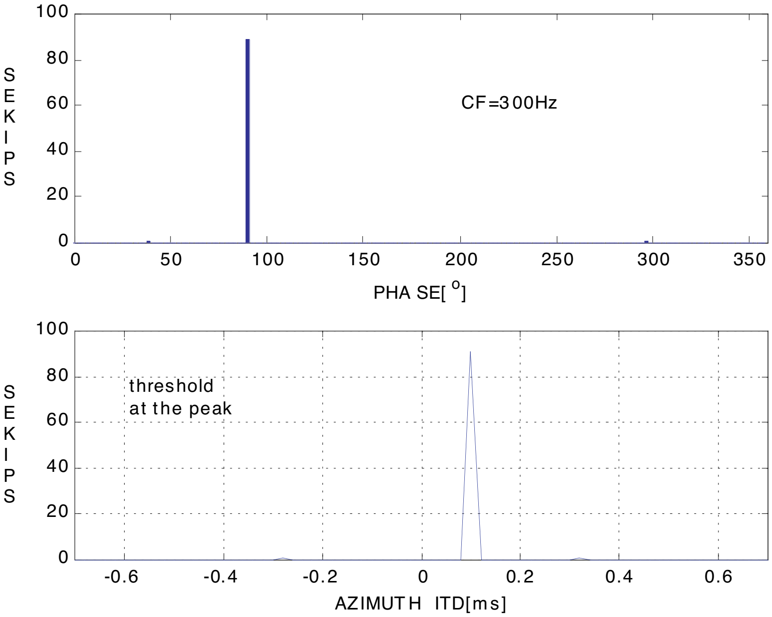


Figure 7 Period histogram of the impulse train firing in synchronization with a certain phase of stimuli and the spike histogram obtained by the simulation indicates the ITD (=100 μs) in azimuth.

Impulse trains firing in synchronization with a particular phase of a stimuli with a frequency of 300 Hz and a 0.3-s duration with an interaural time difference of 100 μs are provided as input to the model. The upper panel in Figure 7 shows the period histogram of the impulse train and the lower panel shows the spike histogram obtained by this simulation. The spikes are concentrated at an ITD of 100 μs in azimuth.

3.3 Nonlinear Output Mechanism

However, it is difficult to set the threshold level precisely at the level of the peak of the potential envelope. It is natural to set it to a level below the peak. In that case, all the detectors whose potential exceeds the threshold level fire and a broad range of firings appear along the azimuth (Figure 8). Accordingly, our model includes a nonlinear output mechanism.

Figure 9 shows the result of a simulation of the nonlinear output mechanism using the same impulse trains as in Figure 7. The upper panel in Figure 9 shows the period histogram of the impulse train and the lower panel shows the spike histogram obtained by this simulation. The nonlinear output mechanism outputs spikes over a broad range along the axis of ITD and the envelope of the spike histogram looks so square that it is difficult to determine the ITD. Although this output mechanism seems inappropriate for detecting ITDs, the output can be improved by using the variability of impulses on auditory nerve fibers.

3.4 Impulse Fluctuation

Auditory nerve fibers do not always fire in synchronization with the same phase of the stimuli; impulse trains from the auditory nerves fluctuate slightly in time (Figure 1) [6]. Since our model uses impulse trains from the auditory peripheral model that fluctuate in

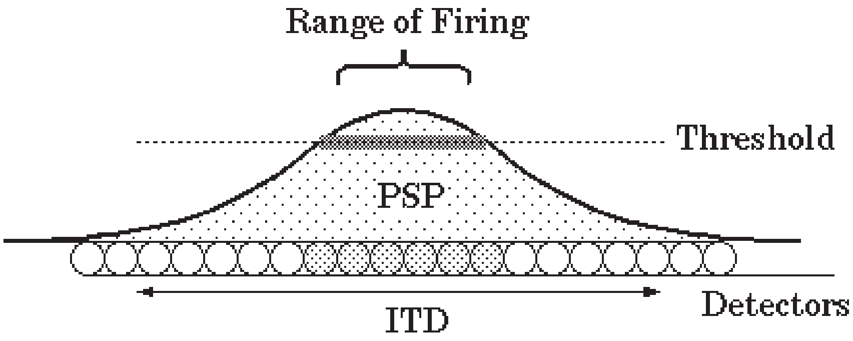


Figure 8 Threshold level below the peak of the potential envelope.

time [9], the model will be affected by the impulse fluctuation. For ITD detection by cross correlation, or the threshold level at the peak of the potential envelope in a coincidence detector circuit in particular, the fluctuation act like noise. Impulse trains having a characteristic frequency of 300 Hz with a large fluctuation in time. duration of 0.3s and with a time difference of 100 μs, are provided as input to the cross correlation model. The upper panel in Figure 10 shows the period histogram of one of those impulse trains with a large fluctuation and the lower panel shows the spike histogram obtained by this simulation. The spike histogram has some peaks but they do not indicate the ITD. Thus, it is difficult to determine the ITD.

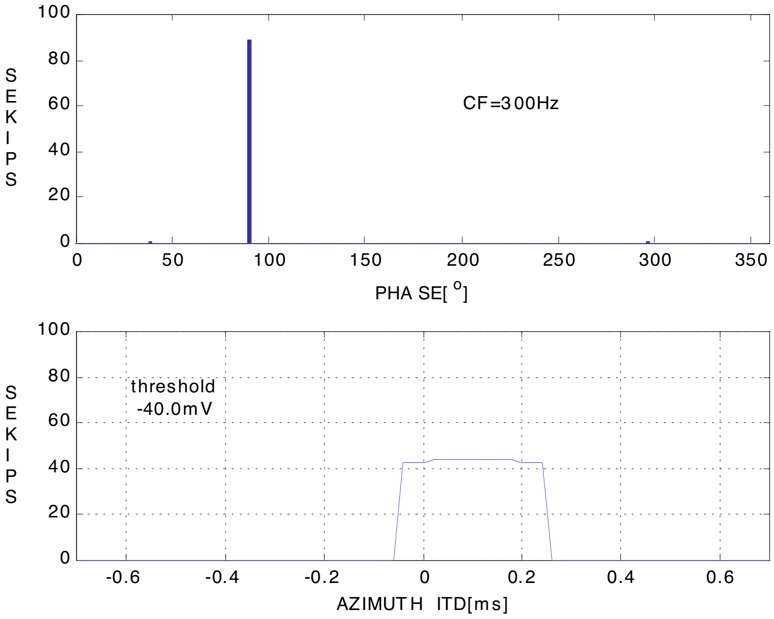


Figure 9 Period histogram of the impulse train firing in synchronization with a certain phase of stimuli and the spike histogram obtained by the nonlinear output mechanism (ITD =100 μs). The envelope of the spike histogram looks so square that it is difficult to determine the ITD

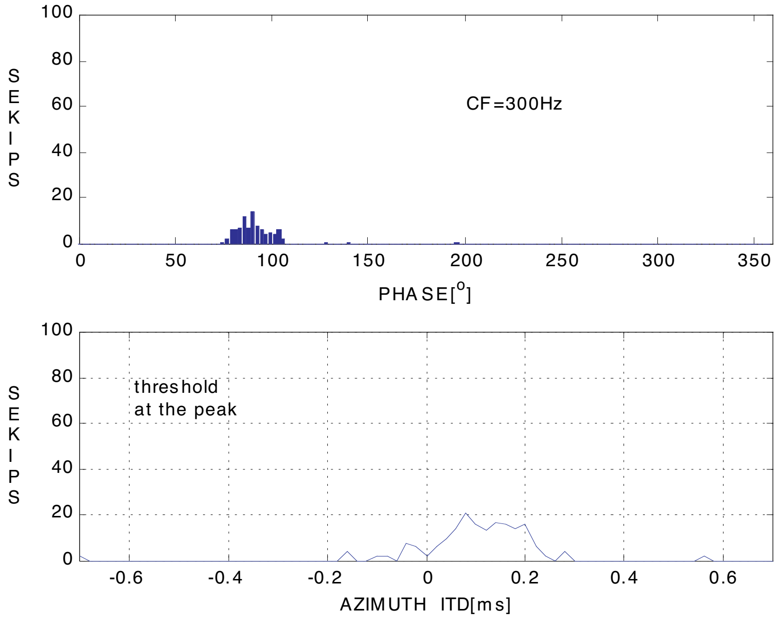


Figure 10 Period histogram of the impulse train with a large fluctuation in time and the spike histogram obtained by cross correlation (ITD =100 μs). The envelope of the spike histogram has several peaks and it is difficult to determine the ITD.

Therefore, since we have impulse trains with a large fluctuation in time as input to our model, we set the firing threshold level below the peak of the potential envelope. Impulse fluctuations affect the location where the peak of the potential envelope appears on a coincidence detector circuit because the detector where impulses from both sides of the circuit encounter each other changes based on the noise. The location of the peak of the potential envelope varies and the firing range also shifts along the axis of ITD (Figure 11A). If the threshold level is set to an appropriate level, the firing ranges will often overlap each other in spite of the variation in peak location, and the detectors in the overlapping area will keep firing. Since impulse fluctuation of an auditory nerve fiber has a normal distribution, it is likely that the actual ITD is included in the response curve.

If a spike histogram is drawn according to the variation in the firing range, it is found that the number of spikes in the overlapping area is greater than in other areas (Figure11B). This means that we use impulse trains fluctuating in time as input data improves the output of the

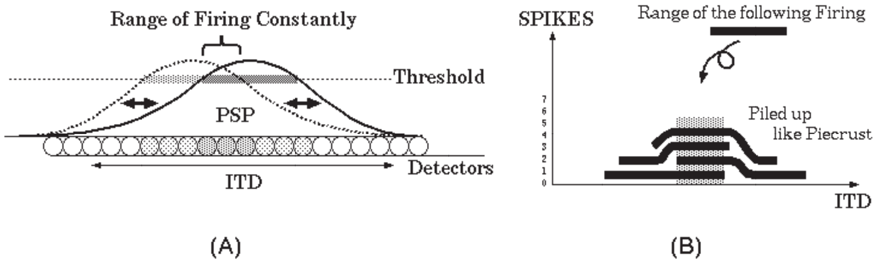


Figure 11 Nonlinear output mechanism.

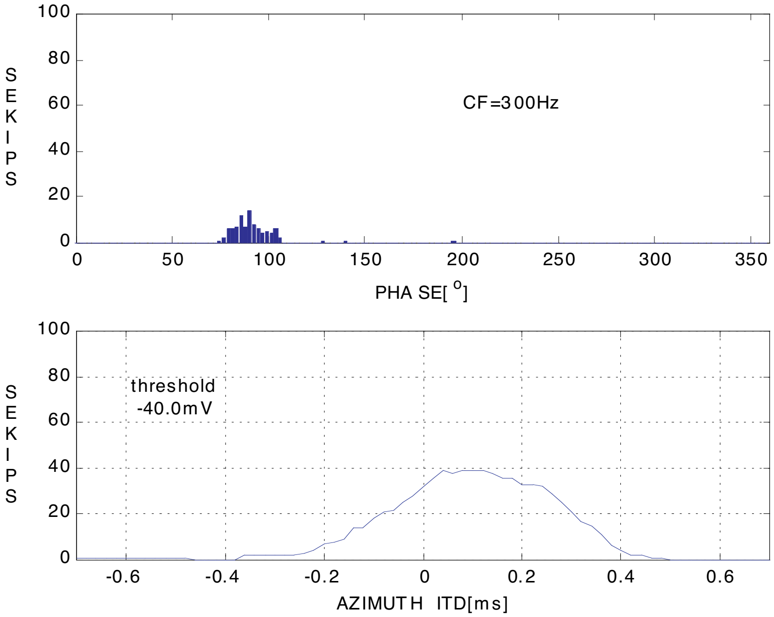


Figure 12 Period histogram of the impulse train with a large fluctuation in time and the spike histogram obtained by the nonlinear output mechanism (ITD =100 μs). The envelope of the spike histogram on the ITD rises.

model compared with using ones with no fluctuation, and the envelopes of spike histograms output from the nonlinear output mechanism tend to have a peak that indicates the ITD or its vicinity. Figure 12 shows the result of a simulation by the nonlinear output mechanism using the same impulse trains as in Figure 10. The upper panel in Figure 12 shows the period histogram of the impulse train with a large fluctuation and the lower panel shows the spike histogram obtained by this simulation. The envelope of the spike histogram as a function of the ITD is rising and beginning to form a peak.

4. Improving the Accuracy

4.1 Emphasizing

It is still difficult to determine the ITD using only one coincidence detector circuit. Actually, there are hundreds of detector circuits for each frequency in the MSO and outputs from these circuits are integrated [2] [17]. Therefore, the model can emphasize the peak indicating the ITD and achieve higher accuracy by having many circuits with different thresholds for each frequency and by integrating all the outputs. To avoid a large increase in the amount of calculation, we used multi-threshold paths for every coincidence detector instead of many circuits. Then, the outputs from all the paths were integrated. This method can emphasize the peaks of the potential to detect the ITD (Figure 13).

We used an inhibition-like model to determine the ITD more effectively. Since the coincidence detector indicating the ITD tends to fire earlier than others in the circuit (Figure 14), the inhibition suppresses the succeeding firings and thus emphasizes the initial firing. Although there are reports about inhibition in the MSO [15] [3], this inhibition-like model is

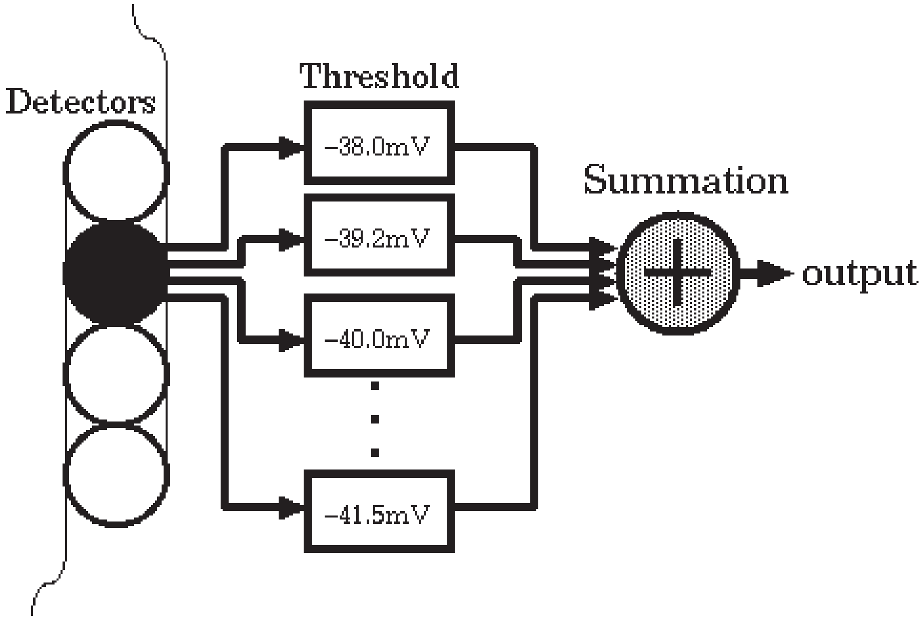


Figure 13 Multi-threshold model.

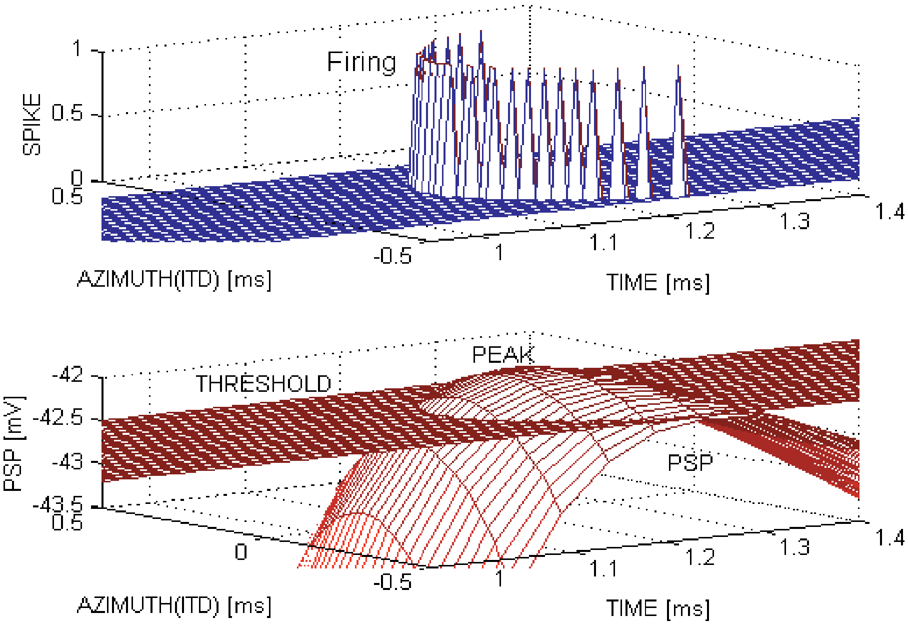


Figure 14 The detector indicating the ITD tends to fire earlier than others in the circuit.

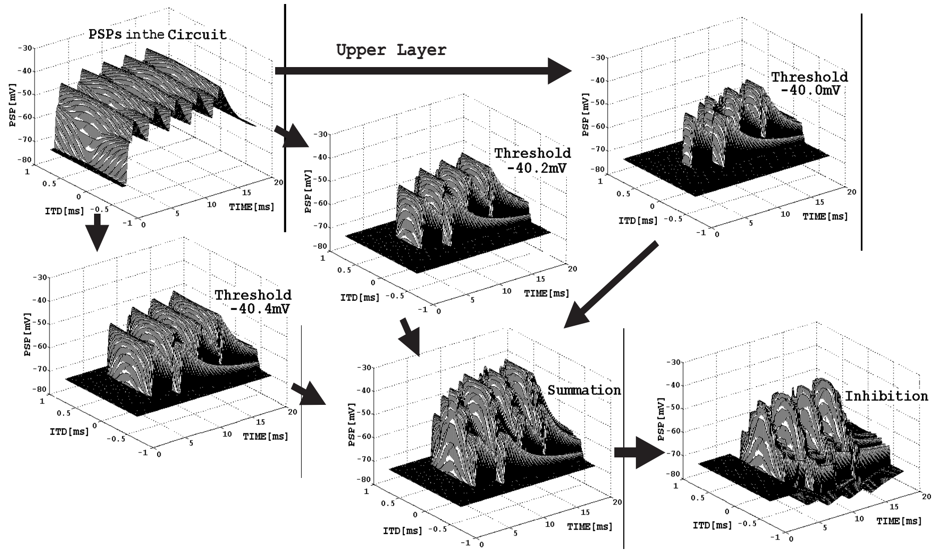


Figure 15 Outputs from the coincidence detector circuit to a upper layer. The potentials are summed by the multi-threshold model and emphasized by the inhibition-like model.

not based on physiological data. Note that this is another way to emphasize the peak and suggests the possibility of a multi-threshold mechanism. Thus, our computational localization model is constructed using models of action potentials and synaptic transmission, the multi-threshold model, and the inhibition-like model. The sound localization model outputs spikes. that indicate the best ITD.

4.2 Simulation Results

Figure 15 shows an example of outputs from the coincidence detector circuit with the multi-threshold and inhibition models. Impulse trains with a small fluctuation — having a characteristic frequency of 300 Hz and about 10-ms duration, or five impulses — were used as input data and the model output the postsynaptic potentials instead of spikes to show the effects of the emphasis. Consequently, the potentials in the detectors indicating the ITD and its vicinity were emphasized by the multi-threshold model and the inhibition-like model.

Next, we examined the potential of the auditory sound localization model to achieve greater accuracy at detecting ITDs. Impulse trains having characteristic frequency of 300 Hz, 0.3-s duration and with a time difference of 100 μs were used as input to the model. Each simulation used three types of impulse trains, which fired in synchronization with a fixed phase of the stimuli, with small and large fluctuations in time. The impulse train with large fluctuation mimics the phase locking of actual auditory nerves [9]. The results of the simulations were as follows; The upper panel in Figure 16 shows the period histogram of the impulse train firing in synchronization with a fixed phase of stimuli. The lower panel shows the spike histogram that result from this simulation. The envelope of the spike histogram has no peak and it is difficult to determine the ITD.

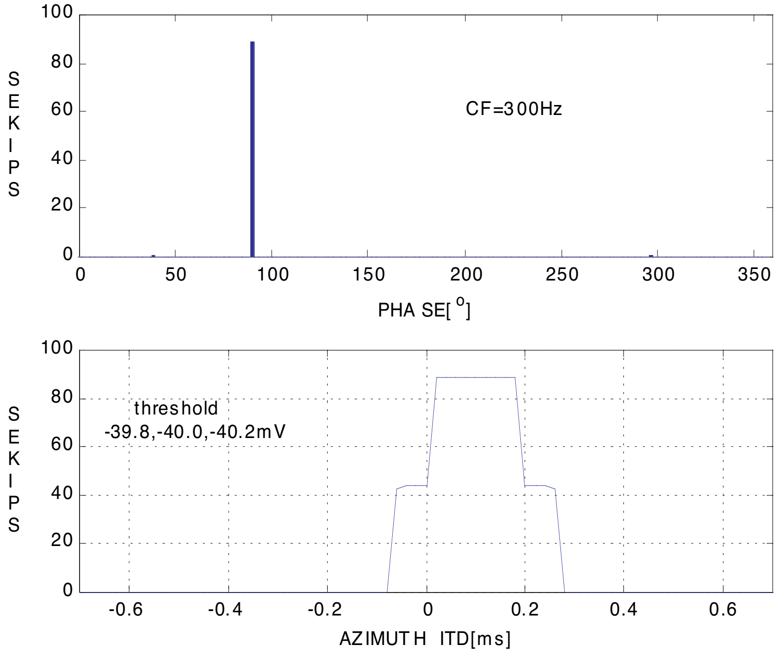


Figure 16 Period histogram of the impulse train firing in synchronization with a certain phase of stimuli and the spike histogram obtained by the simulation (ITD =100 μs). The envelope of the spike histogram has no peak and it is difficult to determine the ITD.

Impulse trains do not always keep firing in synchronization with a fixed phase of the stimuli. The upper panel in Figure 17 shows the period histogram of the impulse train with a small fluctuation in firing time. And the lower panel shows the spike histogram obtained by this simulation. The spikes are distributed around ITD of 100 μs in azimuth. However, the peak indicates the correct ITD.

The upper panel in Figure 18 shows the period histogram of the impulse train with a large fluctuation in firing time. The lower panel shows the spike histogram obtained by this simulation. Even though the spikes are distributed around ITD of 100 μs in azimuth and the envelope is smooth, the peak indicates the correct ITD.

5. Conclusions

A computational model of the auditory sound localization based on the interaural time difference was presented. Nerve impulses and synaptic transmission in the nervous system were modeled computationally and these models were applied to a coincidence detector circuit model to detect ITDs. Impulse trains with fluctuation in time were used as input data and the effects of the impulse fluctuation on the detection of ITDs were investigated.

The simulation results show that the peak indicating the ITD in azimuth obviously sharpens when impulses fluctuating in time are used as input. Using such impulse trains as input data improves the output of the model compared with using ones having no fluctuation. This suggests that impulse fluctuation can contribute to the detection of ITDs in the temporally redundant process and the nonlinear output mechanism.

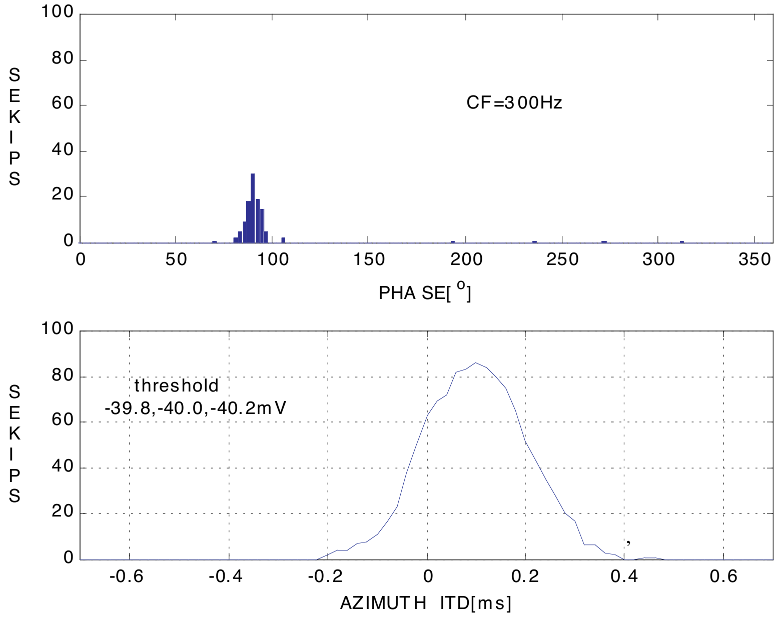


Figure 17 Period histogram of the impulse train with a small fluctuation in time and the spike histogram obtained by the simulation (ITD =100 μs). The peak of the envelope of the spike histogram indicates the ITD.

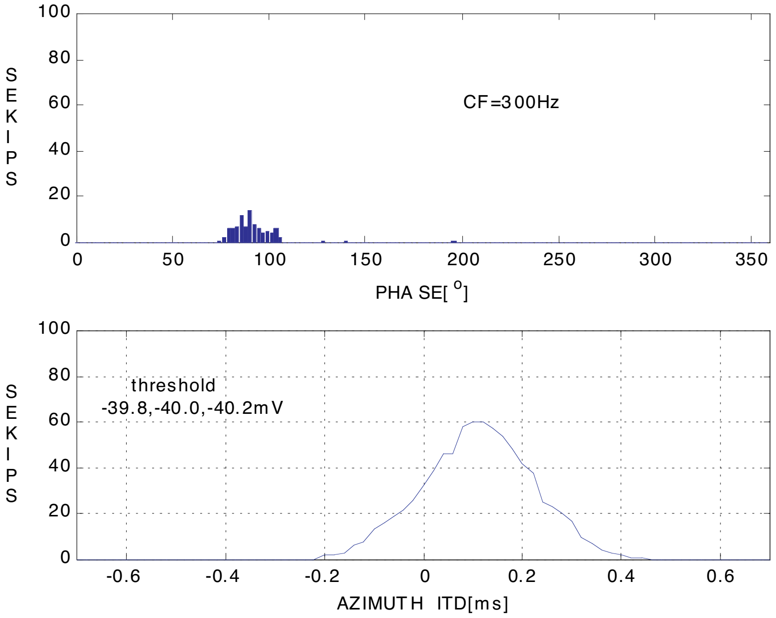


Figure 18 Period histogram of the impulse train with a large fluctuation in time and the spike histogram obtained by the simulation (ITD =100 μs). The peak of the envelope of the spike histogram indicates the ITD.

Acknowledgments

This work was supported by CREST and by Grant-in-Aid for Science Research from Ministry of Education (No.10680374).

References

[1] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. *Molecular Biology of the Cell (3rd ed.)*. New York: Garland, 1994.

[2] Carr, C. E. and Konishi, M. "A circuit for detection of interaural time differences in the brain stem of the barn owl." *J. Neurosci.*, 10: 3227–3246, 1990.

[3] Funabiki, K., Koyano, K., and Ohmori, H. "The role of GABAergic inputs for coincidence detection in the neurons of nucleus laminaris of the chick." *J. Physiol.* 508: 851–869, 1998.

[4] Hodgkin, A. L. and Huxley, A. F. "A quantitative description of membrane current and its application to conduction and excitation in nerve." *J. Physiol.,* 117: 500–544, 1952.

[5] Jeffress, L. A. "A place theory of sound localization." *J. Comp. Physiol. Pychol.*, 45: 35–49, 1948.

[6] Johnson, D. H. "The relationship between spike rate and synchrony in responses of auditory nerve fibers to single tones." *J. Acoust. Soc. Am.* 68, pp. 1115–1122, 1980.

[7] Kuwada, S., Batra, R., and Fitzpatrick, D. C. "Neural processing of binaural temporal cues." In *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson (eds.), Hillsdale, NJ: Lawrence Erlbaum, pp. 399–425, 1997.

[8] Konishi, M. "Listening with two ears." *Sci. Am.*, pp. 34–41, April 1993.

[9] Maki, K. and Akagi, M. "A functional model of the auditory peripheral system." *Proc. ASVA97*, Tokyo, pp. 703–710, 1997.

[10] Mills, A. W. "On the minimum audible angle." *J. Acoust. Soc. Am.*, 30: 237–246, 1958.

[11] Moore, B. C. J. *An Introduction to the Psychology of Hearing*. London: Academic Press, 1997.

[12] Palmer, A. R. "Neural signal processing," in *Hearing,* B. C. J. Moore (ed.), San Diego: Academic Press, pp. 75–122, 1995.

[13] Pickles, J. O. *An Introduction to the Physiology of Hearing*. London: Academic Press, 1998.

[14] Rothman, J. S., Young, E. D., and Manis, P. B. "Convergence of auditory nerve fibers onto bushy cells in the ventral cochlear nucleus: Implications of a computational model." *J. Neurophysiol.* 70: 2562–2583, 1993.

[15] Smith, P. H. "Structural and functional differences distinguish principal from nonprincipal cells in the guniea pig mso slice." *J. Neurophysiol.* 73: 1653–1667, 1995.

[16] Stern, R. M. and Trahiotis, C. "Models of binaural interaction." In *Hearing,* B. C. J. Moore (ed.), London: Academic Press, pp. 347–386, 1995.

[17] Takahashi, T. T. and Konishi, M. "Projections of the cochlear nuclei and nucleus laminaris to the inferior colliculus of the barn owl." *J. Comp. Neurol.,* 274: 190–211, 1988.

[18] Yin, T. C. T., Joris, P. X., Smith, P. H., and Chan, J. C. K. "Neuronal processing for coding interaural disparities." In *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson (eds.), Hillsdale, NJ: Lawrence Erlbaum, pp. 427–445, 1997.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

A COMPUTATIONAL MODEL OF SOUND LOCALIZATION BASED ON NEUROPHYSIOLOGICAL DATA

K. Hartung[1] and S. J. Sterbing [2]

[1]Institute of Communication Acoustics,
[2]Department of Zoology and Neurobiology,
Ruhr University Bochum, 44780 Bochum, Germany

1. Introduction

The perceived direction of an auditory event is determined largely by the sound pressure at the two ears (either humans or animals). Before sound waves, which are emitted by a sound source, reach the ears of a listener, they are shadowed, diffracted and reflected by the head, shoulders and pinna. This leads to directionally dependent changes of the spectrum at the ears. The auditory system uses these specific distortions to estimate the direction of a sound source.

Usually the cues used by the auditory system are divided into cues which only need the evaluation of the sound pressure at one ear (monaural cues) and cues, which need the evaluation of the sound pressure of both ears (interaural cues). If a sound source is positioned out of the median plane, the sound has to travel a shorter path to the ear close to the sound source and a longer way to the opposite ear. This leads to interaural time differences in the signals at the two ears. The shadowing of the head also attenuates the sound pressure at the opposite ear so that the level at the two ears is different (interaural level differences, ILD).

The directional dependent spectral distortion is described by the head-related transfer function (HRTF). The HRTF is defined as the ratio of the Fourier transform of the sound pressure at some position in the ear canal and the Fourier transform of the sound pressure in the center of the coordinate system assuming that head and body are absent [4].

The ratio of the HRTF of the left and the right ear (H_l, H_r) for the same direction is termed the interaural transfer function (ITF):

$$ITF(\omega, r, \varphi, \vartheta) = \frac{P_l(\omega, r, \varphi, \vartheta)}{P_r(\omega, r, \varphi, \vartheta)} = \frac{H_l(\omega, r, \varphi, \vartheta)}{H_r(\omega, r, \varphi, \vartheta)} \tag{1}$$

From the ITF the following interaural parameters can be derived. The interaural level difference (ILD, L_Δ) is

$$ILD(\omega, r, \varphi, \vartheta) = 20\log(|ITF(\omega, r, \varphi, \vartheta)|) \tag{2}$$

The interaural phase difference (IPD, ϕ_Δ) is

$$IPD(\omega, r, \varphi, \vartheta) = arc(ITF(\omega, r, \varphi, \vartheta)) \tag{3}$$

The interaural phase delay (IPDT) is

$$IPDT(\omega, r, \varphi, \vartheta) = \frac{IPD(\omega, r, \varphi, \vartheta)}{\omega} \tag{4}$$

The interaural group delay (IGDT) is

$$IGDT = \frac{dIPD(\omega, r, \varphi, \vartheta)}{d\omega} \tag{5}$$

The IPDT is the time delay of the fine structure of the signals and is used for low frequencies (f < 1.6 kHz). The IGDT is the time delay of the envelopes of the signals and is used for high frequencies. This distinction reflects the fact that in humans and guinea pigs the fine structure of the signal is preserved (“phase-locking“) only for low frequencies. Above that frequency limit only the envelope is coded. The monaural spectra at the ears are the sound source spectrum multiplied by the HRTF of the left and right ear, which are calculated as follows:

$$P_l(\omega, r, \varphi, \vartheta) = S(\omega) \times H_l(\omega, r, \varphi, \vartheta) \tag{6}$$

$$P_r(\omega, r, \varphi, \vartheta) = S(\omega) \times H_r(\omega, r, \varphi, \vartheta) \tag{7}$$

For a long time it has been assumed that the interaural cues determine the lateral position of the hearing event, while the monaural cues determine the perceived elevation and are necessary for front–back disambiguation (overview in [4]). These assumptions were based on psychoacoustic experiments restricted to the median plane or using headphone presentations. It was assumed that in the median plane the interaural cues were zero and only monaural cues provided information about the direction (e.g. [2], overview in [4]). Searle et al. [15] pointed out, that due to small asymmetries of head and pinna, even in the median plane ILD and ITD are dependent on the elevation and that these cues can be used for localization. Other experiments also give evidence that interaural cues might play a role for front–back disambiguation and elevation estimation [17].

In the mammalian brain, the first station for processing interaural cues is found in the superior olivary complex (SOC). For spatial hearing the medial superior olive (MSO) and lateral superior olive (LSO), and medial nucleus of the trapezoid body are important (see [10] for an overview) The MSO receives projections from the ipsi- and contralateral-ventral cochlear nucleus. It is assumed that the MSO computes the interaural time differences by some form of coincidence detection. A MSO neuron responds maximally if it is excited at the same time by an ipsi- and contralateral excitation. The inputs of the individual neurons are delayed by different lengths of fibers so that a neuron is tuned to a specific interaural time difference. The LSO receives excitatory input from the ipsilateral ventral cochlear nucleus and inhibitory input from the contralateral side via the ipsilateral medial nucleus of the trapezoid body (MNTB). The LSO neurons show the strongest response, if the ipislateral ear is excited by a higher sound pressure level or the signal reaches the ipsilateral ear earlier. All nuclei of the brainstem project to the central nucleus of the inferior colliculus (ICc). The sensitivity of ICc neurons to interaural time and level differences has been demonstrated by a number of studies (overview [10]). In free-field experiments, some form of azimuth tuning was found for 52% of the neurons (e.g. [1]). The representation of auditory space, which means the representation of the lateral position and elevation tuning could not be demonstrated with free-field stimulation. Brugge and coworkers [6][7] used virtual sound sources generated with non-individual HRTFs to investigate spatial tuning in the cat primary auditory cortex (AI). They found relatively large receptive fields (half-field, full field). It is not clear, whether this unspecific tuning can be explained by the use of non-individual HRTFs or the ability of cats to use their pinnae actively.

There have been several proposals for the modeling of sound localization. Colburn [8] gives an extensive overview of the different models. Most of the models discussed in Colburn's chapter model only the lateralization (left/right dimension) of the hearing event and are not able to explain elevation perception or front–back discrimination. The majority of the models which try to explain elevation perception and front–back discrimination use only interaural cues. Lim and Duda [12] and Martin [14], for example, use the ILD and ITD cues in different frequency bands and compare the actual parameter vector with a set of a reference vectors, which stand for the specific combination of interaural parameters for a certain direction of incidence. All studies show that the interaural cues provide sufficient cues for the estimation of the elevation and front–back disambiguation. Brainard et al. [5] accounted for the spatial tuning of neurons in the optic tectum of the barn owl by assuming that each neuron is tuned to an optimal ILD and ITD in each frequency channel. The activity of the neuron reaches a maximum, when the actual ILD and ITD match this optimal values. The superposition of the activity of different frequency band can explain the shape of the receptive fields.

2. Methods

Individual head-related transfer functions (HRTFs) of 19 guinea pigs were measured with miniature microphones (Knowles 3046) placed at the entrance of the ear canal (for details see [9]). The animals were placed in the center of an anechoic room with eleven loudspeakers mounted on an arc ranging from -10° to 90°. The ratio of the discrete Fourier transform of the test noise (random phase noise, sampling rate 50 kHz, duration 4096 samples) of the sounds measured at the two ears gives the transfer function. For each direction the ITD, ILD and monaural directivity for third-octave wide bands were computed. Virtual spatial sounds of different bandwidth (50-ms duration, 5-ms rise/fall time, 80 dB SPL) from 122 directions of the entire upper hemisphere were generated off-line and presented via individually calibrated earphones to the anesthetized (ketamine/thiazine) animal. Each position was presented five times in pseudo-random order. The single-unit activity in the central nucleus of the inferior colliculus was recorded with glass microelectrodes (impedance: 3-10 MΩ, filled with 3M KCl). The spike number for each direction was tested for statistical significant differences against the neighboring directions (Kruskal–Wallis Test). The characteristic frequency of each neuron was determined using monaural pure tone stimulation.

3. Results

3.1 Analysis of the Head-Related Transfer Functions

The head-related transfer functions of the guinea pig show asymmetries between the left and the right ear. These asymmetries are most prominent for high frequencies. Furthermore, individual differences were found. The maximum ILD value was 15 dB, maximum ITD value was 500 μs. The directivity for the left and right ear varies with frequency. The maximum is moving from lateral positions (Figure 1 a–d) at low frequencies to frontal directions at high frequencies (Figure 1 e–f).The ITD values are relatively independent of the frequency (Figure 2). The maximum ILDs are increasing with frequency and show large differences in their spatial patterns between different frequency bands (e.g. Figure 3 d,e).

3.2 Physiological Recordings

The central nucleus of the inferior colliculus receives, among others, binaural input from the medial and the lateral superior olive (MSO, LSO) as well as monaural input from the cochlear nucleus (mainly from the dorsal cochlear nucleus). Although little is known about the interaction of these sources, this convergence may be used for the combination of binaural and monaural spatial cues. The majority of neurons in the central nucleus of the inferior colliculus are spatially tuned. This has been revealed by studies using virtual sound source stimulation using white-noise signals. Single unit recordings in the midbrain of the guinea pig showed that approximately 90% of the neurons responded selectively for certain sound source directions [9] [16]. Front–rear discrimination could be shown on single neuron level.

The primary goal of the present study was to investigate, whether the spatial tuning of the ICc neurons can be explained by a superposition of physiological plausible representations of these inputs mentioned above. For that reason virtual sound source signals (VSS) of different bandwidth (white noise, one octave, 1/3 octave) were used to stimulate 46 single neurons. The center frequency of the one-octave and third-octave bands was chosen according to the characteristic frequency of each neuron. The characteristic frequencies were measured using pure tone stimulation of the contralateral ear. The majority (74%) of neurons, which were tuned under the white-noise condition, were also tuned during stimulation with one-octave VSS. The size of the receptive fields (RF) was constant with decreasing bandwidth for 51% of the neurons, increased for 35%, decreased for 7%. Another 7% of the units showed a change of RF position. For stimulation with third-octave bands only 31% of the neurons responded spatial selective. This suggests that narrow-band stimuli cannot elicit spatial tuning on neuronal level. Usually the spike rate increased for narrow-band stimuli.

3.3 Model Structure

Figure 4 shows the structure of the model. The signals are spatially filtered by the HRTF of the left and right ear. Within third-octave bands the ILD and ITD of each frequency band are calculated. It is assumed, that within a frequency band different neurons are tuned to different ITD values. A neuron responds maximally if the presented ITD matches the preferred ITD_{opt} of the neuron. σ_{ITD} determines the tuning sharpness and is the reciprocal of the neuron's characteristic frequency. The activity is modeled with

$$A(ITD, f_i) = e^{-\left(\frac{ITD_{opt}(f_i) - ITD(f_i)}{\sigma_{ITD}(f_i)}\right)^2} \qquad (8)$$

Figure 5 gives an example of the activity function for a low and a high frequency neuron. The ILD-tuning is modeled by a similar equation using a σ_{ILD} value of 10 dB.

$$A(ILD, f_i) = e^{-\left(\frac{ILD_{opt}(f_i) - ILD(f_i)}{\sigma_{ILD}(f_i)}\right)^2} \qquad (9)$$

Monaural inputs also contribute to the binaural response. The activity of each of the monaural inputs is proportional to the power in each 1/3 octave band at the left and right ear. The values are scaled in such way, that for a constant level of the sound source the values are

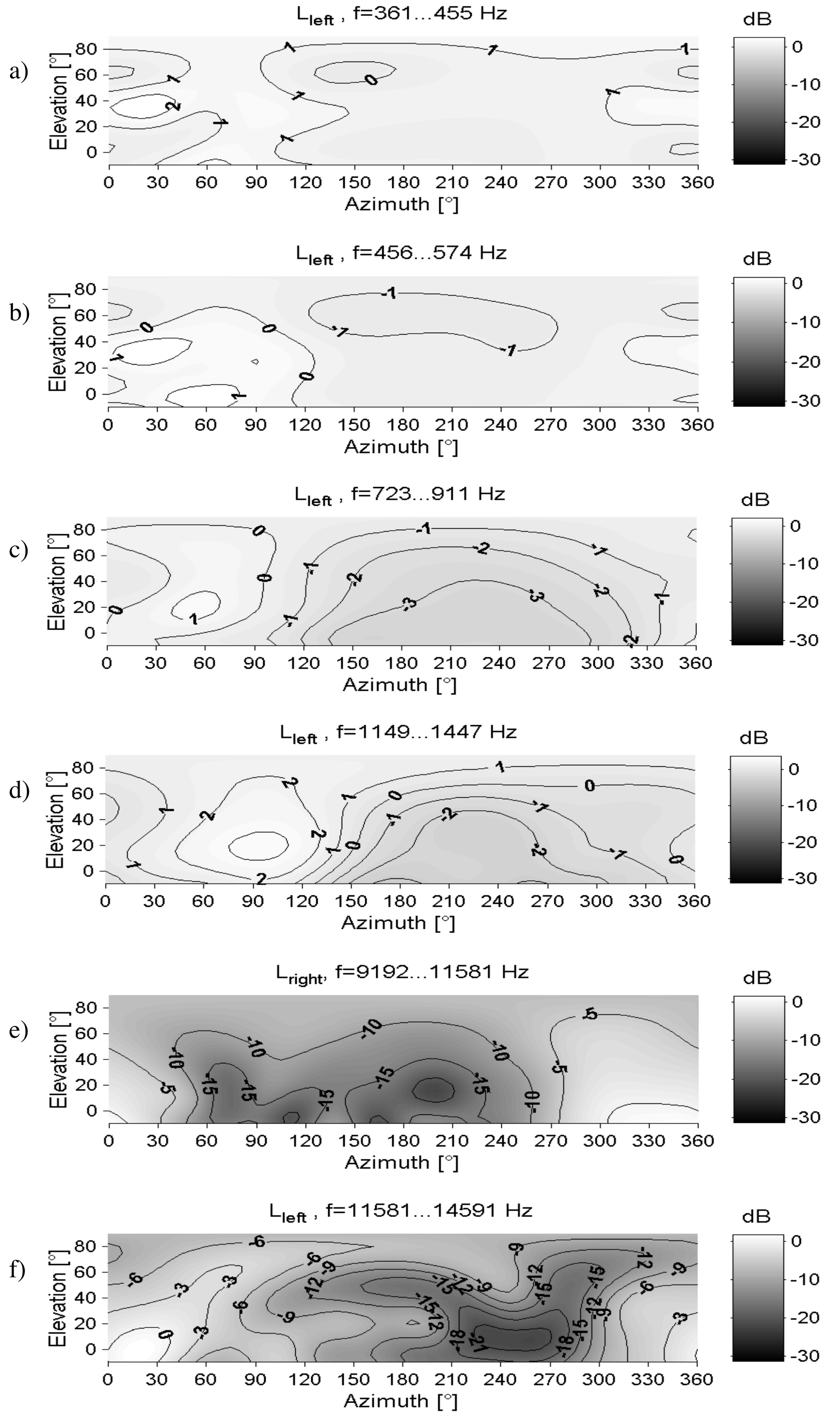


Figure 1 Directionality in different frequency bands (1/3 octave bandwidth); Panel a, b, c, d and f: left ear, panel e: right ear; the lines mark iso-level contours.

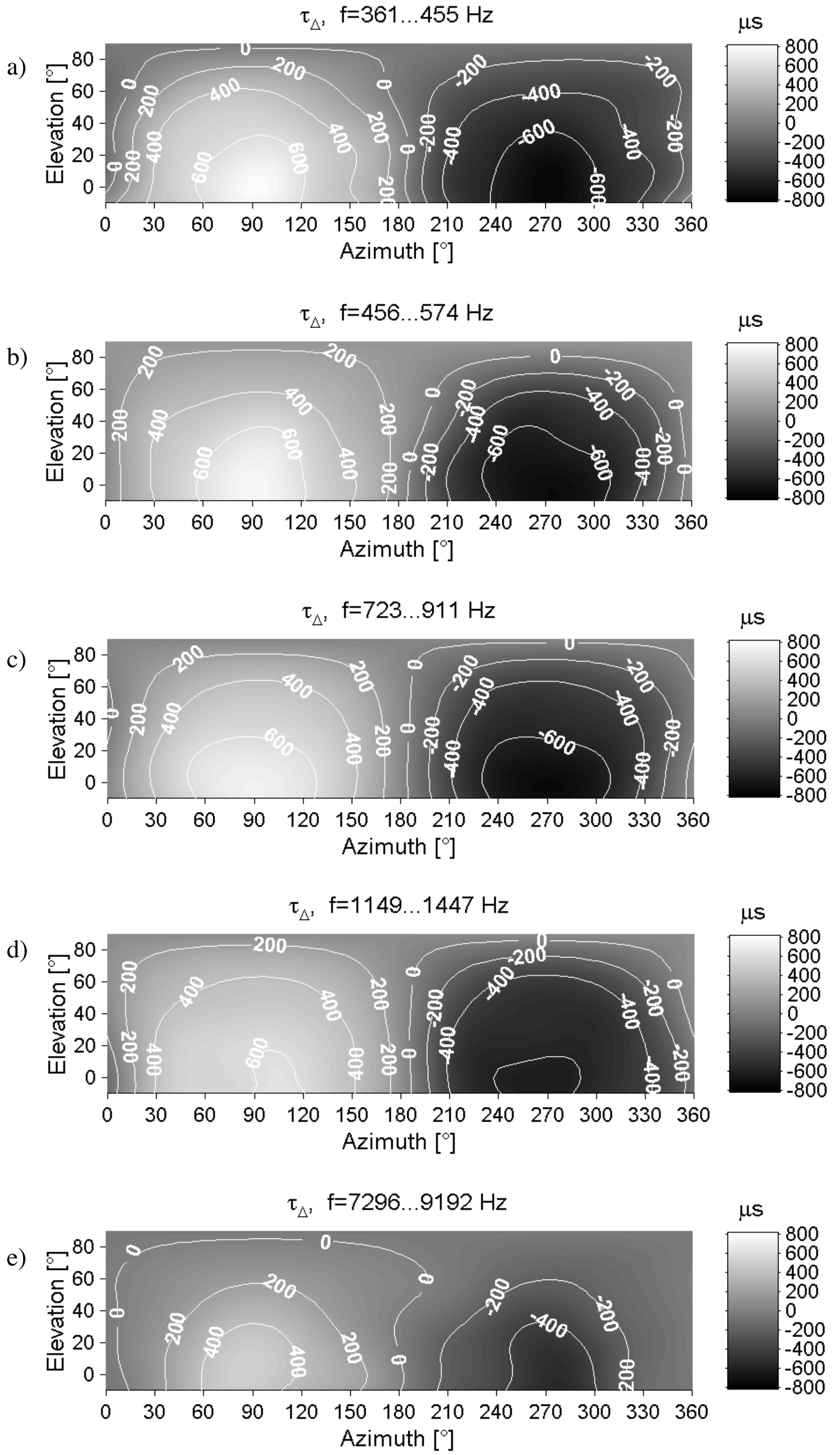


Figure 2 Interaural time differences for different frequency bands.

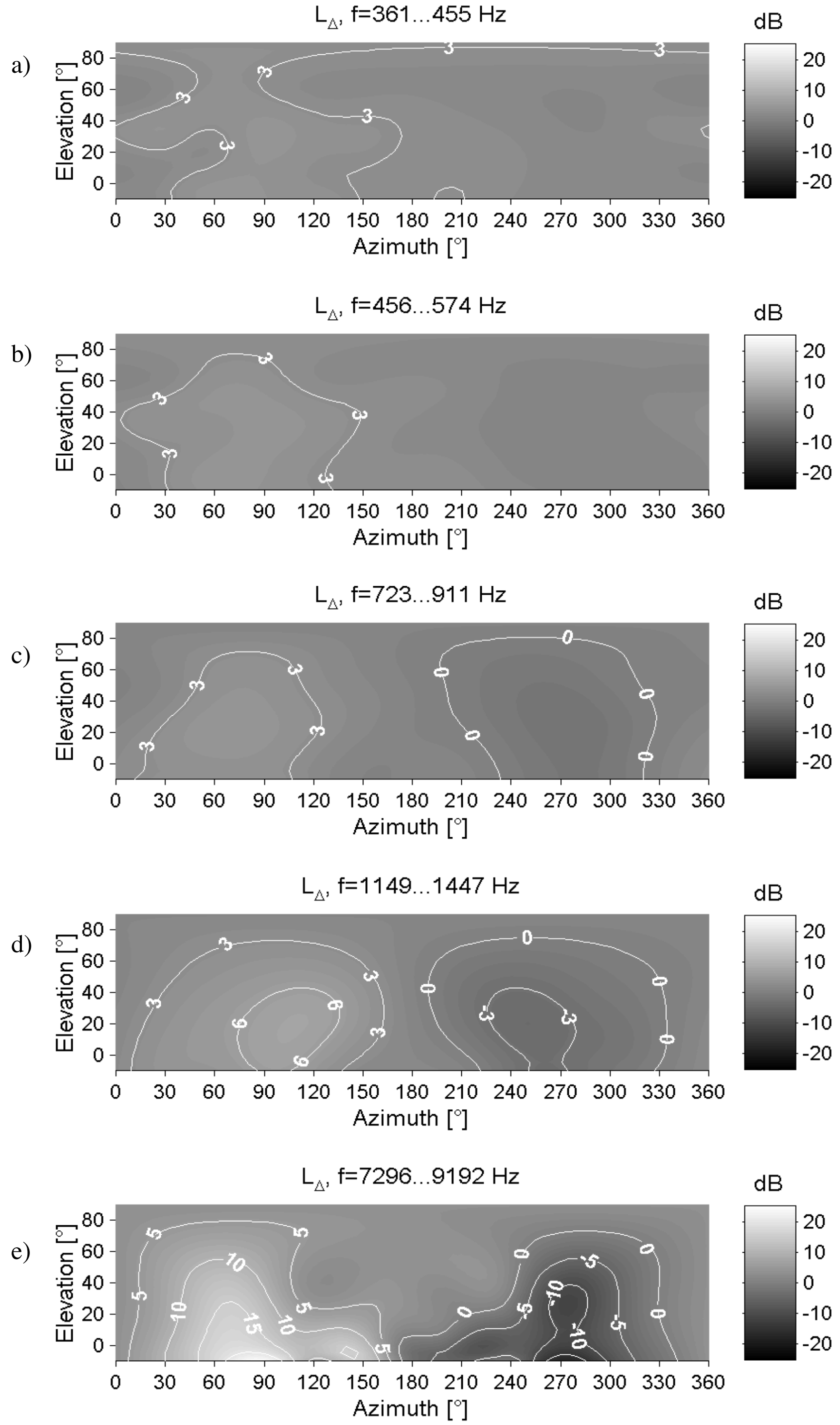


Figure 3 Interaural level differences for different frequency bands. Note that the iso-ILD contours are asymmetrical with respect to the frontal plane.

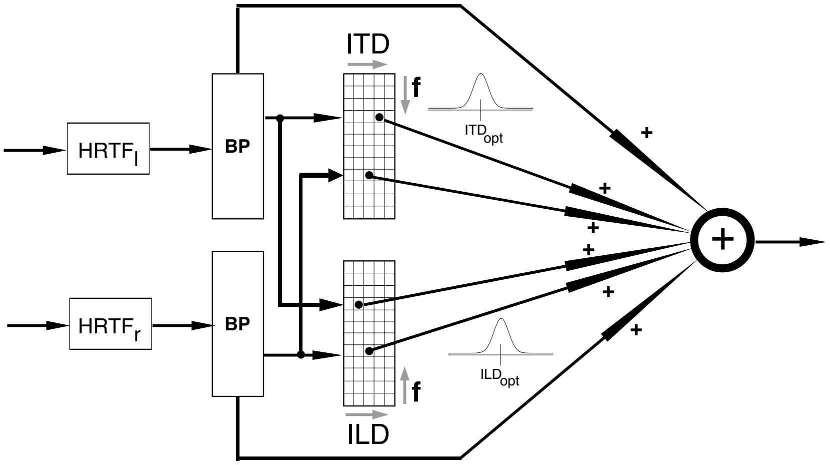


Figure 4 Structure of the model. $HRTF_{l,r}$ — head-related transfer functions, BP — Band-pass filter bank ITD — time difference processors. ILD — level difference processors.

between 0 and 1 for all frequency bands and directions. The activity of the model neuron is the weighted linear superposition of the activities of the above mentioned inputs. The model neuron can receive maximally one ITD, one ILD and one monaural input from the left and the right ear for each band, and integrates these inputs over the whole frequency range.

Figure 6 shows an example of the directional activity of a time and level difference processor. These activities reflect the shape of the ILD and ITD contours of the HRTF. The ILD activity shows prominent asymmetries between sound incidence from the frontal and rear hemisphere, while the ITD activity is symmetric around the interaural axes.

The weights of the different inputs are adjusted in order to minimize the squared error between response of the model and the measured neuronal response over all directions (Figure 7). This least square problem is constrained by the condition, that the weights should not

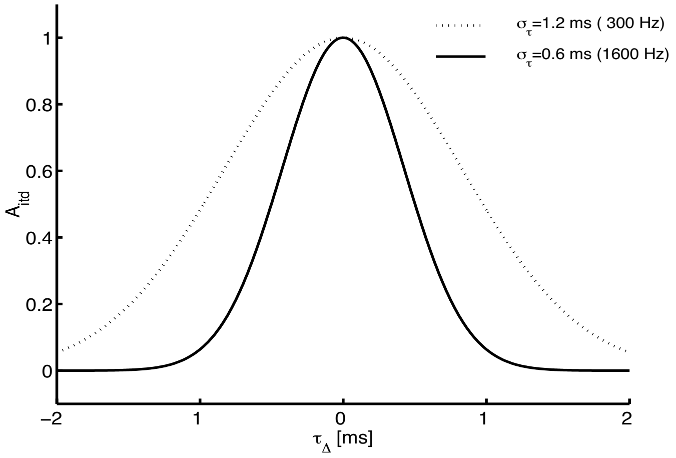


Figure 5 Two examples of ITD-tuning width.

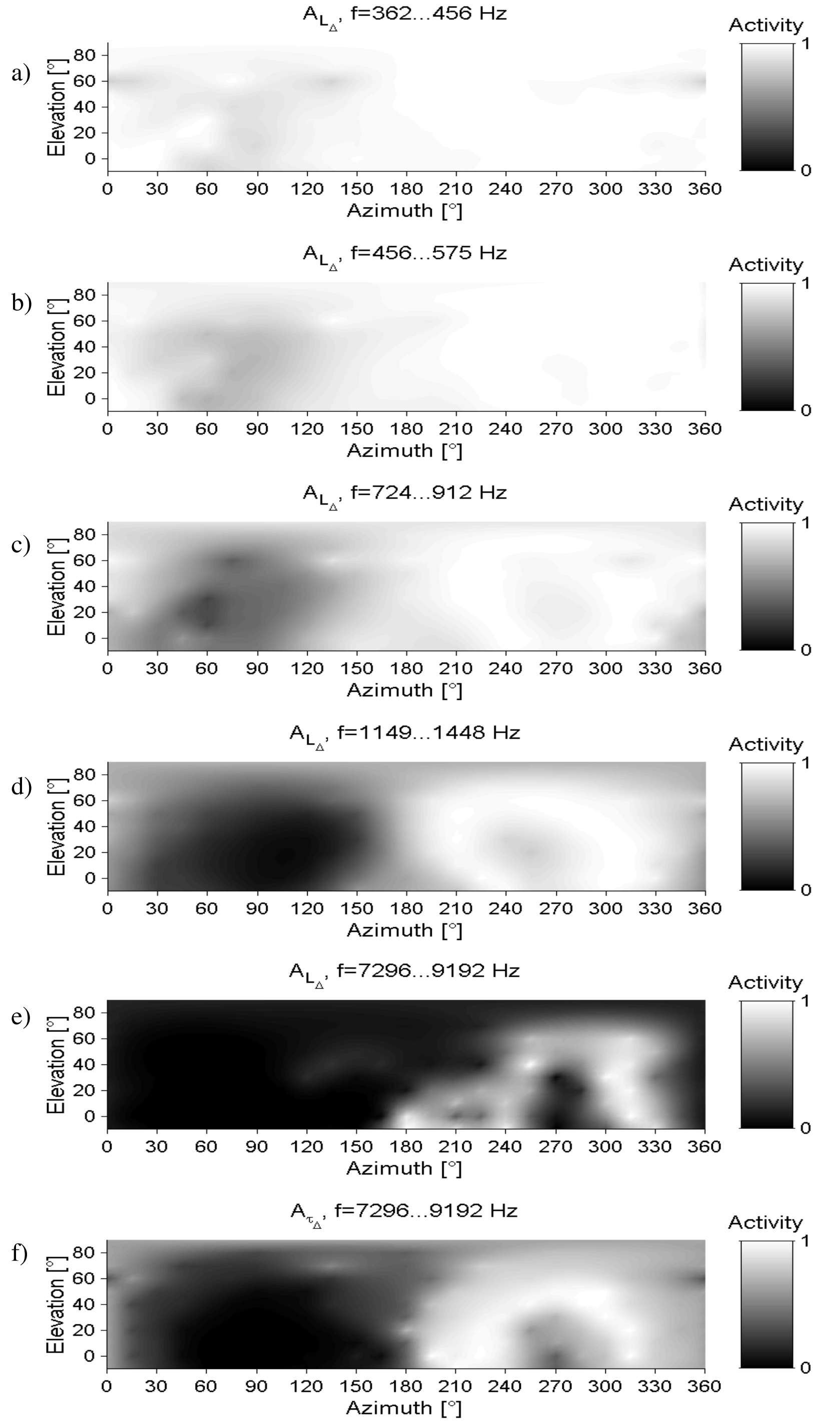


Figure 6 Examples of directional activity of ILD (a-e) and ITD (f) input. The inputs are optimally tuned to sound incidence from 315° azimuth and 0° elevation. The gray scale bar on the right side of each diagram indicates normalized activity. White areas mark directions with maximum activity.

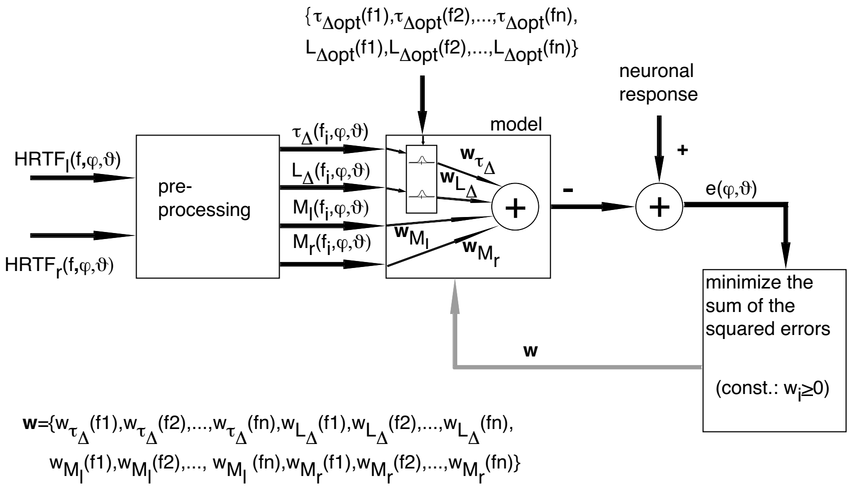

$\{\tau_{\Delta opt}(f1),\tau_{\Delta opt}(f2),...,\tau_{\Delta opt}(fn),$
$L_{\Delta opt}(f1),L_{\Delta opt}(f2),...,L_{\Delta opt}(fn)\}$
neuronal response
model
$HRTF_l(f,\varphi,\vartheta)$
$HRTF_r(f,\varphi,\vartheta)$
pre-processing
$\tau_\Delta(f_i,\varphi,\vartheta)$
$L_\Delta(f_i,\varphi,\vartheta)$
$M_l(f_i,\varphi,\vartheta)$
$M_r(f_i,\varphi,\vartheta)$
w_{τ_Δ}
w_{L_Δ}
w_{M_l}
w_{M_r}
+
-
+
$e(\varphi,\vartheta)$
minimize the sum of the squared errors
(const.: $w_i \geq 0$)
w
$w=\{w_{\tau_\Delta}(f1),w_{\tau_\Delta}(f2),...,w_{\tau_\Delta}(fn),w_{L_\Delta}(f1),w_{L_\Delta}(f2),...,w_{L_\Delta}(fn),$
$w_{M_l}(f1),w_{M_l}(f2),..., w_{M_l}(fn),w_{M_r}(f1),w_{M_r}(f2),...,w_{M_r}(fn)\}$

Figure 7 Estimation of the weights. The optimal ITD and ILDs correspond to the best direction of the measured neuron. The weights are calculated using a constrained least-square approximation.

take any negative values. The optimal ITD ($\tau_{\Delta opt}$) and ILD ($L_{\Delta opt}$) correspond to the values at the preferred direction of the neuron which is modeled.

3.4 Results of Receptive Field Modeling

The spatial activity of the model neurons fit the measured responses very well. Front-rear discrimination and the selectivity for high elevation can be modeled. The distribution of the input weights is different for all tested neurons. The modeled neurons have high weights for the inputs close to the characteristic frequency (CF). The monaural inputs and ITD inputs seem to be less important, but cannot be neglected. Figure 8 shows an example of the measured (a) and modeled (b) spatial response. This neuron responds maximally for sound from 315° azimuth and 0° elevation. Figure 9 shows the estimated inputs. This neuron has a CF of approximately 500 Hz. One can see high weights close to CF for the ILD and monaural inputs of the left ear. Additional ILD inputs from higher frequency bands (0.8, 1.3 and 8.2 kHz) and additional monaural inputs seem to resolve front/back ambiguities. Figure 8c shows summed interaural activity, which reveals clear preference for frontal directions. Figure 8(d) displays the summed activity of the interaural inputs alone, which is relatively ambiguous in terms of front–back discrimination. In many cases it is observed, that in addition to the inputs close to CF further inputs from distant frequency bands have a significant contribution to the output activity. These results are in accordance with the results of the physiological experiments with narrow-band sound sources (1/3 octave, 1 octave). It can be assumed, that the spatial tuning of ICc neurons requires the integration of monaural and interaural cues over a wide frequency range.

3.5 Localization Model

In order to test the efficiency of the representation, a localization model based on a population of many single neurons was created (Figure 10). The localization model consists of great number of model neurons, which are each tuned for different directions (resolution 5°).

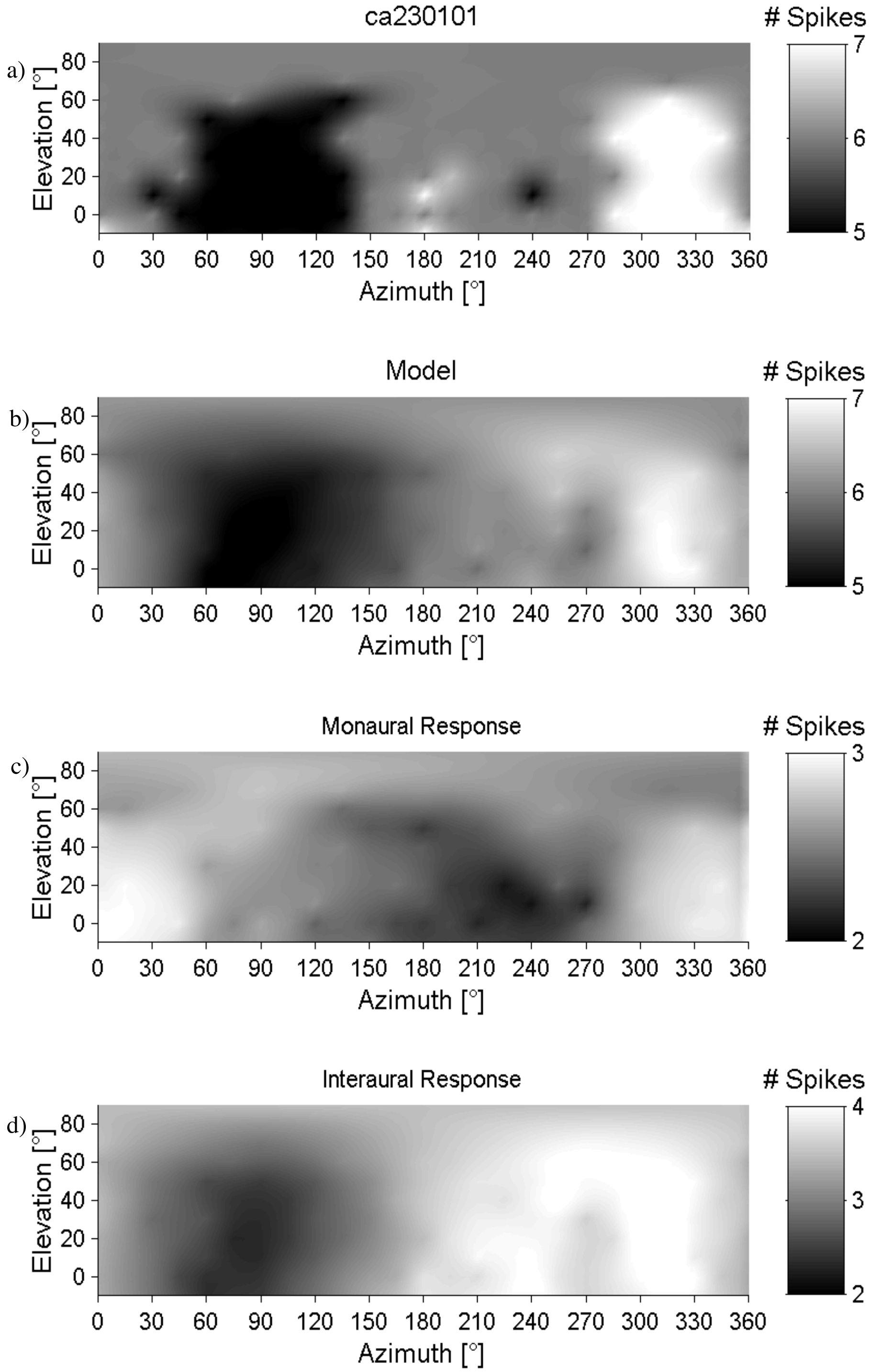


Figure 8 a) neuronal response, b) response of the model, c) response of the model to monaural cues, d) response of the model to interaural cues.

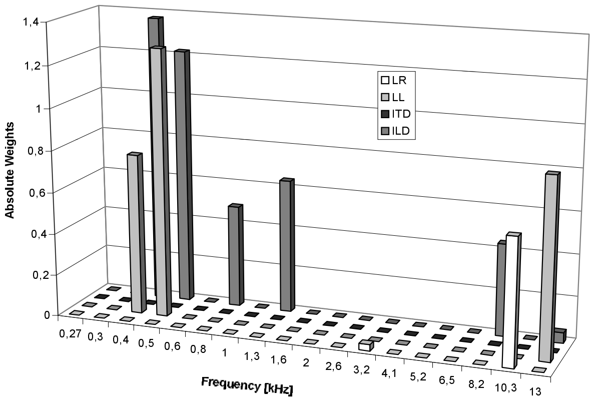


Figure 9 Estimates of the weights for each input. LR — monaural spectrum right ear, LL — monaural spectrum left ear, ITD — interaural time difference, ILD — interaural level difference.

The preprocessing and the representation of the ILDs and ITDs is the same as in the single cell model. The robustness of the representation of the interaural parameters was tested by a superposition of internal noise on the presented ILD (5 dB variance) and ITD (25μs variance) patterns.

A localization test in the horizontal plane using interaural level and time difference cues derived from a catalogue of human HRTF gave the following results: for all directions, except directly on or close to the median plane, the estimated direction matched the presented direction (Figure 11, upper row). The model responses showed more front–back than back–front confusions (Figure 11, lower row)

4. Discussion

The model simulations confirm that the interaural parameters allow a robust estimation of the direction of incidence. In contrast to the parametric models of Janko et al. [11], Lim and Duda [12], or Backman and Karjalainen [3], our approach allows a representation of concurrent, but spectrally non-overlapping sound sources. The results of the narrow-band experiments and the model simulations provide evidence for the assumption that the spatial tuning of a single neuron in the ICc requires the integration over non-adjacent bands in a relatively wide frequency range. This is different than the findings of Brainard and coworkers [5]. The authors modeled the spatial tuning of the neurons in the optic tectum of the barn owl and found that the shape of the spatial receptive fields could be explained by the superposition of the activity of ILD and ITD processors of adjacent frequency bands. They used slightly different activation functions for the interaural cues and did not consider monaural cues. Front–back discrimination was not tested in this study, because the presented directions were restricted to the frontal hemisphere. Our results suggest, that the processing in the ICc of the guinea pig is different to the processing in the optic tectum of the barn owl. It seems that in the ICc different cues are integrated in a very specific manner, which might be more robust against distractors and can explain spatial front–back discrimination. The distribution of the estimated input weights is in agreement with the results of the narrow-band experiments described above and anatomical studies [13]. It should be remarked, that these input weights are only estimates of the real inputs. Further experiments will test the estimates during electrophysiological experiments.

5. Summary

The vast majority of neurons in the central nucleus of the inferior colliculus (auditory midbrain) are spatially tuned, which has been revealed by electrophysiological studies using broadband virtual sound sources for stimulation. Based on the individual head-related transfer functions of each animal the interaural level differences (ILD), interaural time differences (ITD) and the monaural directivity were calculated in 1/3-octave bands for the upper hemisphere. It was assumed, that the neurons received input from ILD and ITD processors and from monaural pathways. The relative weights of these 72 inputs were estimated by a least-squares approximation of the neuronal response. The modeled responses were in good agreement with the measured responses. The weights were different for each of the tested neurons. High weights were found for cues close to the characteristic frequency of the neurons. Based on this single-neuron model a localization model using a population of neurons which were tuned to different directions was tested in a localization task. The model allowed a robust estimation of the direction of the sound source.

Acknowledgments

This work was supported by DFG grant Ho 450/23-1-347.

References

[1] Aitkin, L. and Gates, S. P. "Responses of neurons in inferior colliculus to variations in sound-source azimuth." *J. Neurophysiol.* 52: 1–15, 1984.

[2] Asano, F., Suzuki, Y. and Sone, T. "Role of spectral cues in median plane localization." *J. Acoust. Soc. Am.* 88: 159–168, 1990.

[3] Backman, J. and Karjalainen, M. "Modeling of human direction and spatial hearing using neural networks." *Proc. ICASSP-95*, 1995.

[4] Blauert, J. *Spatial Hearing — The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.

[5] Brainard, M., Knudsen, E. and Esterly, S. "Neural deviation of sound source location: Resolution of spatial ambiguities in binaural cues." *J. Acoust. Soc. Am.* 91: 1015–1027, 1992.

[6] Brugge, J., Reale, R. and Hind, J. "The structure of spatial receptive fields of neurons in primary auditory cortex of the cat." *J. Neurosci.* 16: 4420–4437,1996.

[7] Chen, J., Wu, Z. and Reale, R. "Applications of least-squares FIR filters to virtual acoustic space." *Hear. Res.* 80: 153–166, 1994.

[8] Colburn, H. S. "Computational models of binaural processing." In *Auditory Computation,* H. Hawkins, T. McMullen, A. N. Popper and R. R. Fay (eds.), Springer: New York, pp. 332–400, 1995.

[9] Hartung, K. and Sterbing, S. J. "Generation of virtual sound sources for the electrophysiological characterization of auditory spatial tuning in the guinea pig." In *Acoustical Signal Processing in the Central Auditory System*, J. Syka (ed.), New York: Plenum Press, pp. 408–412, 1997.

[10] Irvine, D. R. F. "Physiology of the auditory brainstem." In *The Mammalian Auditory Pathway: Neurophysiology*, A. N. Popper and R. R. Fay (eds.), New York: Springer, pp.153–231, 1992.

[11] Janko, J., Anderson, T. and Gilkey, R. H. "Using neural networks to evaluate the viability of monaural and interaural cues for sound localization." In *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson (eds.). Hillsdale, NJ: Lawrence Erlbaum, pp. 557–570, 1997.

[12] Lim, C. and Duda, R. "Estimating the azimuth and elevation of a sound source from the output of a cochlear model." *IEEE-Asilomar Conf. Signals, Systems, Computers*, pp. 399–403, 1995.

[13] Malmierca, M. S., Rees, A., LeBeau, F. E. and Bjaalie J. G. "Laminar organization of frequency-defined local axons within and between the inferior colliculi of the guinea pig." *J. Comp. Neurol.* 357: 124–144, 1995.

[14] Martin, K. *A Computational Model of Spatial Hearing*. Master Thesis. Massachusetts Institute of Technology, 1995.

[15] Searle, C., Braida, L., Cuddy, D. and David, M. "Binaural pinna disparity: Another localization cue." *J. Acoust. Soc. Am.* 57: 448–455, 1975.

[16] Sterbing, S., Hartung, K., Hoffmann, K.-P. and Blauert, J. "Auditory spatial tuning of inferior colliculus neurons in the guinea pig." *Soc. Neurosci. Abstr.* 350.7, 1996

[17] Wightman, F. L. and Kistler D. J. "Factors affecting the relative salience of sound localization cues." In *Binaural and Spatial Hearing in Real and Virtual Environments,* R. H. Gilkey and T. R. Anderson (eds.) Hillsdale, NJ: Lawrence Erlbaum pp. 1–24, 1997.

THE BAT AS A COMPUTATIONAL SYSTEM

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

THE BAT AS A COMPUTATIONAL SYSTEM

Steven Greenberg

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA

In German, the bat is known as “Die Fledermaus,” or the “flying mouse.” There is some kernel of truth to the name, as bats are, like mice, mammals, and are also capable of scampering quite swiftly in pursuit of their coming meal. But bats differ from mice in the sorts of auditory specializations used to navigate through the world. In contrast to other terrestrial mammals, bats rely heavily on acoustics to maneuver around, using the ear in a manner analogous to the way vision is used by most other species. Because of this extreme behavioral reliance on audition, bats provide a unique animal model with which to test the relation between neuronal function and sound localization.

Many species of bat use a sonar-like system for locating and characterizing objects during the course of flight. A brief (< 2 ms), high-frequency (30–100 kHz) pulse is emitted several times per second during flight, particularly around dusk when bats commonly feed on moths and other flying insects [3]. When the pulse strikes an object, a portion of the energy is reflected back to the bat. From such parameters as the acoustic time delay and Doppler-shifted spectrum the bat can deduce the object’s distance, size and trajectory. Such information can guide the bat’s flight in pursuit of a prospective meal or help avoid a collision with foliage and other stationary objects (cf. [2] for a relatively up-to-date source of information pertaining to the hearing of bats).

The two chapters in this section describe specific auditory adaptations used to guide the bat’s flight.

Wotton and colleagues investigate the temporal and spectral cues used by echo-locating bats for computing the elevation of an object in space. Some of the cues, principally those based on deep notches in the frequency spectrum (which are largely a function of pinna reflections), are similar to those used by other mammals (cf. [1]) (including humans, cf. [5]). Other cues, based on μs-time jitter in the acoustic signal, may be unique to bats. It is likely that both timing and spectral cues are used in tandem during flight.

Müller and Schnitzler view the bat echo-location system as analogous to optic flow. The latter is a conceptual framework used to quantitatively describe the set of cues and information available to the organism for visual navigation (cf. [4]). One such example is the array of visual cues associated with driving a car. What distinguishes optic flow from conventional visual information processing is the importance of time in continually updating sensory cues. This is essentially a four-dimensional space (where time is the fourth dimension) and therefore a non-trivial one to model.

The authors make a preliminary effort to model the acoustic analog of optic flow using principally temporal cues partitioned into narrow-band spectral channels (analogous to the information available to the bat). In order to constrain the problem to tractable limits, elevation is neglected. They consider amplitude and frequency modulation cues to be of primary importance and propose a neural circuit for extracting such cues. Clearly, the problem of acoustic flow is both an important and challenging one to model but will require a considerable amount of additional research over the coming years in order to provide the sort of computational methods required to predict the bat's behavior under a wide range of conditions.

The papers in this section provide two different approaches to modeling the bat auditory system. Wotton and colleagues take a relatively traditional approach in their search for time-frequency cues associated with a single component of bat flight navigation, namely computation of acoustic-object elevation. They perform controlled behavioral experiments and deduce the relevant cues from the performance limitations.

In contrast, Müller and Schnitzler attempt to model future directions in bat research, as the amount of empirical data currently available to delineate a detailed model of acoustic flow is strictly limited. Current computational methods may not quite measure up to the task of simulating the types of neuronal processing responsible for processing and interpreting the vast array of sensory cues involved in acoustic flow.

Together, the two chapters provide a representative sample of current research using quantitative approaches to study the bat auditory system.

References

[1] Musicant, A. D., Chan, J. C. K. and Hind, J. E. "Direction dependent spectral properties of cat external ear: New data and cross species comparisons." *J. Acoust. Soc. Am.*, 87: 757–781, 1990.

[2] Popper, A. N. and Fay, R. R. (eds.). *Hearing by Bats*. New York: Springer–Verlag, 1995.

[3] Tuttle, M. D. *America's Neighborhood Bats: Understanding and Learning to Live in Harmony with Them*. Austin: University of Texas Press, 1988.

[4] Weber, J. and Malik, J. "Robust computation of optical flow in a multi-scale differential framework." *Intern. J. Comp. Vis.*, 14: 5–19, 1995

[5] Wightman, F. L. and Kistler, D. J. "Sound localization." In *Human Psychophysics*, W. Yost, A. Popper and R. Fay (eds.), New York: Springer–Verlag, pp. 155–192, 1993.

TIME AND FREQUENCY INFORMATION USED IN THE COMPUTATION OF ELEVATION BY AN ECHOLOCATING BAT

Janine M. Wotton[1], Michael J. Ferragamo[1], Rick L. Jenison[2] and James A. Simmons[3]

[1]*Department of Biology, Gustavus Adolphus College, St Peter, MN 56082, USA*
[2]*Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706, USA*
[3]*Department of Neuroscience, Brown University, Providence, RI 02912, USA*

1. Introduction

Mammalian pinnae act as spatially dependent filters for incoming sound (e.g. human — [22][29][30][31][40][41]; cat — [24][26]; ferret — [5]; bats – [9][42]) and the information produced by this filtering is used (at least by humans) to localize the elevation of sound sources [19][40]. The acoustic signal can, in principle, be described using either time or frequency parameters. Either representation could convey the information required for accurate sound localization.

The first detailed study of external-ear cues was based on a time-domain representation of pinna reverberations [1][2] and a number of subsequent analyses have adhered to this approach [6][16][45]. Impulse responses of human [16] and bat [42] external ears vary systematically with sound-source position, and the number and timing of reflected components in the sound reaching the eardrum could provide cues for sound location.

Most studies of the spatial-filtering properties of the external ear have represented localization cues as systematic spectral changes in the acoustic transfer function (measured at the eardrum) with changes in sound-source position [40]. The particular features of the sound spectrum responsible for providing mammalian vertical information are subject to controversy. Some authors stress the importance of spectral peaks [3][17], while others view spectral notches as significant [4][15][26]. Some authors treat all spectral changes as equally important [40][22].

Spatial information can be described as events in the time waveform or changes in spectral magnitude. However, the way in which the brain encodes localization cues is not known. Frequency-based descriptions have been more commonly used because the auditory system is organized in a tonotopic fashion. Neurons are tuned to different frequencies and thus are able to represent spectral notches and peaks. The cochlea extracts the locations of spectral notches and peaks from the amount of excitation occurring at different frequencies and the neural activity at virtually all levels of the auditory system mirrors the spectrum of the acoustic signal. However, this does not necessarily mean that the distribution of neural activity across tuned frequencies is the representation of elevation. At some stage of processing it seems likely that a representation of elevation in spatial coordinates emerges and it is possible that the auditory system's peripheral frequency representation of peaks and notches could easily be transformed into a time-domain metric.

If the coding at the periphery is spectral, as is generally thought, then how might these cues be interpreted and utilized? In order to address these questions we need to determine the spectra received at the eardrum and then conduct behavioral experiments that reveal how

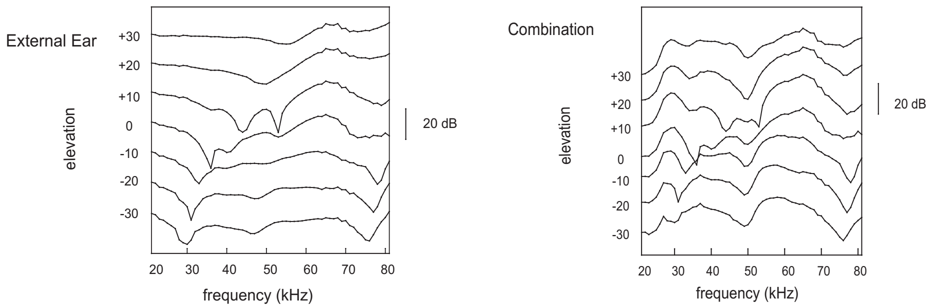


Figure 1 The magnitude spectra of the transfer functions from one cadaver ear of *Eptesicus* are shown in the left panel for seven different sound-source elevations (+30° to -30°) at 0° azimuth [42]. The right panel shows the echolocation combination magnitude spectra created by convolving these ear functions with emission spectra recorded at the same locations [43].

the cues may be processed. Echolocating bats depend entirely on the auditory system to localize targets and thus make an ideal model animal for the study of sound localization. The potential spectral cues received by an echolocating bat, as well as two relevant behavioral experiments, are presented in this chapter. Bats perform either spectral-discrimination tasks or temporal-discrimination tasks while the position of the sound source is varied. In both cases discrimination performance varies with loudspeaker elevation in a manner that conforms to predictions based on the spectra of the external-ear transfer functions.

2. Spectral Cues for Bats

2.1 Combination Spectra

Interpreting changes in spectra is potentially problematic for many animals because the source spectrum is usually unknown [21][41]. However, echolocating animals emit a sound that probes the environment and returns as an echo; thus, both the original source spectrum and the changes in the spectrum are known. For example, the big brown bat, *Eptesicus fuscus*, emits broadband, frequency-modulated (FM) echolocation signals and uses information contained within the echoes to find insects. Echolocating bats acquire information that is a combination of the properties of the sound they emit and the sound received at the eardrum. The external ears of bats impose spatially dependent spectral cues on the echoes [8][10] [11][18][25] [42; cf. Figure 1]. The signal emitted by bats also has directional properties [12][13][14][28] and therefore must be considered in determining the directional information available. Potential localization cues are contained in the combination spectra produced by convolving the magnitude spectra of the emission and the external ear transfer functions. Changes in sound source position result in systematic spectral changes in the combination spectra [43: cf. Figure 1].

The left panel of Figure 1 clearly shows a notch (local minimum) in the magnitude spectra of the external ear transfer functions, which decreases systematically in center frequency (from about 50 kHz to 30 kHz) with decreases in elevation. This notch is referred to as the "primary" notch. A prominent peak (local maximum) between about 60 and 70 kHz, (referred to as the "main" peak hereafter) is visible for the higher elevations. Both the primary notch and main peak are present for a restricted range of elevations. The right panel of

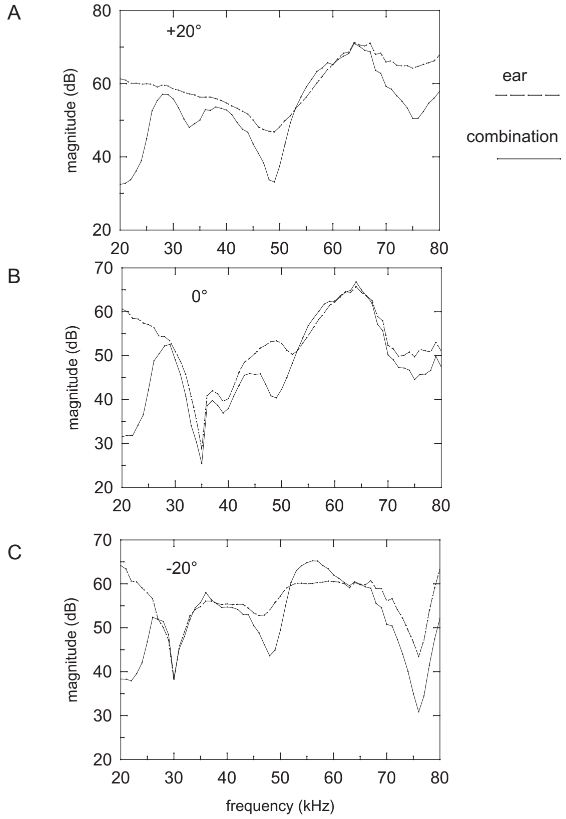


Figure 2 The magnitude spectra of the echolocation combination compared to the magnitude spectra of the ear. To facilitate direct comparison, 60 dB was added to the ear functions at elevations +20° and 0° and 65 dB was added to the function at -20° [43].

Figure 1 shows that the primary notch and main peak have been maintained in the combination spectra. The signal received by the bat has elevation-dependent notches and peaks that could provide cues for localization.

2.2 Spectral Cue Enhancement

Incorporating emission information into the spectra received by the bat may improve localization because the peak and notch information appear to be enhanced in the combination spectra (Figure 2). For all three examples shown in Figure 2, there is a large difference between the spectra in the range of 45 to 55 kHz, with the magnitude of the echolocation combination dropping sharply at ca. 50 kHz. As a consequence, in the combination condition the intensity difference between this frequency region and the main peak is greater and the rate of change in magnitude is increased. The contrast between the primary notch and the surrounding frequencies is enhanced by the sharpening of small, adjacent peaks on either side of the notch.

In echolocation, combination changes in magnitude created by the external ear are imposed on an FM emission. Rapid changes in level created by steep gradients between peaks and notches are imposed on the temporal structure of the emission. The ability of humans to detect glides in frequency or amplitude is very similar, but thresholds tend to be smaller when both types of glides are combined [7][23]. It is possible that the combination of amplitude-modulated and frequency-modulated signals produced in the echolocation combination improve the detectability of the spectral cues.

3. Spectral Discrimination

Behavioral experiments are necessary to reveal the influence of external-ear filtering on the perception and representation of spatial information. Notch information, which has been shown to influence human spatial perception [4][15] and which appears to be important for cat sound localization [26], may also serve as a cue for vertical localization for bats. The echolocation combination spectra of *Eptesicus* contain a primary spectral notch that decreases in center frequency as the elevation of the sound source is decreased. This primary notch is an obvious feature in these transfer functions (the depth usually is 15-20 dB) and appears to provide information about the elevation of a sound source located near and below the horizon [43]. Distortion of the tragus of *Eptesicus* causes a disruption of the bat's vertical angle discrimination [20] and disrupts the systematic changes in primary notch information [42] [43].

A behavioral experiment was designed to examine the influence of sound-source location on the discrimination of spectral notches by bats [43]. Under free-field conditions an echolocating bat broadcasts a sound which it receives several milliseconds later as an echo from a nearby object. If a sound similar in spectral composition and FM structure to the bat's broadcast is delivered at a predetermined delay then the bat seems to accept this as an "echo" and perceives a "phantom" target [32][33]. Bats were trained to discriminate, in a two-alternative, forced-choice paradigm (2AFC) between two computer-synthesized sounds, one with a spectral notch and the other without. The external ear of the bat introduced a spectral notch in both of these signals in addition to the synthesized notch. Effectively, the bats were discriminating between a signal with two notches (one synthesized and the other introduced by the ear) and a signal with only one notch (introduced by the ear). The vertical location of the loudspeakers playing the sounds was changed daily, shifting the frequency of the notch introduced by the ear. At one particular loudspeaker position the external ear notch should coincide in frequency with the synthesized notch; then both signals would each contain a spectral notch at the same frequency. The notch discrimination task should be difficult for the bat to perform at that loudspeaker location. When the sound source is at any other location the bat should be able to perform the discrimination because the synthesized and external ear notches will be at different frequencies.

Changing the frequency of the synthesized notch should change the vertical location at which the synthesized and ear notches coincide. A decrease in the frequency of the synthesized notch should result in a decrease in the elevation of the loudspeaker position at which the task was difficult. The bats' performance followed this prediction based on the external - ear transfer-function information. A decrease in the frequency of the synthesized notch resulted in a decrease in the loudspeaker position that caused confusion for the bats. Figure 3 shows that for each sound source elevation there is only one frequency notch that produces below-threshold (75% correct) performance. The frequency notch that was poorly discriminated systematically decreased as the sound source elevation decreased [43].

The no-notch and the synthesized notch signals are represented in the central auditory pathway by different excitation patterns of frequency-tuned neurons and thus the bat can distinguish between them. Changing the elevation of the sound source alters the bats' discrimi-

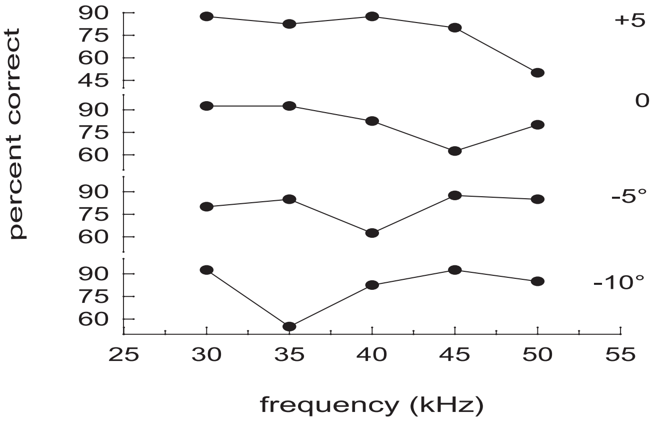


Figure 3 Percentage of correct responses averaged for two bats at four different sound-source elevations as a function of each of the frequency notches tested [44].

nation because the two signals have virtually indistinguishable patterns of excitation at one sound source location. The signals the bat receives have two sources of spectral information — the synthesized spectra and the spectra introduced by the ear. The bat is apparently unable to separate these two independent sources of information and therefore the discrimination is affected.

4. Temporal Discrimination

4.1 Jitter Experiment

An experiment paradigm that examines the perception of very small differences in the delay of echoes has revealed the remarkable temporal acuity of *Eptesicus* [33][34][35][37]. In this "jitter paradigm" bats are trained in a 2AFC task to discriminate between two stimuli [33]. The unrewarded stimulus is a signal delivered to the bat at a fixed delay in response to an echolocation emission. The rewarded stimulus is a signal at one of two different delays and these alternate with each emission (jittered signal), the difference between these two delays is varied over a set of trials.

Changing the elevation of the sound source was the manipulation for this particular experiment. The usual loudspeaker position for this experiment was designated 0°, and the position of both loudspeakers was changed by either 15° above, or 15° below the usual position. Behavioral results of one bat are displayed as the percentage of errors at each delay difference in Figure 4A. The error curves are reflected about the zero-delay difference.

For all loudspeaker elevations there is a peak in the error at zero-delay difference because the two signals are indistinguishable. There is also a side peak of errors occurring at a different delay for each loudspeaker location. The error side peak for this bat at the 0° elevation condition occurs at about 35 μs, and this performance curve has a characteristic shape (with a side peak typically between 30 to 35 μs) that resembles the cross-correlation function of the emission [33] [35] [37].

In the time-domain representation the impulse response shows the reverberation created within the external ear by the multiple sound paths to the eardrum. The impulse has a peak corresponding to the arrival of the incident sound by the most direct path followed by one or

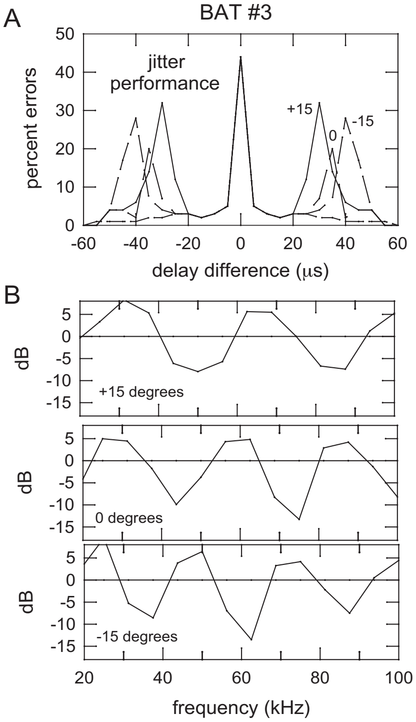


Figure 4 A. Performance of Bat #3 in the jitter experiment at three loudspeaker elevations (+15°, 0°, -15°). The percentage of errors is shown at each delay value tested and reflected around the zero-delay value [38]. B. Amplitude spectra generated for each elevation using the error curves as the time signal.

more delayed versions of the incident sound that reflect from different surfaces of the external ear. (In the frequency domain spectral notches are created from an interference pattern associated with the overlap of signals at a time separation indicated by the impulse response). The time separation of the primary and secondary impulses is determined by the elevation of the sound source [1]. In *Eptesicus* the time separation between the primary and secondary impulses increases smoothly from ca. 25 μs to ca. 40 μs as the vertical angle falls from +15° to - 40° (measured with 0° at the eye-nostril plane) [42].

The behavioral experiment was conducted on a sloped platform and the loudspeaker location, designated 0°, was actually about -10° relative to the eye-nostril plane. The time separation of the two impulse peaks for the external ear filtering at this elevation is about 30 to 35 μs. If the external ear and the emission signal share the same properties described by the impulse responses (or the corresponding spectra) then perceptual effects attributed to the emission signal may actually be due to the external ear or to a combination of emission and ear. If the external ear's reverberation time contributes to the jitter discrimination curve the side peak should either shift to an earlier delay of approximately 25 to 30 μs for the loudspeaker position designated +15° (+5° relative to the eye-nostril plane), or shift to a later delay of approximately 35 to 40 μs for a loudspeaker position designated -15° (-25° relative to the eye-nostril plane). Figure 4A shows that this prediction was borne out. In the behav-

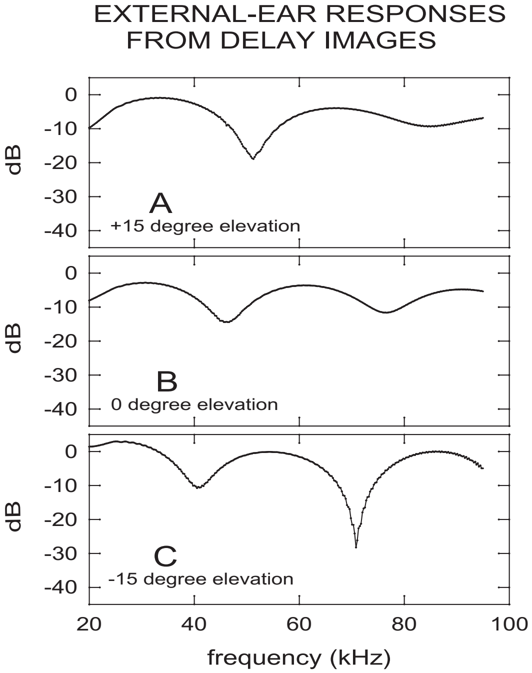


Figure 5 The spectra at each loudspeaker elevation were reconstructed using a weighted-emission signal. The incident echo was then subtracted from each of the reconstructed spectra. The difference spectra are shown for each loudspeaker elevation.

ioral function, the delay of the side peak relative to the main peak corresponds to the delay of the external ear reverberation, and this shift is in the appropriate direction when the elevation of the loudspeaker is changed.

4.2 Spectral Conversion

Each behavioral function in Figure 4A is equivalent to the impulse response of the bat's performance. To view the behavior in the frequency domain, the error curves have been converted into a spectral representation. Each panel in Figure 4B displays the amplitude spectra generated for a loudspeaker elevation using the error curves as the time signal. Broad spectral notches are visible in each function. The center frequency of the first notch changes from ca. 50 kHz to ca. 35 kHz as elevation decreases.

One echolocation emission was introduced to obtain smoother spectra. An emission signal (2-ms duration) was aligned at each of the delays provided by the data points for the jitter performance (Figure 4A), these signals were then added together to reconstruct spectra at each loudspeaker location. The amplitude of the emission signal was weighted to be proportional to the error performance for each delay value. The incident echo measured at 0° relative to the eye-nostril plane was subtracted from each function. The difference spectra are shown in Figure 5 for each loudspeaker location. Each function has two spectral notches that vary systematically in center frequency with elevation. The notch frequency decreases as the elevation decreases. This spectral pattern corresponds to the filtering created by the bats' external ear and is maintained in the echolocation combination (see Figure 1).

5. Conclusions

Previous studies have shown the importance of external ear filtering for locating the elevation of sound sources [1][22][31][39]. Cues could take the form of elevation-dependent spectral changes displayed in transfer functions or changes in the time separation of peaks in the impulse responses. Models have focused on understanding mechanisms that encode sound-source elevation along the frequency dimension in the central auditory system [22] [39][40][3][15]. The notch/no-notch experiment showed that spectral changes imposed by the external ear are important to bats in localizing the elevation of sounds. However, the jitter paradigm reveals a similar influence of sound-source location on discrimination that is based on the bat's perception of time. This experiment suggests that once the initial coding is parceled into parallel frequency channels the subsequent central auditory display of sound elevation could be along temporal coordinates.

The results of the spectral discrimination experiment can be explained in terms of the effects of elevation-dependent filtering of the external ear. Spectral notches created by the external ear coincide with synthesized notches and make the spectral discrimination difficult. Because the spectral manipulation is sufficient to influence the bat's performance it is tempting to conclude that the processing is purely spectral, ignoring the reciprocal changes in the temporal domain.

The jitter experiment is a first step in examining this temporal component in the representation of elevation. The side-peak of errors occurs at delays that correspond to the delays in the impulse response at the appropriate elevations. However, the reverberation delays displayed in the external ear impulse responses of *Eptesicus* are shorter than the integration time for echo reception (300 to 350 μs) [36]. The bat would not receive these impulse peaks as separate events, rather the echoes would arrive at slightly different, but overlapping times and the bat would receive a single compound waveform. Spectral effects created from the interference pattern of the overlapping echoes provide the only representation of the reverberation delays available to the bat. Although the jitter experiment is a temporal discrimination paradigm and the results mirror the impulse response predictions, the initial auditory representation is apparently spectral.

Interference patterns are sculpted into echoes reflected from closely spaced multiple-component objects and from the convolutions and ridges of the pinnae. The frequency of notches and peaks is dependent upon the target's shape and/or elevation and thus can provide a signature for these features. Simmons and colleagues [37] presented compound echoes in a jitter experiment to demonstrate that bats converted the spectral interference patterns of the echo into an estimate of the time separation between echoes from individual target components. Application of the jitter techniques to sound sources that change in elevation reveals that the side peak in the jitter performance curves in Figure 4A behaves as though its location registers the timing of the external ear reverberation. A spectral correlation and transformation model describes the computation for converting the spectral pattern due to ear reverberations into echo arrival time [27].

References

[1] Batteau, D. W. "The role of the pinna in human localization." *Proc. Royal Soc. Lond.*, 8168: 158–180, 1967.

[2] Batteau, D. W. "Listening with the naked ear." In *The Neuropsychology of Spatially Oriented Behavior*, S. J. Freedman (ed.), Homewood: Dorsey Press, 1968.

[3] Blauert, J. "Sound localization in the median plane." *Acustica* 22: 205–213, 1969.

[4] Bloom, P. J. "Creating source elevation illusions by spectral manipulation." *J. Audio Engineering Soc.*, 25: 560–565, 1977.

[5] Carlile, S. "The auditory periphery of the ferret. II: The spectral transformations of the external ear and their implications for sound localization." *J. Acoust. Soc. Am.* 88: 2180–2204, 1990.

[6] Chen, J., VanVeen, B. D. & Hecox, K. E. "External ear transfer modeling: A beamforming approach." *J. Acoust. Soc. Am.* 92: 1933–1944, 1992.

[7] Dooley, G. J. and Moore, B. C. J. "Duration discrimination of steady and gliding tones: A new method for estimating sensitivity to rate of change." *J. Acoust. Soc. Am.* 84: 1332–1337, 1988.

[8] Fuzessery, Z. M. "Speculations on the role of frequency in sound localization" *Brain Behav. Evol.* 28: 95–108, 1986.

[9] Fuzessery, Z. M. "Monaural and binaural spectral cues created by the external ears of the pallid bat." *Hearing Res.* 95: 1–17, 1996.

[10] Grinnell, A. D. "The neurophysiology of audition in bats: directional localization and binaural interaction." *J. Physiol. Lond.* 167: 97–113, 1963.

[11] Guppy, A. and Coles, R. B. "Acoustical and neural aspects of hearing in the Australian gleaning bats *Macroderma gigas* and *Nyctophilus gouldi*." *J. Comp. Physiol. A.* 162: 653–668, 1988.

[12] Hartley, D. J. and Suthers, R. A. "The sound emission pattern and the acoustical role of the noseleaf in the echolocating bat, *Carollia perspicillata*." *J. Acoust. Soc. Am.* 82: 1892–1900, 1987.

[13] Hartley, D. J. and Suthers, R. A. "The sound emission pattern of the echolocating bat, *Eptesicus fuscus*." *J. Acoust. Soc. Am.* 85: 1348–1351, 1989.

[14] Hartley, D. J. and Suthers, R. A. "Sonar pulse radiation and filtering in the mustached bat, *Pteronotus parnellii rubiginosus*." *J. Acoust. Soc. Am.* 87: 2756–2772, 1990.

[15] Hebrank, J. and Wright, D. "Spectral cues used in the localization of sound sources on the median plane." *J. Acoust. Soc. Am.* 56: 1829–1834, 1974.

[16] Hiranaka, Y. and Yamasaki, H. "Envelope representations of pinna impulse responses relating to three-dimensional localization of sound sources." *J. Acoust. Soc. Am.* 73: 291–296, 1983.

[17] Humanski, R. A. and Butler, R. A. "The contribution of the near and far ear toward localization of sound in the sagittal plane." *J. Acoust. Soc. Am.* 83: 2300–2310, 1988.

[18] Jen, P. H.-S. and Chen, D. "Directionality of sound pressure transformation at the pinna of echolocating bats." *Hearing Res.* 34: 101–118, 1988.

[19] Kistler, D. J. and Wightman, F. L. "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction." *J. Acoust. Soc. Am.* 91: 1637–1647, 1992.

[20] Lawrence, B. D. and Simmons, J. A. "Echolocation in bats: The external ear and perception of the vertical position of targets." *Science*, 218: 481–483, 1982.

[21] Middlebrooks, J. C. and Green, D. M. "Sound localization by human listeners." *Ann. Rev. Psych.*, 42: 135–159, 1991.

[22] Middlebrooks, J. C. "Narrow-band sound localization related to external ear acoustics." *J. Acoust. Soc. Am.* 92: 2607–2624, 1992.

[23] Moore, B. C. J. and Sek, A. "Detection of combined frequency and amplitude modulation." *J. Acoust. Soc. Am.* 92: 3119–3131, 1992.

[24] Musicant, A. D., Chan, J. C. K. and Hind, J. E. "Direction-dependent spectral properties of cat external ear: New data and cross-species comparisons." *J. Acoust. Soc. Am.* 87: 757–781, 1990.

[25] Obrist, M. K., Fenton, M. B., Eger, J. L. and Schlegel, P. A. "What ears do for bats: A comparative study of pinna sound pressure transformation in *Chiroptera*." *J. Exp. Biol.* 180: 119–152, 1993.

[26] Rice, J. J., May, B. J., Spirou, G. A., and Young, E. D. "Pinna-based spectral cues for sound localization in cat." *Hearing Res.* 58: 132–152, 1992.

[27] Saillant, P. A., Simmons, J. A., Dear, S. P. and McMullen, T. A. "A computational model of echo processing and acoustic imaging in frequency-modulated echolocating bats: The spectrogram correlation and transformation receiver." *J. Acoust. Soc. Am.* 94: 2691–2712, 1993.

[28] Schnitzler, H.-U. and Grinnell, A. D. "Directional sensitivity of echolocation in the horseshoe bat, *Rhinolophus ferrumequinum*. I. Directionality of sound emission." *J. Comp. Physiol. A.*, 116: 51–61, 1977.

[29] Searle, C. L., Braida, L. D., Cuddy, D. R. and Davis, M. F. "Binaural pinna disparity: Another auditory localization cue." *J. Acoust. Soc. Am.* 57: 448–455, 1975.

[30] Shaw, E. A. G. "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane." *J. Acoust. Soc. Am.* 56: 1848–1861, 1974.

[31] Shaw, E. A. G. "External ear response and sound localization." In *Localization of Sound: Theory and Applications*, R. Gatehouse (ed.), Groton CT: Amphora, pp. 30–41, 1982.

[32] Simmons, J. A. "The resolution of target range by echolocating bats." *J. Acoust. Soc. Am.* 54: 157–172, 1973.

[33] Simmons, J. A. "Perception of echo phase information in bat sonar." *Science*: 204, 1336–1338, 1979.

[34] Simmons, J. A. "A view of the world through the bat's ear: The formation of acoustic images in echolocation." *Cognition*, 33: 155–199, 1989.

[35] Simmons, J. A., Ferragamo, M., Moss, C. F., Stevenson, S. B., and Altes, R. A. "Discrimination of jittered sonar echoes by the echolocating bat, *Eptesicus fuscus*: The shape of target images in echolocation." *J. Comp. Physiol. A* 167: 589–616, 1990.

[36] Simmons, J. A., Freedman, E. G., Stevenson, S. B., Chen, L. and Wohlgenant, T. J. "Clutter interference and the integration time of echoes in the echolocating bat, *Eptesicus fuscus*." *J. Acoust. Soc. Am.* 86: 1318–1332, 1989.

[37] Simmons, J. A., Moss, C. F. and Ferragamo, M. "Convergence of temporal and spectral information into acoustic images of complex sonar targets perceived by the echolocating bat *Eptesicus fuscus*." *J. Comp. Physiol. A.*, 166: 449–470, 1990.

[38] Simmons, J. A., Wotton, J. M., Ferragamo, M. J. and Moss, C. F. "Vertical localization of targets by echolocating bats: transformation of external-ear cues into auditory images by *Eptesicus fuscus*." submitted for publication.

[39] Wightman, F. L. and Kistler, D. J. "Headphone simulation of free-field listening. I. Stimulus synthesis." *J. Acoust. Soc. Am.* 85: 858–-867, 1989.

[40] Wightman, F. L. and Kistler, D. J. "Headphone simulation of free-field listening. II: Psychophysical validation." *J. Acoust. Soc. Am.* 85: 868–-878, 1989.

[41] Wightman, F. L. and Kistler, D. J. "Sound localization." In *Human Psychophysics*, W. A. Yost, A. N. Popper and R. R Fay (eds.). NewYork: Springer-Verlag, pp. 155–192, 1993.

[42] Wotton, J. M., Haresign, T. and Simmons, J. A. "Spatially dependent acoustical cues generated by the external ear of the big brown bat, *Eptesicus fuscus*." *J. Acoust. Soc. Am.* 98: 1423–1445, 1995.

[43] Wotton, J. M., Jenison, R. L., and Hartley, D. J. "The combination of echolocation emission and ear reception enhances directional spectral cues of the bigbrown bat, *Eptesicus fuscus*." *J. Acoust. Soc. Am.* 101**:** 1723–1733, 1997.

[44] Wotton, J. M., Haresign, T. Ferragamo, M. J. and Simmons, J. A. "The influence of sound source elevation and external ear notch cues on the discrimination of spectral notches by the big brown bat, *Eptesicus fuscus*." *J. Acoust. Soc. Am.* 100: 1764–1776, 1997.

[45] Wright, D., Hebrank, J. H. and Wilson, B. "Pinna reflections as cues for localization." *J. Acoust. Soc. Am.* 56: 957–962, 1974.

COMPUTATIONAL ASSESSMENT OF AN ACOUSTIC FLOW HYPOTHESIS FOR CF-BATS

Rolf Müller, Hans-Ulrich Schnitzler

Animal Physiology, Tübingen University
Morgenstella 28, D-72076, Tübingen, Germany

1. Introduction

Biological sonar systems reside on organisms which are often endowed with remarkable mobility. A typical scenario for sensing therefore includes relative motion between transducers and reflectors while the measurement is in progress. This introduces a time variance into the sensory input, which is a mixed blessing; on the one hand, additional degrees of freedom are introduced, which increase the number of signal parameters or measurements as well as the amount of computation required to extract the desired information. On the other hand, the ensemble of salient stimulus features may be enriched substantially by motion-induced cues, which would not have been available otherwise.

In the study of visual perception the use of motion-generated cues has been extensively pursued, developing it into a research area with an impressive outflow of insight sustained by mature scientific theory and methodology. Particularly noteworthy are the advances in understanding conceptual and computational aspects of the problem [8][6]. Optic flow is defined as the projection of the three-dimensional field of relative velocities between observer and points in space onto a two-dimensional image plane. It may be used to recover the egomotion of the observer (e.g. [7]) as well as structural characteristics of the environment (e.g. [14]). Furthermore, a sensory variable critical to immediate control of action, the "time-to-contact" (usually designated τ) can be computed directly from the expansion rate of the optic flow field (e.g. [9]). First order estimates of time-to-contact provided in this way may form the basis for an assortment of braking strategies [10].

Motion-dependence can be found in acoustic as well as in optic signal properties. For instance, the readily perceived Doppler effect constitutes a genuinely motion-related property, which, unlike optic flow, would be absent in a series of static snapshots taken at successive points on a trajectory. Therefore, looking for analogies in perceptual utilization of motion-specific optic and acoustic cues seems obvious. Furthermore, the search for an acoustic analog to optic flow as well as an auditory analog to perception of optic flow may even offer the promise of a unified view of sensory system function.

2. An Acoustic Analog To Optic Flow?

Any expectation of a tight one-to-one correspondence between acoustic and optic flow and the related perceptual mechanisms may be dismissed readily as overly optimistic, since obvious limitations to a hypothetical analogy can be derived from three lines of argument:

1) physical properties of light and sound
2) dimensionality of the sensor arrays
3) different behavioral salience of the vision and audition modalities

2.1 Physical Properties of Light and Sound

Although light and sound share a common wave nature, there is a radical difference in wavelength: for visible light it is on the order of magnitude of 10^{-7} *m*, bat echolocation operates in the range 10^{-3}–10^{-2} *m* and human hearing is sensitive to sound wavelengths of 10^{-2}–10 *m*. Many structures which matter in the daily life of bats or humans are close to the wavelength of sound but thousands of wavelengths across optically. Consequently, diffraction is much more prominent in the formation of acoustic wave fields [11] and spatial resolution is in general inferior to optical sensing. Thus, an eventual acoustic flow field evaluation will have to be based on structural features that are not resolved as well as its optic counterpart.

2.2 Dimensionality of the Sensor Arrays

The auditory sensory epithelia do not represent projected spatial dimensions like the retina does, but form a map of signal frequency instead. The optic flow field, on the contrary, is the projection of a three-dimensional velocity field onto the two spatial dimensions of the retinal image. Since the spatial sampling by the mammalian hearing system is limited to two points (one at each ear), a hypothetical acoustic analog to optic flow would have to include a transformation of velocities in the “acoustic array” into non-spatial signal dimensions.

2.3 Different Behavioral Salience of the Sensory Modalities Vision and Audition

While vision as well as audition are conveying spatial percepts, in humans a general division of labor between the two seems to exist. Vision plays the leading role in spatial perception and hearing serves primarily, but not exclusively, functions in the temporal domain, e.g. alerting to approaches from outside the visual field [5]. The specific role of audition will doubtlessly limit the applicability of acoustic flow concepts to human perception, but does not apply to the situation in bats, where audition constitutes a sufficient spatial sense of its own.

From the given restrictions it is evident that analogies of an eventual acoustic flow concept to optic flow will be confined to a fairly abstract and generalized level. Bearing this precaution in mind, some speculative analogs between optic and acoustic flow may be formulated, examples of which are listed in Table 1.

3. Hypothetical Flow Field Parameters in CF-Bats

The species of so-called cf-bats (cf is an abbreviation for “constant frequency”) emit echolocation pulses dominated by prolonged (duration up to $\approx$100 *ms*), narrow-band signals characterized by a constant carrier frequency and a shallow sloping envelope. Such narrowband signals are ill-suited for target localization based on a combination of binaural azimuth measurements and time-of-flight based range estimates. Pulse trains are usually delivered at a high duty cycle (up to $\approx$ 80%) and hence give rise to an almost continuous stream of echoes as the bats are moving through their reflector-rich forest habitats. Taking into account the peculiar signal design, the non-zero derivatives with respect to the trajectory of the carrier frequency, induced by Doppler shifts, and envelope amplitude, caused by a time-varying sound channel gain, constitute promising candidates for sensory variables conveying spatial information. Here, we restrict ourselves to localization within a plane and express target positions in Cartesian coordinates d, h (Figure 1).

Table 1: Some speculative analogs between optic and acoustic flow. Abbreviations: ITD = interaural time difference, f_e = emitted frequency, v speed of observer or target, c speed of sound, f observed frequency

Optic	Acoustic	Information content
Local vs. global flow	Doppler shift of some or (almost) all echoes	Object vs. egomotion
Expansion rate	Changes in e.g. intensity, ITD, pitch, perceived extent of one or more sources	Time-to-contact
Focus of expansion occupied	Cone of high collision risk (static: frequency band close to $f_e(1+2v/c)$, dynamic: df/dt	Impending collision
Smooth velocity gradient along the image of a surface	Shape of the spectral profile	Surface slant
Discontinuities in flow	Tracking of individual echoes, auditory streaming	Layout of scene, depth structure

Following the approach of "τ-variables," which specify time-to-contact [10], the proportional changes defined as

$$\Psi = \frac{\partial f_d}{\partial h}/f_d;\ \Theta = \frac{\partial P}{\partial h}/P \qquad (1)$$

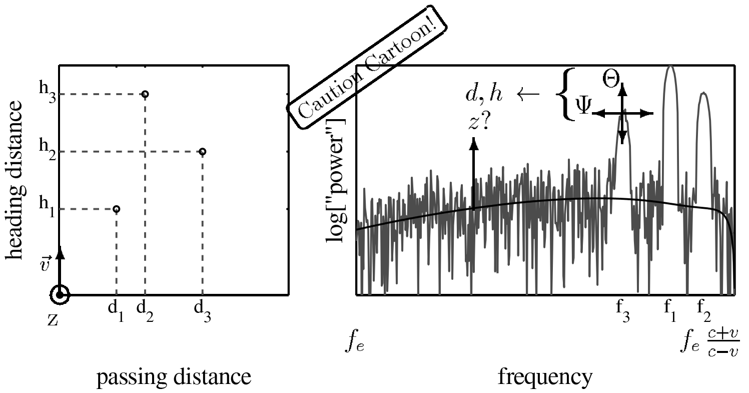


Figure 1 Cartoon illustrating the simplified scenario and the hypothetical flow field variables considered in this study: Target positions d, h in a plane aligned with regard to the animal's velocity vector $\vec{v}$ (left graph) may eventually be recovered using the proportional changes in echo carrier frequency, Ψ, and amplitude, Θ (right graph). If, in addition, a third dimension, z (height above ground), is required, other information sources (e.g. the global spectral shape of noisy returns from the ground) would have to be evaluated.

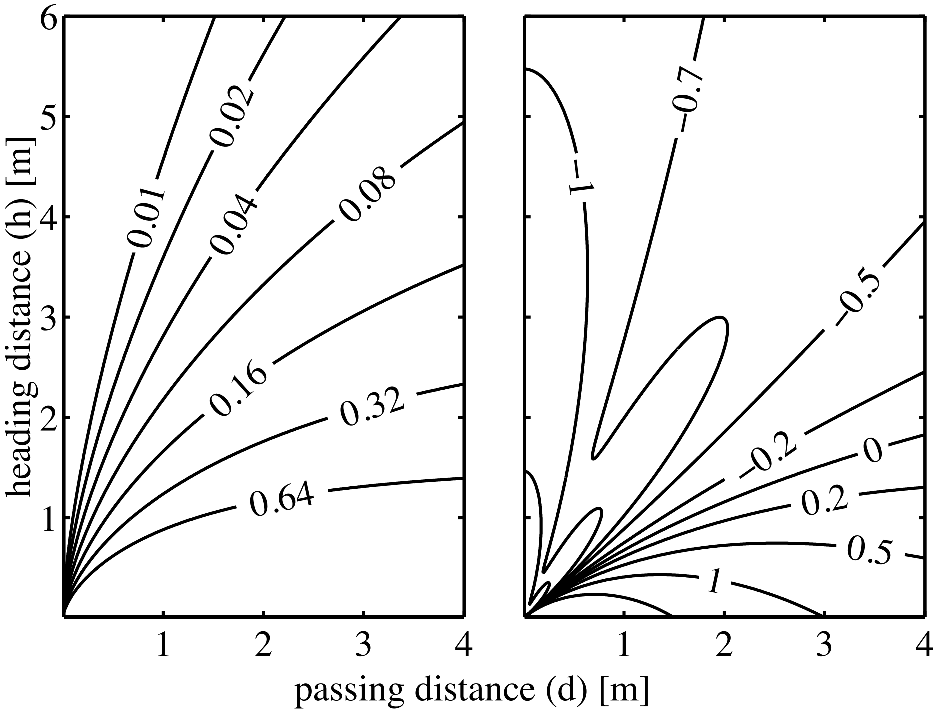


Figure 2 Values of the putative flow variables Ψ [m^{-1}] (left) and Θ [m^{-1}] (right) as a function of position within the right frontal hemifield. The definition of the coordinates is provided in Figure 1.

are considered (Figure 1). f_d is the Doppler frequency and P the amplitude of the sound pressure envelope. Usage of the ratios (Equation 1) is beneficial, because in a simplified world, where targets are point scatters, wave fields have spherical geometry (with some superimposed directivity due to the sonar system's transducers) and the Doppler effect follows the approximate narrow-band model f_d=2v/$c$cosφ (φ being the target bearing), non-trivial constants, like the target scattering coefficient, cancel.

4. Characterizing Flow Fields

Acoustic flow fields for the variables Ψ and Θ are scalar fields, which may be rated in terms of perceptual salience and estimation accuracy by the values of the variables themselves (Figure 2), the magnitude of the respective gradients (Figure 3), the angle subtended by the gradient vectors of the two fields (Figure 4) and the number of solutions d, h obtained. The fields depicted are given by

$$\Psi(d, h) = \frac{d^2}{h(d^2 + h^2)} \quad [m^{-1}] \tag{2}$$

and

$$\Theta(d, h) = -\frac{2h}{d^2 + h^2} - \frac{\ln(10)\alpha(f)h}{10\sqrt{d^2 + h^2}} + \frac{\partial\Phi}{\partial h}/\Phi \quad [m^{-1}] \tag{3}$$

In Equation 3, ln is the natural logarithm, $\alpha(f)$ is the absorption coefficient (in dB/m) and

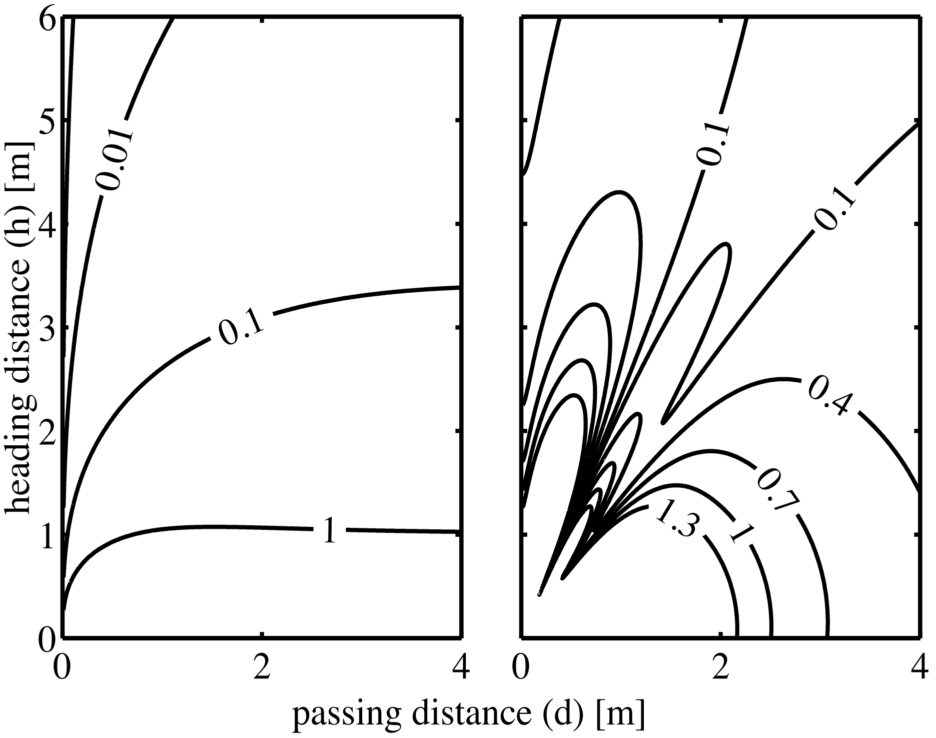


Figure 3 Gradient magnitudes for the scalar fields shown in Figure 2: $|\nabla\Psi|[m^{-2}]$ (left) and $|\nabla\Theta|[m^{-2}]$ (right). The definition of the coordinates is provided in Figure 1.

$0 \leq \Phi(d, h) \leq 1$ the joint directivity function of emitter and receiver. A detailed discussion of these relationships is provided in [12]. The following aspects appear noteworthy:

All functions are bilateral symmetric around the *h*-axis (Figure 2, Figure 3 and Figure 4 are therefore restricted to the right-hand hemifield.) Consequently, target position estimates based on Ψ and Θ display the same symmetry properties and require disambiguation. This could be achieved either by means of other cues or by testing the updated estimates after a tentative steering maneuver against the predictions from the previous set of estimates.

The indented shape of the isocontours in Figure 2 and Figure 3 is a consequence of the assumed joint directivity of receiver and emitter $\Phi(d, h)$, which was modeled as a sum of two Gaussians using data reported in [4]. As obvious from inspection of the isocontour-shapes, multiple solutions (up to three per hemifield) for target positions will result in the regions of the indentations. The restricted extent of these regions, however, renders these effects a negligible nuisance.

For bats inspecting their environment on the wing the forward direction is expected to be particularly salient, since eventual obstacles straight ahead would require immediate action. It should be noted however, that the gradients for both flow variables (Figure 3) are not favorable of a good resolution along the *h*-axis, although the gradients are orthogonal for $d = 0$ (Figure 4).

Using the criteria briefly introduced here, the relative errors in measurements of Ψ and Θ should be fairly small (on the order of a few percent) to allow crude localization [12]. Ultimately, the numbers given in Figure 2 and Figure 3 should be compared to the perceptual

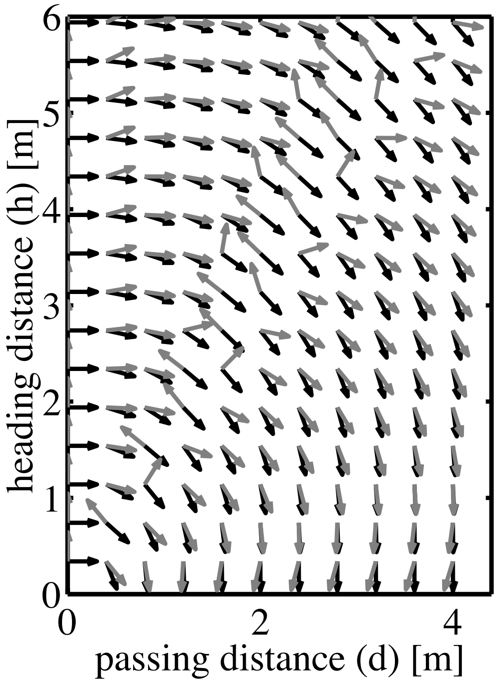


Figure 4 Gradient directions for Ψ (black) and Θ (gray). The definition of the coordinates is provided in Figure 1.

thresholds found in cf-bats in order to determine the precise localization accuracy that is conveyed by these variables.

5. Constraints On Extraction

5.1 Implementation Strategies

Obtaining estimates $\hat{\Psi}(t)$, $\hat{\Theta}(t)$ for $\Psi(t)$ and $\Theta(t)$ amounts to jointly estimating amplitude and frequency modulations as well as the values of the corresponding carriers. There are two possible implementation strategies for achieving this goal: a bank of matched filters and a continuous, "single-track" estimator. A bank of matched filters tests multiple hypothesis for the parameter value in parallel, whereas the "single-track estimator" directly transforms the input signal into an estimate for the parameter of interest. Of course, each of these two implementations could be supplemented with a conversion stage, which mimics the other's output. For instance, the map provided by the bank of feature detectors could be read out, substituting map position with an estimated value of the parameter. Interpolation could be included to smooth the resulting continuous estimate. Along the same line of argument, the output of the continuous estimator could be passed through a quantizer, which converts each interval of input values into the position of the respective thresholding element. The difference between the two implementation strategies lies neither in the output format nor in the basic underlying estimation theory, but in the design considerations for the actual implementation. The latter will be considered here.

Feature maps in the auditory system have been found in numerous experiments, this is particularly true for the echolocation system of bats (see e.g. the title of [15] for an unequivocal statement regarding the significance of these findings). Since such maps would constitute the native output format of a matched filter bank, assuming the latter to be the universal neural estimator implementation is tempting. In modeling of visual perception a bank of matched filters is a widely accepted approach (motion energy filters [1]), and is also corroborated by experimental findings. The same approach has been suggested for modeling auditory perception [18]. However, the continuous, single-track estimator can be found to be present in the brain, too, particularly for the purpose of effective perception-action coupling as for instance required in neural circuits controlling protective reflexes. Single-track estimators may be favorable, when both the implementation of the estimator, as well as deriving a motor-control signal from its native output format, is straightforward.

As stated above, the echoes $u_{in}(t)$ received in a scenario which includes egomotion, are simultaneously modulated in amplitude $a(t)$ and frequency $\Omega(t)$.

$$u_{in}(t) = a(t)\cos\left(2\pi\int_{-\infty}^{t}\Omega(\tau)d\tau\right) \tag{4}$$

The bandpass-filtered signal $u_{out}(t)$ (filtered with transfer function $H(f)$) is given by

$$u_{out}(t) = a(t)|H[f(t)]|\cos\left(2\pi\int_{-\infty}^{t}\Omega(\tau)dt + \arg\{H[f(t)]\}\right) \tag{5}$$

Under the narrow-band assumption the envelope is the product of the amplitude modulation, $a(t)$, and the magnitude transfer function $|H[f(t)]|$. Measuring $|u_{out}(t)|$ in a single channel does not allow us to distinguish between amplitude and frequency modulation. A straightforward approach a biological system might take is to estimate the carrier frequency as the moment of the excitation pattern and use this estimate to correct for the magnitude transfer function of the filter. The separation of the two forms of modulation makes a 2-dimensional matched filter superfluous and there is no need to maintain a separate template for each combination of amplitude and frequency modulation. Thus, a simple single-track estimator is given preference as a more parsimonious hypothesis.

Utilization of the flow parameters described here is entirely hypothetical and so is any proposed neural substrate. The sole purpose of the processing scheme shown in Figure 5 is to demonstrate that the outlined estimation procedure may be readily reconciled with the basic principles of neural function. The schematic assumes analog waveforms to be decoded from the spike train of a single neuron, with synapses acting as reconstruction filters [2]. If a population code was to be assumed instead, each neuron-symbol would have to be replaced with an ensemble of neurons, leaving the rationale of the diagram unchanged. Computing the first moment calls for a neural implementation of the expression

$$\hat{f}_c(t) = \frac{\sum_{n=1}^{N} x_n(t)f_n}{\sum_{n=1}^{N} x_n(t)} \tag{6}$$

where x_n is the excitation amplitude of the n-th bandpass channel and f_n its center frequency. The denominator in Equation 6 could be computed by a neuron, which sums all its inputs with unity weight. In order to obtain the numerator, synaptic weights are set to represent the center frequency of the presynaptic bandpass channel. Thus, the output of this neuron repre-

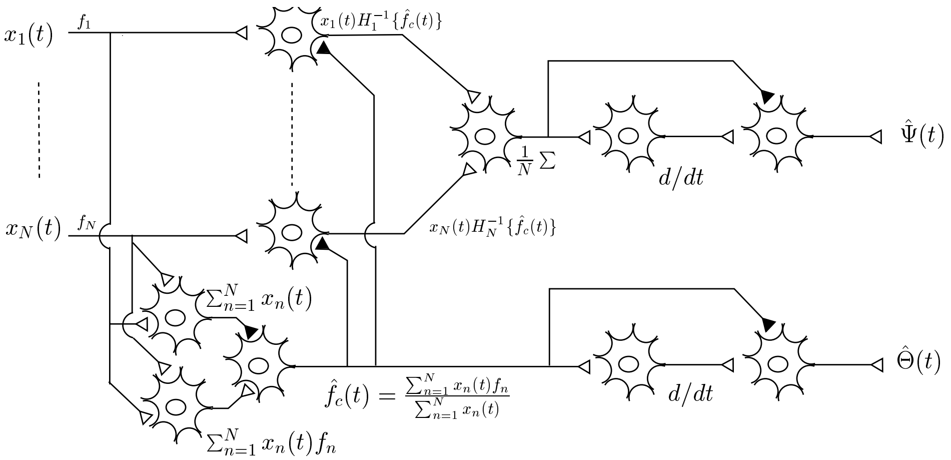

$x_1(t)$
f_1
$x_1(t)H_1^{-1}\{\hat{f}_c(t)\}$
$\frac{1}{N}\sum$
d/dt
$\hat{\Psi}(t)$
$x_N(t)$
f_N
$x_N(t)H_N^{-1}\{\hat{f}_c(t)\}$
$\sum_{n=1}^{N} x_n(t)$
$\hat{f}_c(t) = \frac{\sum_{n=1}^{N} x_n(t) f_n}{\sum_{n=1}^{N} x_n(t)}$
d/dt
$\hat{\Theta}(t)$
$\sum_{n=1}^{N} x_n(t) f_n$

Figure 5 Schematic illustration of how continuous estimation of the proposed flow parameters Ψ and Θ could be implemented in accordance with some basic principles of neural function. x_n is the excitation amplitude of the n-th bandpass channel. Black "synapses" are thought to be inhibitory, the ratios in the accompanying expressions may thus be realized as subtractions of log-compressed signals. This compression is a procedural detail, which is left out in the expressions for sake of readability.

sents an estimate of $\hat{f}_c(t)$, the carrier frequency of the input signal, without any need for the physical units of the synaptic strengths to be those of a frequency. Divisions (like the one in Equation 6) can easily be implemented neurally by a subtraction of log-compressed signals. The logarithm only maintains the sign, but not the magnitude of the derivatives. This would have to be taken into account, if further upstream an estimate for the derivative of the uncompressed function is required. In this case a neuron with a transfer function, which is exponential over a limited range of input values, is required, forming a compander together with the compression stages. A neural implementation of a division is also needed to correct for the influence of the channels' transfer functions on the envelope amplitude (Figure 5). This step requires a neural element, which computes the n-th transfer function $H_n(f)$ at the estimated carrier frequency $\hat{f}_c(t)$. A parsimonious solution is to resort to the transfer function of the synapse over which $\hat{f}_c(t)$ is fed back to the envelope-detected output of the bandpass channels. Once the influence of the individual transfer functions has been corrected for, envelope estimates can be collapsed across channels in order to reduce the variance (to the extent to which the channel outputs are independent). The concluding steps are computation of the derivatives (e.g. by a neuron with a differentiator transfer function) and formation of the ratio described in Equation (1) (another log-compression and subtraction).

5.2 Interferences and Demodulation Distortions

Even though the joint directivity of the emitter and receiver of cf-bats [4] limits the volume of the bats' resolution cell (defined as the volume which contributes to the received echo), the presence of more than one reflector within this cell can hardly be declared an uncommon occurrence. The bandwidth of the auditory filters are in such a situation easily permissive of overlapping auto-representations of the individual echoes. This leads to the formation of oscillatory cross-terms (Figure 6) in the envelope-detected filter outputs

$$|y(t)| = \sqrt{\sum_n {a_n}^2 + 2\sum_{n \neq m} a_n a_m \cos(\Delta\omega_{mn} t + \Delta\varphi_{mn})} \tag{7}$$

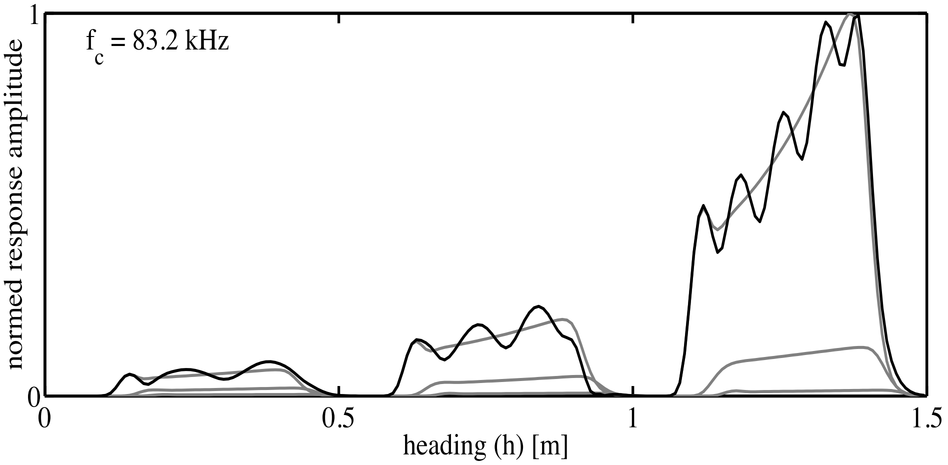


Figure 6 Example of a filter bank channel's output in the presence of 3 targets. Grey lines are the responses (envelopes) to echoes from each target presented in isolation. The oscillatory, black curve is the response to the superposition of the 3 echoes. f_c denotes the center frequency of the filter, its equivalent rectangular bandwidth was 95.9 *Hz*. The model used an IIR implementation of 4th order Gammatone filters described by Slaney [17].

where *y(t)* is the analytic signal formed from the echo, $\Delta\omega_{mn}$=ω_m-ω_n and $\Delta\varphi_{mn}$=φ_m-φ_n are the differences in frequency and phase, respectively and $a_n\cos(\omega_{nt}+\varphi_n)$ is the filter output of the n-th input signal prior to demodulation. The oscillatory nature of the cross-terms is detrimental to the computation of derivatives, which will hardly reflect the changes in gain or Doppler shift pertinent to the sound channel, when there is interference.

A fairly reasonable model of auditory demodulation is given by half-wave rectification followed by low-pass filtering [16]. Such a demodulation procedure does not result in a "clean" frequency shifting of the signal spectrum, if the input consists of several superposed echoes. Therefore, the initial bandpass filtering stage is critical to reduction of both interference and demodulation distortions. In case of the interference, the magnitude transfer functions are clearly not narrow enough in order to provide an adequate remedy. Demodulation distortions are easier to ameliorate than interferences [13]. Therefore, an additional filtering stage, e.g. as discussed in [3], may be employed to separate at least the envelope of one strong signal of interest from attenuated distractors.

6. Conclusions

The potential existence of an auditory analog to visual flow field perception is an enticing concept. However, analogies have to be confined to a somewhat abstract level, since some important physical and perceptual aspects of the two phenomena differ fundamentally. Echolocating animals, which can employ sonar as a sufficient spatial sense, constitute the most promising candidates for utilization of such analogies.

The model system of cf-bat echolocation allows for the formulation of an operational acoustic flow hypothesis. The resulting scalar fields for the flow variables proportional change in Doppler shift, Ψ, and envelope amplitude Θ can be readily characterized for a simplified acoustica scenario. The properties of these fields are found to be commensurate with the requirements of at least crude localization within a plane. This may be sufficient to meet the requirements of, for instance, obstacle avoidance, where conservative choice of safety margins may compensate for a large estimator variance.

Extraction of the hypothetical flow parameters directly from the signal is feasible by a simple signal processing scheme. There is no need for a bank of matched filters containing a template for every resolved combination of the two parameters. If transfer functions of connections between neurons could be made use of, only a small number of neurons would be needed in order to carry out the required transformations. A continuous estimate of the relevant parameters obtained in this way should provide a convenient substrate for the generation of motor-control signals.

The auditory spectrogram in cf-bats, as represented by the available neurophysiological data, is susceptible to the formation of cross-terms between multiple echoes. Although application of additional smoothing low-pass filters is capable of providing a partial remedy, the oscillatory nature of these cross-terms will degrade the usability of the flow parameters discussed. It is therefore hardly conceivable that acoustic flow could provide an analog to the visual perception of a densely occupied large scale flow field. Coping with a few echoes spaced favorably in frequency should be achievable, though. The latter scenario is also in accordance with sonar in air, providing only a limited field of view, thereby windowing in on a small illuminated volume.

References

[1] Adelson, E. H. and Bergen, J. R. “Spatiotemporal energy models for the perception of motion.” *J. Opt. Soc. Am. A*, Vol. 2, pp, 284–299, 1985.

[2] Bialek, W., Rieke, F., de-Ruyter-van-Steveninck, R. R. and Warland, D. “Reading a neural code.” *Science*, 252: 1854–1857, 1991.

[3] Dau, T., Kollmeier, B. and Kohlrausch. A. “Modeling auditory processing of amplitude modulation: Detection and masking with narrow-band carriers.” *J. Acoust. Soc. Am.*, 102: 2892–2905.

[4] Grinnell, A. D. and Schnitzler, H.-U. “Directional sensitivity of echolocation in the Horseshoe Bat *Rhinolophus ferrumequinum*: II. Behavioral directionality of hearing.” *J. Comp. Physiol. A*, Vol. 116, pp. 63–76, 1997.

[5] Guski, R. “Acoustic tau: an easy analogue to visual tau?” *Ecol. Psych.,* 4: 189–197, 1992.

[6] Hildreth, E. C., “The neural computation of the velocity field.” In *Vision and the Brain, the Organization of the Central Visual System,* B. Cohen and I. Bodis-Wollner (eds.), New York: Raven Press, pp. 139–164, 1990.

[7] Hildreth, E. C. “Recovering heading for visually-guided navigation.” *Vision Res.,* 32: 1177–1192, 1992.

[8] Koenderink, J. J. “Optic flow.” *Vision Res.,* 26: 161–180, 1986.

[9] Lee, D. N. “The optic flow field: The foundation of vision.” *Phil. Trans. Roy. Soc. Lond. B,* 290: 169–179,1980.

[10] Lee, D. N., van der Weel, F. R. and Hitchcock, T., Matejowsky, E., Pettigrew, J. D. “Common principle of guidance by echolocation and vision.” *J. Comp. Physiol. A,* 171: 563–571, 1992.

[11] Morse, P. M. *Vibration and Sound* (2nd ed.). New York: McGraw-Hill, 1948.

[12] Müller, R., Schnitzler, H.-U. “Acoustic flow perception in cf-bats: properties of the available cues.” *J. Acoust. Soc. Am.,* 105: 2958–2966, 1999.

[13] Müller, R. and Schnitzler, H.-U. “Acoustic flow perception in cf-bats: Extraction of flow parameters.” *J. Acoust. Soc. Am.,* 108: 1298-1307, 2000.

[14] Rieger, J. H. and Lawton, D. T. “Processing differential image motion.” *J. Opt. Soc. Am. A,* 2: 354–359, 1985.

[15] Riquimaroux, H., Gaioni, S. J., Suga, N. “Cortical computational maps control auditory perception.” *Science,* 251: 565–568, 1991.

[16] Schroeder, M. R. and Hall, J. L. “Model for mechanical to neural transduction in the auditory receptor.” *J. Acoust. Soc. Am.,* 65: 1055–1060, 1974.

[17] Slaney, M. “An efficient implementation of the Patterson-Holdsworth auditory filter bank.” Apple Technical Report 1993-35, Cupertino, CA: Apple Computer, 1993.

[18] Todd, N. P. M. “A model of auditory image flow: Architecture.“*J. Acoust. Soc. Am.,* Vol. 103: 2844, 1998.

AUDITORY PROCESSING OF PITCH

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

AUDITORY PROCESSING OF PITCH

Malcolm Slaney

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120, USA

1. Modeling Pitch

Why should we study pitch? Why have people paid so much attention to pitch over the years? Most important to the answer, pitch is one of the most salient perceptions we have of a sound. Pitch is "that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale" [1].

Pitch means different things to different people. The definition in the previous paragraph most closely approximates how a musician defines pitch and is based on human perception [11]. People who work with speech define pitch to be the frequency at which the human glottis opens and closes during speech—a definition based on a production model of pitch. Engineers often define pitch to be the fundamental frequency in a Fourier analysis. For simple sounds the pitch of a signal is well described by the glottis and the fundamental frequency, but these physical measurements are only approximations to the "true" pitch. Thus, we use the definition based on human perception [10].

Pitch is an interesting aspect of sound both because it is salient and because it has the potential to tell us so much about how sound is processed by the brain. Most individuals can make a reliable judgment about which of two sounds has the higher pitch. Thus, psychoacousticians use different kinds of sounds to probe the auditory system and tease out its secrets. In the best experiments, sounds with a clear pitch difference might rule out one approach and lend support to another

At this time, models of pitch are commonly based on either spectral or temporal descriptions of the sound. This distinction is hard to resolve because the two representations are related in a linear system by a Fourier transform. By applying a simple linear transform, we can make in one mathematical domain any decision that we can make in the other. Yet the difference between the two approaches has a profound effect on the neural machinery needed to perceive the auditory signal. I favor a model of pitch based on temporal processing, but that preference is by no means universal. The auditory system is not a linear system, so perhaps some combination of the two approaches is best [3], [7], [8].

2. Spectral Pitch

We know that the cochlea disperses the frequencies of a sound so that different inner hair cells respond most vigorously to different frequencies. Spectral models of pitch use this distributed representation, in which the firing rate of a neuron encodes the amplitude associated with a specific frequency region, to form a rate profile describing the sound's spectrum.

Given the frequencies with the strongest response, there are computational models that will give a pitch estimate that closely approximates human perception and performance [5]. (Cohen [5] is only the most recent work from this perspective; earlier work is cited in this paper.)

The spectral models, which determine pitch by the frequencies present in the sound, are easy to understand. They are most useful for accounting for the pitch of sounds that have aurally resolved frequency components—i.e., those with component frequencies for which the cochlea can distinguish clearly between adjacent partials. This is also the region where the pitch percept is strongest [12].

3. Temporal Pitch

Temporal models of pitch ignore the spectral profile in favor of the timing information in auditory firings. The cochlea is not a perfect Fourier analyzer, yet throughout the auditory system many neurons do an amazing job of preserving the temporal information. Thus, the inner-hair cells that respond most vigorously to 1-kHz stimuli also tend to fire at the same point in the input waveform. An ensemble of neurons will tend to fire at intervals of 1 ms [13].

It is well established (for example, by Cariani [4]), that neural firings contain the temporal information that models can use to make pitch decisions that match those made by human judgments. The temporal models of pitch look most promising when the sound has high harmonics—that is, where the cochlea is no longer able to resolve the frequency components [9]. Likewise, temporal models work well in many interesting cases where the detail and robustness of the timing information are an advantage, such as double-vowel simulations of the cocktail-party effect [2].

4. Section Overview

The papers in this section discuss two aspects of pitch perception: dichotic pitch and neural machinery that estimates pitch.

Akeroyd and Summerfield describe an interesting problem that combines pitch and binaural hearing: the dichotic-pitch problem. Most pitch sensations can be heard with either ear alone or in stereo, but the pitch percept is independent of the perceived location of the sound. To a first approximation, the machinery that humans use to estimate pitch and to localize the sound are independent.

There are sounds that evoke a sense of pitch when played to the two ears concurrently, yet fail to evoke one when played to only one ear. Somehow. the auditory system combines the information presented to the two ears to derive a common pitch. The best known example of such a sound is the Huggins pitch [6].

Akeroyd and Summerfield describe a model that predicts both the pitch and perceived spectral location of a Huggins pitch signal. Their approach assumes that the sound is heard by the two ears as two different auditory objects; the auditory system cannot analyze the perception as though there were only a single pitch at a single location. Akeroyd and Summerfield first use a correlogram to estimate the pitch and spatial location of the tonal signal, then subtract the tonal sound's effect from the correlogram before determining the location of the noise percept. In effect, they separate the sound into distinct objects before they determine the individual locations.

Cai, McGee, and Walsh describe a model that tests whether octopus cells in the cochlear nucleus could be extracting the information needed to make a temporal-pitch judgement.

Octopus cells, in particular, are often identified as potential temporal integrators because they gather information from many different frequencies and their responses are aligned precisely in time. Could they comprise the machinery that converts temporal information into pitch?

The Cai model uses input from a broad range of auditory-nerve fibers. The authors apply the model to harmonic complexes, over a wide range of frequencies, then measure the overall response and compare it to that of humans. In only certain cases does the octopus-cell response approximates human perception.

Computational models of pitch are not yet complete. These two works are good examples of current research.

References

[1] American National Standards Institute. "Acoustical Terminology." *ANSI S1.1-1994*.

[2] Assman, P. F. and Summerfield, Q. "Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies." *J. Acoust. Soc. Am.* 88: 680–697, 1990.

[3] Bilsen, F. A. and Ritsma R. J. "Repetition pitch mediated by temporal fine structure at dominant spectral regions." *Acoustica* 19:114–116, 1967.

[4] Cariani, P. "Neural timing nets for auditory computation." This volume, 2001.

[5] Cohen, M. A., Grossberg, S., and Wyse, L. L. "A spectral network model of pitch perception." *J. Acoust. Soc. Am.* 98:862–879, 1995.

[6] Cramer, E. M. and Huggins, W. H., "Creation of pitch through binaural interaction." *J. Acoust. Soc. Am.* 30:412–417, 1958.

[7] Evans, E. F. "Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons." *Audiology* 17: 369–420, 1978.

[8] Goldstein, J. L. and Srulovicz, P. "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement." In *Psychophysics and Physiology of Hearing*, E. F. Evans and J. P. Wilson (eds.), London: Academic Press, 1977.

[9] Meddis, R. and Hewitt, M. J. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification." *J. Acoust. Soc. Am.* 89: 2866-–2882, 1991.

[10] Moore, B. C. J. *An Introduction to the Psychology of Hearing*. London: Academic Press, 1997.

[11] Pierce, J. R. *The Science of Musical Sound*. New York: W. H. Freeman, 1992.

[12] Ritsma, R. J. "Frequencies dominant in the perception of the pitch of complex sounds." *J. Acoust. Soc. Am.* 41: 191–198, 1967.

[13] Shamma, S. A. "Speech processing in the auditory system. I: The representation of speech sounds in the responses of the auditory nerve." *J. Acoust. Soc. Am.* 78: 1612–1621, 1985.

A COMPUTATIONAL AUDITORY MODEL OF THE LATERALIZATION OF THE HUGGINS PITCH

Michael A. Akeroyd and A. Quentin Summerfield

MRC Institute of Hearing Research,
University Park, Nottingham, NG7 2RD, United Kingdom

1. Introduction

The Huggins pitch is the prototypical example of a dichotic pitch: a pitch that can only be heard when a specially crafted, flat-spectrum noise is presented to both ears simultaneously. If only one ear (it does not matter which) is stimulated then just the "shshsh" of the noise is heard, but if both ears are stimulated a faint tone is also heard against the background of noise. Although auditory scenes consisting of a tone and a noise are rare in everyday life, a dichotic pitch is of interest because the percept of pitch is entirely due to binaural analysis. Thus, an understanding of the perceptual characteristics of dichotic pitches — their pitch, spatial position, and loudness — may illuminate the action of the binaural auditory system. This chapter deals with the spatial position of the Huggins pitch. Current models of spatial position can successfully account for many cases of single sounds (e.g., [24], 25], [27]), but they fail with respect to the Huggins pitch because two sounds are heard simultaneously. This fundamental effect is accounted for in our model by separating the auditory scene into two distinct objects, the Huggins pitch and the noise, before computing the spatial position of each object. This new model is offered as an alternative to the only other model of the spatial position of the Huggins pitch, the "central-spectrum model" (e.g., [21]).

Huggins created his pitch by presenting a white noise over headphones, but with one channel passed through an allpass filter (Figure 1, left panel) [3], [14]. The filter did not change the amplitude spectrum of the noise but introduced a progressive change in the phase spectrum: 0 to 360° over a narrow band (ca. 60 Hz) of frequencies centered on 600 Hz (middle panel). This phase change created a progressive shift in interaural time delay (ITD) near 600 Hz because the other channel was unfiltered (right panel). The phase change could not be detected monaurally because a phase-shifted white noise sounds identical to any other white noise. However, the ITD shift can be detected by binaural analysis. The result is a percept of a faint, 600-Hz tone amidst the noise. The term "Huggins pitch" refers to this tone (the term "Huggins-pitch stimulus" refers to the entire acoustic stimulus as presented: i.e., the interaurally phase-shifted noise).

The Huggins pitch is heard within the head, as is usual with headphone-presented sounds. Some listeners hear the pitch to the far left of the head while others hear it to the far right. However, all listeners hear the background noise at the center of the head. Importantly, the spatial position within the head (i.e., the lateral position) of the Huggins pitch can be varied by manipulating the interaural configuration of the noise. For example, if either channel is inverted, the Huggins pitch is heard near the center of the head and the background noise is heard diffusely across the head. The perceived pitch is still 600 Hz; all that has changed is its lateral position. Our discussion of lateralization will concentrate on these two variants of

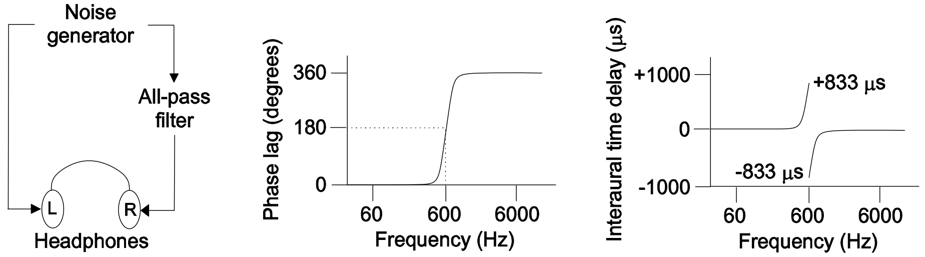


Figure 1 Left panel: schematic illustration of the apparatus for creating a 600-Hz Huggins pitch. Middle panel: phase response of the all-pass filter. Its amplitude response is flat in the illustrated range. Right panel: interaural time delays introduced by the all-pass filter.

the Huggins pitch which, following standard binaural terminology, are termed an N_0 Huggins pitch and an N_π Huggins pitch, respectively.

Raatgever, Bilsen and colleagues ([9], [20], [21], [22]) argued that their central-spectrum model could account for the pitch and lateralization of most dichotic pitches, including the Huggins pitch. Their model incorporated a Fourier transform to represent the initial frequency analysis of the inner ear, thus providing much finer frequency selectivity than is observed physiologically or psychophysically [17]. When implemented using a computational filterbank incorporating realistic frequency selectivity, however, the central-spectrum model is inaccurate. Culling et al. ([5], [6]) demonstrated that it predicts the wrong pitch for certain dichotic pitches and no pitch at all for others. Instead, Culling et al. showed that these cases can be explained using an alternative framework: the "modified equalization–cancellation" model [4]. This latter model correctly predicts the pitch of a range of dichotic pitches. It can also account for the binaural masking level difference and the binaural intelligibility level difference, as well as explain why listeners cannot group simultaneous sounds by common ITD but can group by common interaural decorrelation [26]. It does not, however, deal with the lateralization of dichotic pitches. Thus, for the problem of explaining the lateralization of a dichotic pitch the central-spectrum model is the only available account to date. However, if the central-spectrum model is to be questioned, on the grounds that it fails to correctly predict the perceived pitch of some dichotic pitches, then we are placed in the unattractive situation of having no adequate theory of lateralization. The aim of the present work is to explore a new approach to accounting for the lateralization for dichotic pitch.

In Section 2 we detail the resulting "reconstruction–comparison model" and show that it can account for the lateral position of N_0 and N_π Huggins pitches. In Section 3 we briefly describe the central-spectrum model. In Section 4 we illustrate the predictions of the two models for Raatgever and Bilsen's [21] experimental data concerning the effect of the ITD of the background noise on the lateral position of the Huggins pitch. In Section 5.2 the modified equalization–cancellation model is briefly outlined.

2. Description of the Reconstruction–Comparison Model

2.1 Introduction

Modern models of lateralization are based on Jeffress' [13] hypothesis that a neural network converts interaural time delays into a place code [2]. Spike trains from the left and right auditory nerves converge on a set of left–right coincidence detectors. An internal time delay is imposed on one of these spike trains. This internal time delay is progressively varied

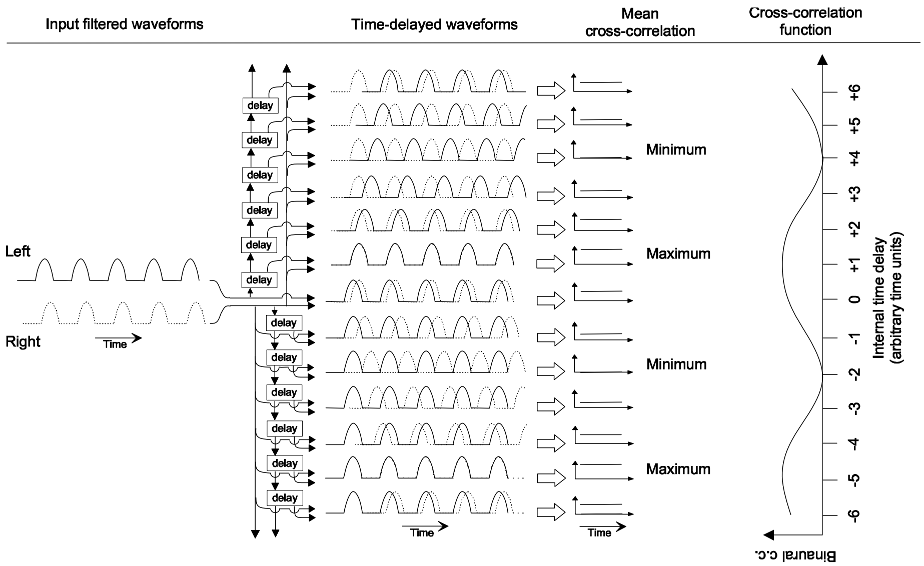


Figure 2 Schematic illustration of the Jeffress–Colburn binaural cross-correlation network. The inputs are the left waveform (solid line) and right waveform (dotted line) after auditory filtering of the incident sound. In this example, the incident sound is a sine wave with the right channel assigned an ITD of 1 time unit with respect to the left channel. The input waveforms arrive at cross-correlators after traversing additional internal delays of ±1 time unit, ±2 time units, and so on. The cross-correlators multiply the two delayed waveforms together. Where the delayed waveforms are in-phase, the mean cross-correlation is at a maximum. Where they are out-of-phase, the mean cross-correlation is at a minimum. The delay of the cross-correlator giving the maximum marks the interaural delay of the incident sound because that internal delay exactly compensates for the ITD. Note that the cross-correlation function is periodic in that additional peaks occur whenever the internal delay puts the delayed waveforms in-phase. Each peak is separated by the period of the center frequency of the auditory filter - 6 time units in this example.

across the network. The largest number of coincidences occur where the internal time delay exactly compensates for the ITD of the incident sound. The process is illustrated in Figure 2. The probability of a spike occurring is shown instead of the spikes themselves, thus allowing analyses to be based on cross-correlation rather than on coincidence detection. The end result is the binaural cross-correlation function.

The internal-delay axis defines the perceptual dimension of lateral position. Lateral position is presumed to be determined by the internal delays of the peaks in the cross-correlation function. All sounds exhibit multiple peaks in the cross-correlation function because of the short-term periodicities in their waveforms, but most sounds are heard in just one position, rather than many (c.f. [23]). For a sound such as a pure tone or the Huggins pitch, the lateral position is primarily determined by the peak closest to 0 μs. The internal delay of this peak in the cross-correlation function corresponds to the lateral position of the sound. A sound giving a peak at a positive delay is heard on the right of the head, at a negative delay is heard on the left of the head, and at zero is heard at the center of the head.

There is an independent set of coincidence detectors for each frequency channel. The resulting pattern of binaural cross-correlation versus frequency is termed a "cross-correlogram" (cf. Figure 3). Many results on the lateralization of single sounds presented in quiet can be explained using a complex decision strategy based on an across-frequency comparison of the cross-correlogram ([25], [27]). Nonetheless, a simpler strategy, based on averaging the cross-correlogram across frequency, works well for many single sounds (e.g., [24]). We used a similar strategy in the reconstruction–comparison model.

The reconstruction–comparison model incorporates three stages. First, it generates a cross-correlogram for the Huggins-pitch stimulus. Second, it separates this correlogram into two cross-correlograms, one for the Huggins pitch, the other for the background noise. This stage includes the reconstruction (or "filling in") of the notch in the noise cross-correlogram created by the removal of the frequency channels carrying the Huggins pitch. Third, it compares the Huggins-pitch cross-correlogram with the noise cross-correlogram by subtracting one from the other. It is the position of the peaks in the resulting, "remainder" cross-correlogram that determines the lateral position of the Huggins pitch. Each stage is described in the following sections (2.2, 2.3 and 2.4) using a 600-Hz N_0 Huggins pitch as the example. The application of the model to a 600-Hz N_π Huggins pitch is outlined in Section 2.5.

The model thus deals with the problem of the simultaneous perception of two distinct sounds, the Huggins pitch and the noise, by splitting the auditory scene into two separate objects and then computing the lateralization of the Huggins pitch. In a manner similar to certain other accounts of the lateralization of multiple sound sources, the reconstruction–comparison model determines which sound sources are present before it determines where these sound sources are located [7], [8], [12], [28].

2.2 Construction of the Cross-Correlogram of the Huggins-Pitch Stimulus

The first stage of the reconstruction–comparison model is illustrated schematically in Figure 3. The incident waveforms at the left and right ears are filtered using a matched pair of gammatone filterbanks [18], spanning center frequencies of 100 Hz to 1200 Hz at a resolution of 5 filters per equivalent rectangular bandwidth ("ERB") [10]. The output of each filter is halfwave rectified and then logarithmically compressed. The binaural cross-correlation functions are measured for each left/right pair of frequency channels resulting in the binaural cross-correlogram. The range of internal delays spans –5,000 μs to +5,000 μs at a resolution of 50 μs (each simulation was run using stimuli digitally sampled at 20,000 samples per second). A weighting function is applied to emphasize the contribution of frequency channels near 600 Hz and to attenuate the contribution of channels at lower and higher frequencies [19] [21].

A gammatone filterbank is commonly used for rapid computational modeling of the frequency selectivity of the inner ear. The frequency range and filters-per-ERB were chosen as a compromise between speed of modeling and resolution of frequency channels. The halfwave rectification and logarithmic compression simulate the transduction by the inner hair cells. Thus, their output represents the probability of a spike event in an auditory neuron (cf. Figure 2). The frequency-weighting simulates the dominant effect that frequencies near 600 Hz have on the lateralization of a wideband noise [19] and is included for consistency with the central-spectrum model [21].

The cross-correlogram illustrated in Figure 3 is for an N_0 noise without any Huggins pitch. The ridge at an internal delay of 0 μs marks its ITD; when projected onto the frequency-versus-internal-delay plane it lies on a straight line at 0 μs. The other ridges mark the multiple peaks resulting from the short-term periodicity in the filter outputs; when projected onto the frequency-versus-internal-delay plane they lie on hyperbolic curves.

The left panel of Figure 4 illustrates the cross-correlogram for an N_0 Huggins-pitch stimulus. The interaural phase shift gives a near-flat cross-correlation function at 600 Hz, thus

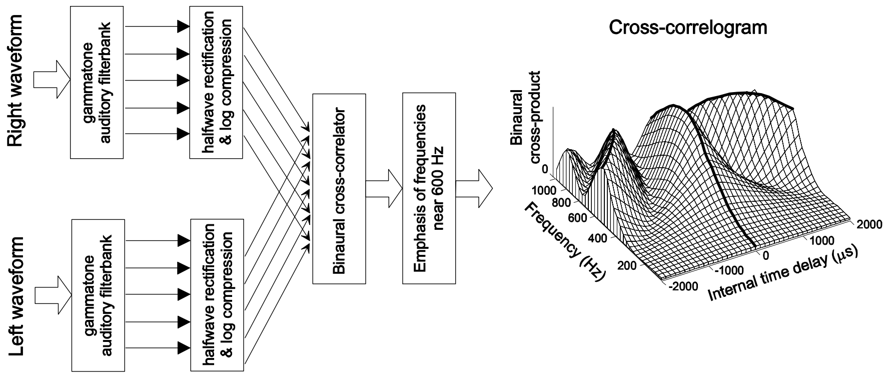

Right waveform
Left waveform
gammatone auditory filterbank
gammatone auditory filterbank
halfwave rectification & log compression
halfwave rectification & log compression
Binaural cross-correlator
Emphasis of frequencies near 600 Hz
Cross-correlogram
Binaural cross-product
0
1000
800
600
400
200
Frequency (Hz)
-2000
-1000
0
1000
2000
Internal time delay (μs)

Figure 3 Schematic illustration of the first stage of the reconstruction–comparison model. For visual clarity, just five frequency channels are shown. In the cross-correlogram, the spacing of frequency channels is reduced to 2.5 per ERB, the resolution of the internal delay is reduced to 100 μs and the span of internal delays is limited to ±2000 μs (these values are also used in all subsequent illustrations). The cross-correlogram is of an N_0 noise; the ridges are marked in bold.

partially filling in the valleys between the ridges of the noise. The right panel of Figure 4 shows the result of averaging this cross-correlogram across frequency (cf. [24]). The straight ridge gives rise to the large peak at an internal delay of 0 μs, whereas the hyperbolic ridges cancel out. This averaging strategy was first used by Shackleton et al. [24] in their model of the lateralization of single sounds. They showed that it gave a successful account of that problem but it fails in this instance for the Huggins pitch. The peak in the across-frequency average is at 0 μs, corresponding to the center of the head, and so it marks the lateral position of the noise, rather than the Huggins pitch. The failure occurs because the noise dominates the across-frequency average.

2.3 Separation into Two Cross-Correlograms and Reconstruction of the Noise Cross-Correlogram

The second stage of the model separates the initial cross-correlogram into two separate cross-correlograms, one for the Huggins pitch and the other for the background noise. The cross-correlation functions in frequency channels occurring within ±0.5 ERBs of the center frequency of the Huggins pitch are placed within the "Huggins-pitch cross-correlogram" (Figure 5, left panel) while the remaining functions are put into the "noise cross-correlo-

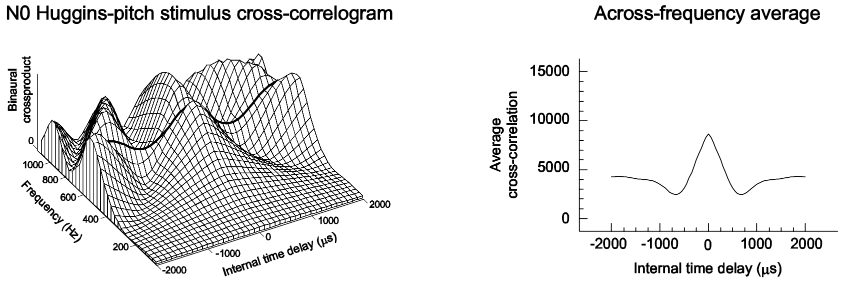
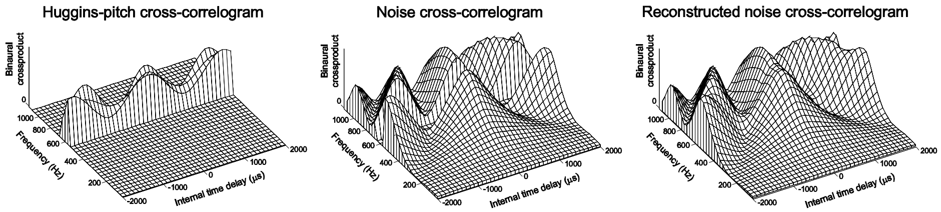


Figure 4 Left panel: the cross-correlogram for a 600-Hz N_0 Huggins-pitch stimulus. The 600-Hz frequency channel is highlighted in bold. Compare with the cross-correlogram for an N_0 noise shown in Figure 3. Right panel: the across-frequency average of the cross-correlogram.


Huggins-pitch cross-correlogram

Noise cross-correlogram

Reconstructed noise cross-correlogram

Figure 5 Left panel: the Huggins-pitch cross-correlogram. Middle panel: the noise cross-correlogram. Note the 1-ERB wide notch centered on 600 Hz. Right panel: the noise cross-correlogram after reconstruction by linear interpolation. Note that the notch is filled-in.

gram" (middle panel). The results are a Huggins-pitch cross-correlogram that covers exactly 1 ERB and a noise cross-correlogram with a corresponding notch. This notch is then filled in by linear interpolation, yielding a "reconstructed noise cross-correlogram" (right panel).

The use of two separate cross-correlograms implements the idea that the Huggins pitch and the background noise are separate auditory objects. The range of channels assigned to the Huggins-pitch cross-correlogram is a free parameter in the model. We used a value of 1 ERB because it is reasonable that a tonal object should have a narrow bandwidth. A computational strategy for measuring the center frequency of the dichotic pitch is sketched in section 5.2. The reconstruction strategy is crucial for the success of the model. It implements the idea that auditory analysis recreates the parts of the noise hidden by the Huggins pitch.

2.4 Comparison of Cross-correlograms and the Calculation of the Lateral Position of the Huggins Pitch

The third stage of the model compares the Huggins-pitch cross-correlogram with the reconstructed noise cross-correlogram by subtracting the latter from the former. Only positive remainders are retained (Figure 6, left panel). This "remainder cross-correlogram" is then averaged across frequency (right panel), and the peak closest to 0 μs is selected. Its internal delay corresponds to the lateral position of the Huggins pitch.

In the illustration there are two peaks in the remainder cross-correlogram, one at –850 μs and the other at +800 μs. It should be noted that these peak positions were determined using a single, 500-ms duration noise as the input to the model. The exact positions vary somewhat because of random fluctuations inherent in any noise. Repeating the simulation 100 times showed that the peaks fell into two ranges, of –900 to –750 μs and +750 to +850 μs. Therefore, it is reasonable to assume that the two peaks are located at about –800 μs and +800 μs and thus are equally distant from 0 μs. These internal delays correspond to lateral positions on the far left and far right of the head, respectively. Thus, the reconstruction–comparison model can account for why an N_0 Huggins pitch is lateralized to the edges of the head. In order to explain why some listeners hear the Huggins pitch on the left but others hear it on the right, the further assumption is needed, that each listener's auditory system contains a bias towards the left or right, perhaps because of small left/right asymmetries in the auditory pathway. This bias will result in one peak being selected in preference to the other.

2.5 The Reconstruction–Comparison Model Applied to the N_π Huggins Pitch.

Figures 7, 8, and 9 illustrate the set of cross-correlograms for a 600-Hz N_π Huggins pitch. In the final, remainder cross-correlogram, three peaks are created within the illustrated internal-delay range. The middle peak is selected as it is closest to 0 μs. In the illustration it

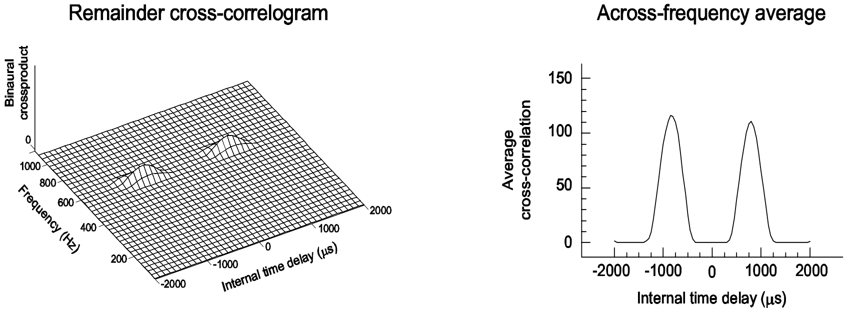


Figure 6 Left panel: The remainder after subtracting the reconstructed noise cross-correlogram from the Huggins-pitch cross-correlogram. Only positive remainders are shown. The peaks look small because the vertical scale is the same as in Figures 4 and 5. Right panel: the cross-correlogram after averaging across frequency. The reconstruction–comparison model proposes that the lateral position of the Huggins pitch corresponds to the internal delay of the peak closest to 0 μs.

is located at an internal delay of +100 μs. Again, the exact position varies somewhat: repeating the simulation 100 times shows that the range of peak positions varies from –100 μs to +100 μs. These values bracket 0 μs, and thus the reconstruction–comparison model can account for why an N_π Huggins pitch is lateralized close to the center of the head.

3. Intermission: the Central-Spectrum Model

The analyses above demonstrate that the reconstruction–comparison model can account for the lateralization of both N_0 and N_π Huggins pitches. It therefore provides an alternative to Raatgever and Bilsen's central-spectrum model [21]. Before comparing the predictions of the reconstruction–comparison and central-spectrum models, we briefly describe the concept of a central spectrum.

A central spectrum is an across-frequency slice of a cross-correlogram – a plot of cross-correlation versus frequency for a fixed internal delay. There is one central spectrum for each value of internal delay. Raatgever and Bilsen [20] [21] proposed that a selection mechanism chooses one or more central spectra, citing harmonicity, modulation depth, the pronouncement of a peak and spectral pattern recognition as possible criteria for selection. A peak in the chosen spectrum is heard as a dichotic pitch, and the internal delay of the chosen spectrum corresponds to the lateral position of the dichotic pitch.

Raatgever and Bilsen [21] argued that the selection mechanism would choose two central spectra, at ± 800 μs, for a 600-Hz N_0 Huggins pitch, but would choose a single central spec-

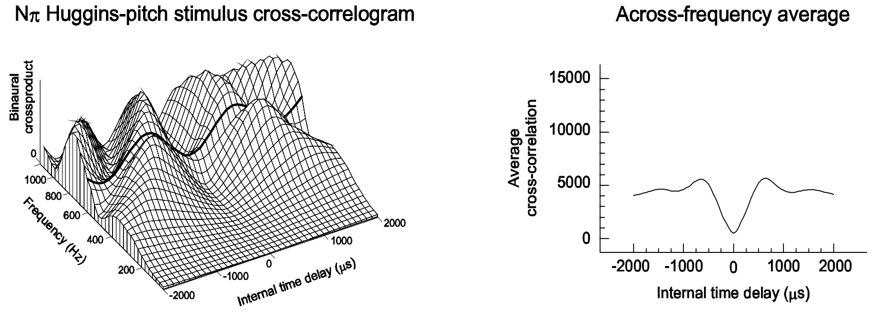


Figure 7 Same as in Figure 4, but for a 600-Hz N_π Huggins pitch.

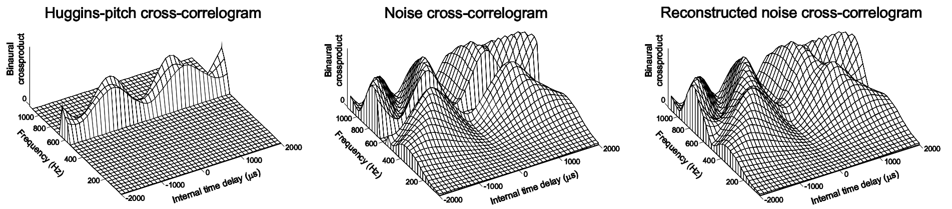


Figure 8 Same as in Figure 5, but for a 600-Hz N_π Huggins pitch.

trum, at 0 μs, for a 600-Hz N_π Huggins pitch. The choice is made because these central spectra contain pronounced peaks at 600 Hz (Figure 10) and because they are "characteristic for the pitch perceived" [21, p. 432]. Because the internal delay of the chosen spectrum is proposed to correspond to the lateral position, without any intermediate processing or analysis to affect the directness of the correspondence, the choice of central spectrum is critical for the prediction of lateral position. This choice is correct for N_0 and N_π Huggins pitches, in that the internal delays of the chosen spectra do indeed correspond to their lateral positions. However, without a principled method for specifying the choice this result may be fortuitous. The criteria for choosing a central spectrum must be explicitly defined before the model's account of lateralization can be tested rigorously.

4. An Application of the Reconstruction–Comparison Model

Raatgever and Bilsen [21] noted that the introduction of an external ITD to a Huggins-pitch stimulus should result in the corresponding translation of its cross-correlogram along the internal-delay axis. The spectral profile of the central spectrum chosen by the selection mechanism would be unchanged, in that the translation will not affect the variety of criteria upon which the choice is made, but its internal delay will change by an amount equal to the introduced ITD. Raatgever and Bilsen thus argued that the central-spectrum predicts that the lateral position of the Huggins pitch depends on the ITD of the stimulus. Numerically, the ITD corresponding to the lateral position is expected to vary 1:1 with the applied ITD. They collected experimental data in broad support of this prediction. Raatgever and Bilsen's prediction, their data and our own predictions using the reconstruction–comparison model are described in the following sections.

4.1 The Lateralization of an N_0 Huggins Pitch

The left panel of Figure 11 reproduces Raatgever and Bilsen's [21] results. In their

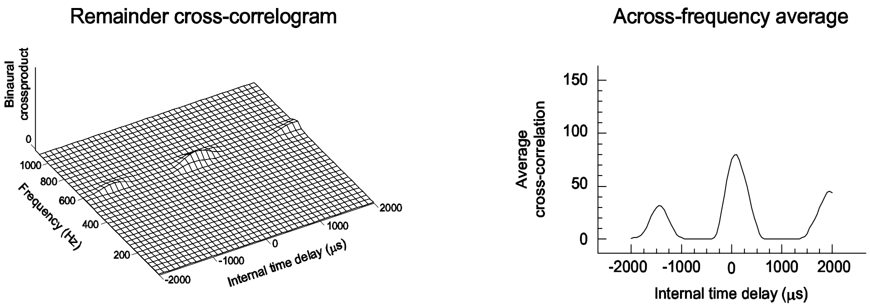


Figure 9 Same as in Figure 6, but for a 600-Hz N_π Huggins pitch.

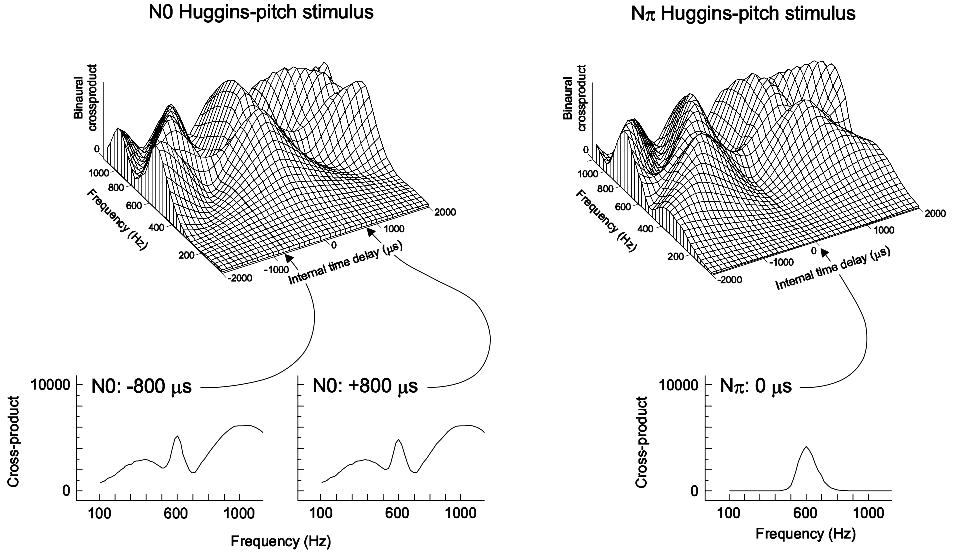


Figure 10 Left column: the selected central spectra for a 600-Hz N_0 Huggins pitch. Right column: the selected central spectrum for a 600-Hz N_π Huggins pitch. The central spectra are marked in bold on the cross-correlograms of the Huggins-pitch stimuli shown in the top row. Note that the cross-correlograms and central spectra are based on the output of a gammatone filterbank (cf. Figure 3) instead of the Fourier transform used by Raatgever and Bilsen [21].

method they required listeners to manipulate the lateral position of a 600-Hz N_0 Huggins pitch so that it was heard in the same position as a comparison white noise. The manipulation was to vary the external ITD applied to the Huggins-pitch stimulus ("matched ITD"). The lateral position of the comparison noise was randomly varied across successive trials by manipulating its ITD ("comparison ITD"). Each symbol in Figure 11 plots the matched ITD from each trial for two listeners (circles and asterisks). The solid lines show Raatgever and Bilsen's own predictions for the central-spectrum model.

Two effects are of interest. First, the matched ITD clearly depends upon the comparison ITD, thus demonstrating that the lateral position of the Huggins pitch can be varied by manipulating the interaural configuration of the noise carrier. The pattern of data points broadly supports the predictions of the central-spectrum model, although the scatter is large. Second, the sign of the matched ITD could be ambiguous across listeners: compare the circles with the asterisks at a comparison ITD of 0 μs. This ambiguity indicates that one listener applied a positive ITD in order to shift the lateral position of the Huggins pitch to the center of the head but the other listener applied a negative ITD. Thus, the first listener originally heard the Huggins pitch on the left side of the head but the second listener originally heard it on the right.

The right panel of Figure 11 shows the predictions of the reconstruction–comparison model. The method was to first measure the internal delay of the closest-to-0-μs peak in the N_0 Huggins pitch and then to compute the ITD offset required to shift this peak in order for its internal delay to be equal to the comparison ITD. The positive/negative ambiguity in the closest-to-0-μs peak (cf. Figure 6) is illustrated by plotting the results from positive-internal-delay peaks as open circles and negative-internal-delay peaks as filled circles. The model was run 10 times for each value of the comparison ITD, each time using an independent

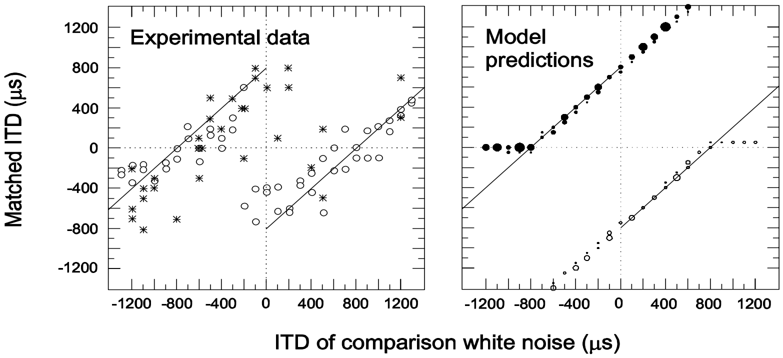


Figure 11 Left panel: experimental data pertaining to the lateralization of a 600-Hz N_0 Huggins pitch, redrawn from Figure 7 of Raatgever and Bilsen [21] (reproduced with permission). The task was to vary the ITD applied to a 600-Hz Huggins pitch until it was heard in the same place as a comparison white noise of known ITD. The results are shown for two listeners (circles and asterisks). The solid lines show the predictions of the central-spectrum model. Right panel: predictions from the reconstruction–comparison model. The model was run 10 times for each value of comparison ITD; the radius of each circle represents the number of times each matched-ITD was found.

noise. The radius of each circle is proportional to the number of times that each matched ITD was found. The diagonal lines again show Raatgever and Bilsen's [21] predictions for the central-spectrum model.

For comparison ITDs within ±800 μs, the predictions of the reconstruction–comparison model fall close to the diagonal lines. For these stimuli, therefore, the reconstruction–comparison model makes the same predictions as the central-spectrum model. By extension, the reconstruction–comparison model can explain the general pattern observed in the experimental data.

There are, however, two differences between the predictions and the experimental data. First, the scatter of predicted data points is smaller than the scatter of the listeners' matches. This result occurs because the reconstruction–comparison model is deterministic (the random fluctuations inherent in a noise create what scatter there is). In order to generate a scatter comparable to that observed experimentally it would be necessary to add an internal noise to the model.

Second, if the comparison noise has an ITD less than ca. –800 μs, or greater than ca. +800 μs, the reconstruction–comparison model predicts that the matched ITD to be ca. 0 μs. This prediction follows from the fact that the period of the 600-Hz auditory filter is 1667 μs. Because the multiple peaks in any cross-correlation function are separated by the period of the auditory filter (cf. Figure 2), the multiple peaks in the 600-Hz cross-correlation function are separated by about 1600 μs. Thus, there must be a peak somewhere in the range –800 to +800 μs. This peak will always be selected by the closest-to-0-μs strategy. Consequently, the largest possible lateralization of a 600-Hz Huggins pitch corresponds to an internal delay of about –800 or +800 μs, and thus it cannot be accurately matched to a comparison ITD less than about –800 μs or greater than about +800 μs. The best that the model can do is to leave the ITD of the Huggins-pitch stimulus fixed at 0 μs. Thus, the matched ITD is predicted to be 0 μs. Raatgever and Bilsen's [21] experimental data do not support this prediction. One strategy for bringing the model into correspondence with the data is to allow the selection of a peak other than that one closest to 0 μs. Nonetheless, the reconstruction–comparison model offers a successful account of the primary effects observed as the ITD of an N_0 Huggins pitch is varied.

4.2 The Lateralization of an N_π Huggins Pitch

In a second set of conditions Raatgever and Bilsen [21] repeated the experiment but required listeners to manipulate a 600-Hz N_π Huggins pitch instead of an N_0 Huggins pitch. The left panel of Figure 12 reproduces their results along with their predictions from the central-spectrum model. Two effects are of interest. First, the general pattern is shifted by ca. 800 μs (approximately one half-period of 600 Hz) from the general pattern illustrated in Figure 11. Second, the data points again broadly support the predictions of the central-spectrum model.

The right panel of Figure 12 shows the predictions of the reconstruction–comparison model. For the majority of comparison ITDs, the predictions of the reconstruction–comparison model fall close to the diagonal lines marking the predictions of the central-spectrum model. By extension, the reconstruction–comparison model can explain the primary effects observed in the data.

5. Discussion

Raatgever and Bilsen's ([21]) central-spectrum model is the only comprehensive account of the pitch and lateralization of dichotic pitch to have been published. Culling et al. [5] [6] demonstrated that this model cannot account for the pitch of certain dichotic pitches, showing instead that their modified equalization–cancellation model could account for these and other cases. We therefore questioned the account of lateralization of a dichotic pitch given by the central-spectrum model. Our objective in this chapter was to evaluate a computational model of lateralization that might serve as an alternative to the central-spectrum model. As demonstrated in Section 4, the resulting reconstruction–comparison model can account for the major features of the experimental data on lateralization of a Huggins pitch.

The modified equalization–cancellation model and the reconstruction–comparison model remain separate because each was designed to explain a different perceptual characteristic of a dichotic pitch. Neither model, considered alone, can offer the comprehensive account of pitch and lateralization that was the objective of the central-spectrum model. In the remainder of this discussion we outline a future strategy for combining the two into a

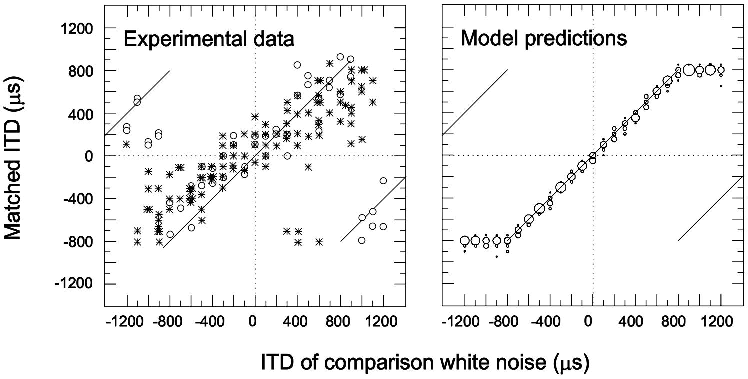


Figure 12 As per Figure 11, but for a 600-Hz N_π Huggins pitch.

single, comprehensive model. It is based on the fact that the outputs of the left-right pair of auditory filters centered on the Huggins pitch are interaurally decorrelated.

5.1 The Source of the Interaural Decorrelation

Interaural decorrelation is the opposite of interaural correlation ("loosely defined as the point-by-point correlation coefficient computed for a stimulus segment after an appropriate [internal] delay is imposed on one of the inputs to maximize the correlation" [11, p. 302-303]). For example, if the input to a left-right pair of auditory filters is a noise band with an ITD of –300 μs, then their outputs also have an ITD of –300 μs, and so delaying one output by +300 μs will compensate for this ITD. This delay will bring the outputs into exact, sample-for-sample, correspondence, and so they would be interaurally correlated. If, instead, the input is a noise band with an ITD progressively increasing with frequency from –300 μs to +300 μs, no single delay can compensate. The correspondence between the two outputs is thus minimal, and so the outputs are interaurally decorrelated.

This second example is deliberately similar to that of a 600-Hz Huggins pitch. The introduction of a 60-Hz-wide interaural phase shift creates a progressive shift in ITD (cf. Figure 1). The bandwidth is of the same order of magnitude as the bandwidth of an auditory filter (for the 600-Hz auditory filter the [equivalent-rectangular] bandwidth is approximately 90 Hz) and so the outputs of the left-right 600-Hz auditory filters will be interaurally decorrelated. The decorrelation will be largest at 600 Hz, as that is the center frequency of the interaural phase shift. Consequently, the frequency of the Huggins pitch can be measured using interaural decorrelation.

5.2 The Measurement of Interaural Decorrelation

The reconstruction–comparison model requires the frequency of the Huggins pitch to be known. If it were not known then the model would be unable to separate the Huggins-pitch cross-correlogram from the noise cross-correlogram because it would not know which frequency channels to separate. In our description of the model (Section 2.3) we assumed that this frequency is known. Preferably, the modified equalization–cancellation model would be used to measure the interaural decorrelation.

The modified equalization–cancellation model is based on a frequency-analysis stage similar to that used in the reconstruction–comparison model (cf. Figure 3). It equalizes the amplitude of the outputs of corresponding left and right frequency channels and then cancels the two outputs by subtracting one from the other, sample by sample, as a function of an internal delay. In essence, the model performs cross-subtraction instead of cross-correlation. The smallest subtraction remainder in each frequency channel is measured; the plot of smallest remainder versus frequency is termed the "recovered spectrum." Culling and Summerfield [4] have shown that the recovered spectrum is a plot of interaural decorrelation versus frequency. The recovered spectra for a 600-Hz N_0 and a 600-Hz N_π Huggins pitch are shown in Figure 13. Both spectra contain clear peaks at 600 Hz. Culling et al. [5], [6] demonstrated that the pattern of peaks in the recovered spectrum successfully predicts the perceived pitch of several forms of dichotic pitch, including the Huggins pitch. For the Huggins pitch the peak is at 600 Hz, so determining the frequency of the pitch and allowing the reconstruction-cancellation model to know which frequency channels to separate.[1]

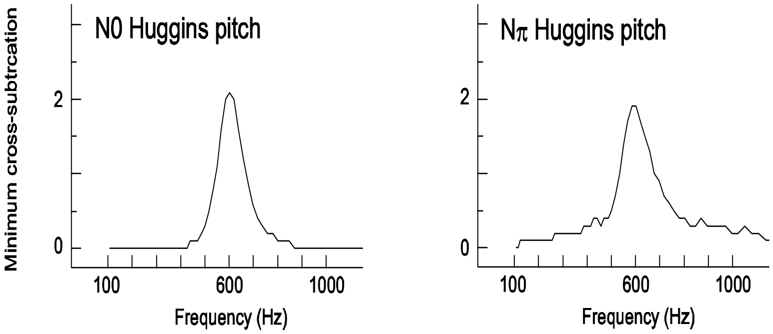


Figure 13 Plots of the smallest remainder after cross-subtraction versus frequency ("recovered spectra") for the 600-Hz N_0 and N_π Huggins pitches, created by a computational implementation of the modified equalization–cancellation model [4]. Note that the implementation is based on the peripheral processing model shown in Figure 3 and thus differs from Culling and Summerfield's [4] description primarily in using halfwave rectification and logarithmic compression instead of Meddis' [15] [16] hair-cell model.

6. Conclusion

We have described a new model of the lateralization of the Huggins pitch. It is based on three fundamental ideas. The first is that the auditory system creates two objects, one for the Huggins pitch and one for the noise. The second is that the parts of the noise hidden by the Huggins pitch are then reconstructed. The third is that the lateralization of the Huggins pitch is subsequently determined by a comparison of the Huggins pitch with the background noise. This "reconstruction–comparison model" can account for the overall pattern of listeners' judgements pertaining to the lateralization of the Huggins-pitch stimulus as a function of ITD.

The reconstruction–comparison model is offered as an alternative approach to accounting for the lateralization to that provided by Raatgever and Bilsen's central-spectrum model (e.g., [21]). It remains to be demonstrated that a comprehensive account of both the lateralization and the pitch of a dichotic pitch can be obtained by combining the reconstruction–comparison model with Culling and Summerfield's modified equalization–cancellation model [4] [5] [6] [26].

Acknowledgements

We thank Trevor Shackleton and John Culling for their comments and advice throughout this research. A further development and detailed description of the ideas in this paper will be found in [1].

Note

1. Because of the random fluctuations inherent in a noise, it is expected that there would be a small variation in the exact frequency of the peak in the recovered spectrum. In turn, this would create a small variation in which frequency channels are separated, so leading to an increase in the predicted scatter of the lateralizations (cf. Figures 11 and 12).

References

[1] Akeroyd, M. A. and Summerfield, A. Q. "The lateralization of simple dichotic pitches." *J. Acoust. Soc. Am.*, 108: 316–334, 2000.

[2] Colburn, H. S. “Theory of binaural interaction based on auditory nerve data. II. Detection of tones in noise.” *J. Acoust. Soc. Am.*, 61: 525–533, 1977.

[3] Cramer, E. M. and Huggins, W. H. “Creation of pitch through binaural interaction.” *J. Acoust. Soc. Am.*, 30: 413–417, 1958.

[4] Culling, J. F. and Summerfield, A. Q. “Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay.” *J. Acoust. Soc. Am.*, 98: 785–797, 1995.

[5] Culling, J. F., Summerfield, A. Q. and Marshall, D. H. “Dichotic pitches as illusions of binaural unmasking. I. Huggins’ pitch and the ‘binaural edge pitch’.” *J. Acoust. Soc. Am.*, 103: 3509–2526, 1998.

[6] Culling, J. F., Marshall, D. H. and Summerfield, A. Q. “Dichotic pitches as illusions of binaural unmasking. II. The Fourcin pitch and the dichotic repetition pitch.” *J. Acoust. Soc. Am.*, 103: 3527–3539, 1998.

[7] Darwin, C. J. “Auditory grouping.” *Trends in Cognitive Science*, 1: 327–333, 1997.

[8] Darwin, C. J. and Hukin, R. W. “Auditory objects of attention: The role of interaural time differences.” *J. Exp. Psych: Human Perception and Performance,* in press.

[9] Frijns, J. H. M., Raatgever, J. and Bilsen, F. A. “A central spectrum theory of binaural processing. The binaural edge pitch revisited.” *J. Acoust. Soc. Am.,* 80: 441–451, 1986.

[10] Glasberg, B. R. and Moore, B. C. J. “Derivation of auditory filter shapes from notched-noise data.” *Hear. Res.*, 47: 103–138, 1990

[11] Grantham, D. W. “Spatial hearing and related phenomena.” In *Hearing*, B. C. J. Moore (ed.), London: Academic Press, pp. 297–345, 1995.

[12] Hill, N. I. and Darwin, C. J. “Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning.” *J. Acoust. Soc. Am.*, 100: 2352–2364, 1996.

[13] Jeffress, L. A. “A place theory of sound localization.” *J. Comp. Physiol. Psych.*, 41: 35–39, 1948.

[14] Licklider, J. C. R. “Auditory frequency analysis.” In *Proc. Third Symp. Inform. Theory*, C. Cherry (ed.), New York: Academic Press, pp. 253–268, 1956.

[15] Meddis, R. “Simulation of mechanical to neural transduction in the auditory receptor.” *J. Acoust. Soc. Am.*, 79: 702–711, 1986.

[16] Meddis, R. “Simulation of auditory-neural transduction: Further studies.” *J. Acoust. Soc. Am.*, 74: 750–753, 1988.

[17] Moore, B. C. J. *An Introduction to the Psychology of Hearing*. London: Academic Press, 1997.

[18] Patterson, R. D., Allerhand, M. H., and Giguère, C. “Time-domain modeling of peripheral auditory processing: A model architecture and a software platform.” *J. Acoust. Soc. Am.*, 98: 1890–1894, 1995.

[19] Raatgever, J. *On the Binaural Processing of Stimuli with Different Interaural Phase Relations*. Doctoral Dissertation, Delft University of Technology, 1980.

[20] Raatgever, J. and Bilsen, F. A. “Lateralization and dichotic pitch as a result of spectral pattern recognition.” In *Psychophysics and Physiology of Hearing*, E. F. Evans and J. P. Wilson (eds.), London: Academic Press, pp. 443–453, 1977.

[21] Raatgever, J. and Bilsen, F. A. “A central spectrum theory of binaural processing. Evidence from dichotic pitch.” *J. Acoust. Soc. Am.*, 80: 429–441, 1986.

[22] Raatgever, J. Bilsen, F. A. and Mungra, R. “New experiments beyond the traditional Fourcin pitch range.” In *Proc. 16th Int. Cong. Acoust. and 135th Acoust. Soc. Am., Volume 1,* pp. 165–166, 1998.

[23] Shackleton, T. M., and Bowsher, J. M. “Binaural effects of the temporal variation of a masking noise upon the detection thresholds of tone pulses.” *Acustica*, 69: 218–225, 1989.

[24] Shackleton, T. M., Meddis, R. and Hewitt, M. J. “Across frequency integration in a model of lateralization.” *J. Acoust. Soc. Am.*, 91: 2276–2279, 1992.

[25] Stern, R. M., Zeiberg, T. and Trahiotis, C. “Lateralization of complex binaural stimuli: A weighted image model.” *J. Acoust. Soc. Am.*, 84: 156–165, 1988.

[26] Summerfield, Q. and Akeroyd, M. A. “Computational approaches to modeling auditory selective attention: Monaural and binaural processes.” In *Course Reader for the NATO ASI on Computational Hearing, volume 2,* S. Greenberg and M. Slaney (eds.), pp. 743–800, 1998.

[27] Trahiotis, C. and Stern, R. M. “Across-frequency interaction in lateralization of complex binaural stimuli.” *J. Acoust. Soc. Am.*, 96: 3804–3806, 1994.

[28] Woods, W. S. and Colburn, H. S. “Test of a model of auditory object formation using intensity and interaural time difference discrimination.” *J. Acoust. Soc. Am.*, 91: 2894–2902, 1992.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

PROCESSING OF PITCH INFORMATION IN COMPLEX STIMULI BY A MODEL OF OCTOPUS CELLS IN THE COCHLEAR NUCLEUS

Yidao Cai, JoAnn McGee, Edward J. Walsh

Developmental Auditory Physiology Laboratory
Boys Town National Research Hospital
555 North 30th Street, Omaha, NE 68131, USA

1. Introduction

It is widely accepted that pitch is an important attribute of speech. However, how pitch information is processed in the central auditory system is largely unknown. Various models of pitch perception have been proposed on the basis of psychophysical studies [7][9] [17][19], but the physiological mechanisms underlying the process are poorly understood. For example, although responses of neurons in the cochlear nucleus complex (CN) to complex stimuli have been studied [12], their role in processing pitch information remains unclear.

One of the principle neuronal classes in the posteroventral cochlear nucleus is the octopus cell. Cells in this category respond to tonal stimulation mainly at the time of stimulus onset and are thus commonly referred to as "onset" responders. Octopus cells are thought to play an important role in pitch perception [6][10][12]. They carry precise temporal information in the timing of action potentials comprising spike trains resulting from acoustic stimulation. This property is determined, at least in part, by a low input impedance and a low-threshold potassium (K^+) channel [2][6][7].

In this study, we used a computer model of octopus cells to examine the processing of pitch information contained in complex stimuli. Inputs to the model were auditory-nerve fiber spike trains recorded from anesthetized cats. Harmonic or "inharmonic" stimuli, similar to those used in psychophysical studies, were used to record data from the cat auditory nerve as well from the cochlear nucleus of the gerbil. Our simulation results demonstrate that octopus cells take advantage of converging inputs from an array of auditory-nerve fibers spanning a wide frequency range to process pitch information. This finding is consistent with experimental studies of CN neurons *in vivo* and support the hypothesis that interspike interval information is a correlate of pitch.

2. Methods

2.1 The Model

The model used in this investigation was developed explicitly to study the mechanism(s) whereby onset responses are generated by octopus cells [3]. As shown in Figure 1, the model consists of a soma, an axon, and four identical dendrites. The axon and soma are each represented by a single compartment and each dendrite is represented by 20 compartments. The axon and soma compartments contain Hodgkin–Huxley-like sodium (Na^+) and K^+ channels. The soma compartment contains two additional active mechanisms: a low-threshold K^+

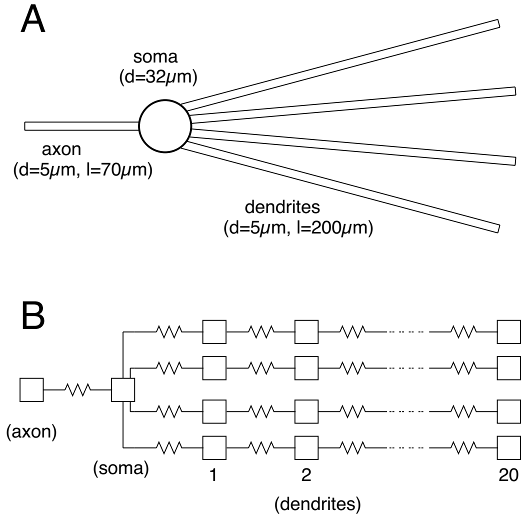


Figure 1 A. The octopus cell model. B. Compartmental representation of the model.

channel, K_{LT}, and a Cs^{+}-sensitive, hyperpolarization-activated inward rectifier, I_h. The dendrites are passive and have a space constant of 354 μm. The resting membrane potential is −62 mV.

2.2 *The Stimuli*

The inputs to the model were auditory-nerve (AN) fiber spike trains, collected from adult cats. In addition to tone bursts of different frequencies, responses to harmonic and "inharmonic" complex stimuli, similar to those used in psychophysical studies [11][15-20], were recorded. As shown in Figure 2, two groups of stimuli were employed: in one group (A-E) the stimuli contained three components centered at 1000 Hz, and in the other group (F-H), the stimuli contained 6 components between 1000 and 2000 Hz with 200-Hz spacing. Except for the "inharmonic" complex shown in Figure 2E, which is a frequency-shifted version of B, all stimuli produce the same pitch of 200 Hz (equivalent to the fundamental frequency), although the fundamental component is present in only two stimuli (A and F). Psychophysical studies have shown that the fundamental component is not essential for pitch perception (the so-called phenomenon of the "missing fundamental"). The frequency-shifted stimulus (E) was chosen because this stimulus, with a frequency spacing of 200 Hz, yields a pitch slightly higher than 200 Hz. The amplitude-modulated (AM) stimuli (C and D) have the same frequency components but their temporal waveforms are inverted versions of each other. Due to the rectifying characteristic of the inner-hair-cell transfer function, only the positive portion of the waveform is utilized when generating the spike events on the AN fibers. Psychophysically both stimuli sound the same despite the difference in temporal waveforms. The stimulus shown in Figure 2H was generated using the same six frequency components as those in Figure 2G, except that the phase of each component was randomized between 0 and 360 degrees, while their counterparts in Figure 2F and G all have a starting phase of zero degrees. Compared to the zero-phase versions, the random-phase version lacks temporal periodicity in its waveform. Psychophysical studies suggest that phase has no effect on the perception of pitch [11]. Except for the AM stimuli, all individual components have the same amplitude and the sound pressure level (SPL) pertains to the entire waveform.

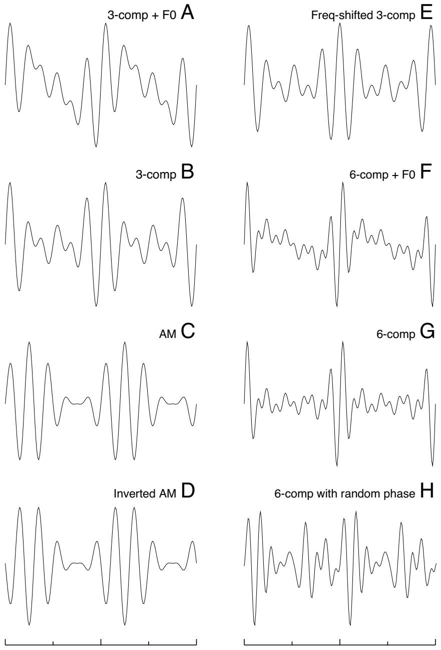


Figure 2 Temporal waveforms of stimuli used in the collection of auditory-nerve spike trains. The “3-comp” refers to a three-component complex of 800, 1000 and 1200 Hz, while the “6-comp” refers to a six-component complex of 1000–2000 Hz with 200-Hz spacing between the harmonics. Except for the frequency-shifted, three-component complex (E, 850, 1050, 1250 Hz), all stimuli have a fundamental period of 5 ms. Waveforms are normalized according to their respective peak magnitude and are shown for a 10-ms time window. Except for the AM stimuli (C and D), all frequency components in each complex have the same magnitude. F_0 refers to the fundamental frequency (200 Hz).

2.3 Auditory Nerve and Cochlear Nucleus Neuron Spike Train Data

Standard experimental procedures were used to collect data from the auditory nerve of cats and the cochlear nucleus of gerbils. Adult cats were deeply anesthetized using sodium pentobarbital (40 mg/kg, i.p.), the pinna was removed, and a craniotomy was performed to gain access to the posterior fossa. Cerebellar tissue was aspirated to expose the root of the auditory nerve. A glass microelectrode filled with 2 M KCl, and with an impedance of 15–20 MΩ, was inserted into the auditory nerve. Before recording, a routine calibration curve was obtained. Stimuli were generated using both the amplitude and phase characteristics of the calibration. The stimuli were 50 ms in duration, had a repetition interval of 120 ms, and were presented 50 times. The same stimulus conditions were used to collect data from each AN fiber encountered.

The procedure for collecting responses from CN neurons was similar, but adult gerbils were used. Sodium pentobarbital (50 mg/kg, i.p.) was used in conjunction with ketamine HCl (30 mg/kg, i.m.) to anesthetize the animals. The cochlear nucleus was exposed using standard surgical procedures, and recordings made from the posteroventral division. The stimuli used to collect data from the AN were also used in recording from the CN. The care and use of the animals were approved by the Boys Town IACUC.

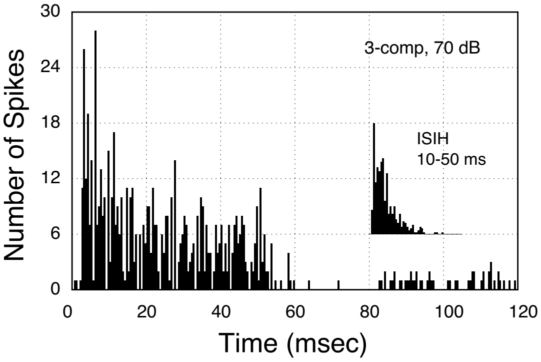


Figure 3 Responses of an auditory-nerve fiber (CF = 1050 Hz, SR = 63 spikes/s, threshold = 13 dB SPL) to a three-component complex (cf. Figure 2B) at 70 dB SPL. The ISIH for spikes falling between 10 and 50 ms from stimulus onset is shown in the insert. The histogram bin width is 0.5 ms. The maximum bin height in the ISIH corresponds to 60 occurrences and the abscissa ranges between 0 and 40 ms.

2.4 Application of Auditory Nerve Inputs to the Model

Figure 3 shows the responses of a typical AN fiber to a three-component harmonic complex (cf. Figure 2B) presented at 70 dB SPL. Although synchronization to the fundamental frequency (200 Hz) can be seen, the temporal response is “noisy” because of synchronization to individual harmonic components. Spike trains from six AN fibers were chosen as inputs to the model in order to mimic the natural input to octopus cells (incorporating a broad range of characteristic frequencies [CFs] and high spontaneous rates [SRs]). The fibers selected have CFs between 1050 Hz and 3250 Hz and they all have high SRs. To increase the number and temporal variability of inputs, we distributed each spike train over multiple locations. By varying the number of the starting trial and applying the trials in a circular manner (e.g., trials 7, 8,..., 50, 1,..., 6), we ensured that no two inputs were identical at a given moment in time. A total of 120 inputs were distributed across different locations of the model: 40 at the soma, and the remainder at different dendritic compartments.

Only excitatory inputs were used because there is little evidence suggesting the existence of inhibitory inputs onto octopus cells. In addition, a previous study showed that inhibitory inputs are not needed to generate the basic onset response pattern observed in octopus cells [3]. The dynamics of the synaptic conductance were modeled by an alpha function. The maximum synaptic conductance, adjusted to produce little or no spontaneous firing, had a value of 3.68 nS.

2.5 Simulation

Simulations were performed on a PC running Linux (a PC-based version of UNIX), with a program developed in our laboratory [2]. Simulation of 50 trials (120 ms per trial) requires about 13 minutes to complete when running on a Pentium-133 computer. The parameters used in this study were almost identical to those used in our previous current injection simulations [3], except for the implementation of faster K_{LT} kinetics, an adjustment that resulted from knowledge gained from simulations using spike trains collected with tone bursts of different frequencies. Octopus cells are typically associated with both O_I (a response peak at stimulus onset with little or no steady-state response) and O_L response patterns (with steady-state responses, usually > 10 spikes/s.). We adjusted the model parameters such that the model neuron produced phase-locked responses (entrainment) at lower frequencies (< 500 Hz). This

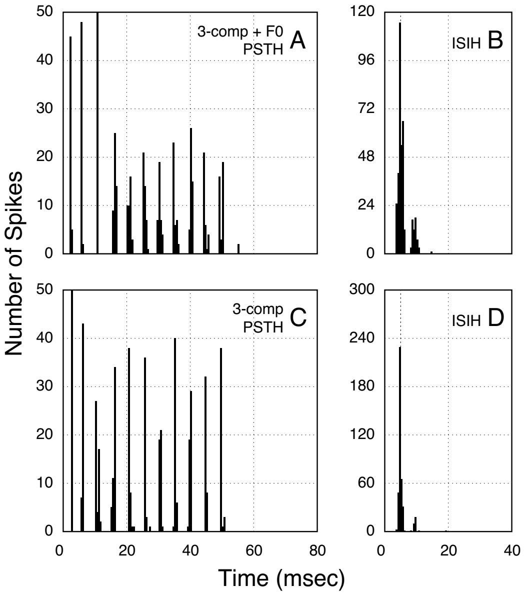


Figure 4 Responses of the model to three-component complexes with (upper panels) and without (lower panels) the fundamental component f_0. The repetition interval was 120 ms, but only the initial 80 ms of the PSTHs is shown for clarity. The stimulus level used to collect the spike trains was 70 dB SPL. The bin width of the histograms is 0.5 ms. For ISIHs a window of 4.5–60 ms was used to exclude the initial peak in the PSTHs from analysis and a vertical dotted line is drawn at 5 ms (corresponding to the fundamental frequency of 200 Hz). This convention applies to all subsequent figures.

pattern is typical of responses to low-frequency stimuli. The model neuron mainly responded at stimulus onset when high-frequency stimuli were used. At 3 kHz the model produced an O_I pattern at 70 dB SPL (35 dB above rate threshold), with a steady-state rate of 4.5 spikes/s. After the parameter set was established, simulations were performed with spike trains collected using complex stimuli.

3. Results

3.1 Harmonic Complexes With and Without the Fundamental

In Figure 4 are illustrated the responses produced by the model when inputs were spike trains collected using the three-component harmonic complexes. The responses are presented as post-stimulus time histograms (PSTHs) and interspike interval histograms (ISIHs). At low sound pressure levels, the model typically responds with a single peak in the PSTH (at a threshold of 30-35 dB SPL). At higher sound pressure levels, the model neuron also produced spikes during the steady-state portion of the stimulus (Figure 4A and C). We only present results obtained at 70 dB SPL in this and all subsequent figures, since it is unlikely that temporal responses containing an onset spike alone convey pitch information.

The responses of the model to the three-component harmonic stimuli, with or without the fundamental component, are very similar: strong responses (high, driven discharge rates) were evoked by both stimuli during the steady-state portion of the stimulus (Figure 4A and C), and ISIHs exhibited a prominent peak (Figure 4B and D). Average interspike intervals in both cases were about 5 ms, which corresponds to the fundamental frequency (200 Hz) of

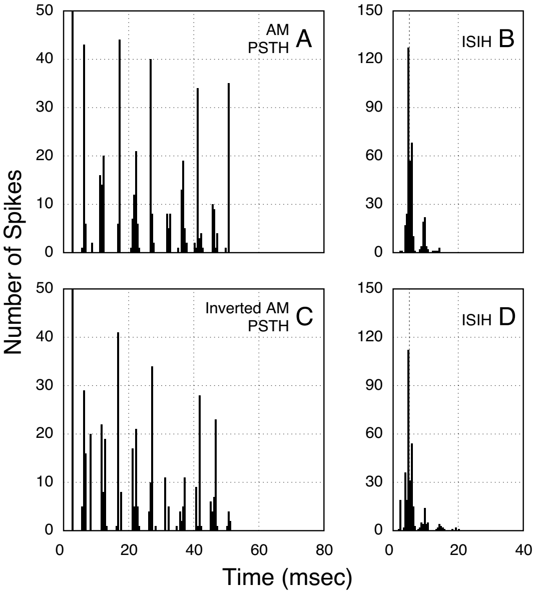


Figure 5 Responses of the model to an amplitude-modulated stimulus (upper panel) and its inverted version (lower panel). The stimulus level was 70 dB SPL.

the stimuli. Compared to the inputs (i.e., the AN spike trains; cf. Figure 3), responses produced by the model showed greatly increased synchrony to the fundamental component.

When the fundamental component was not present in the stimulus, response peaks during the steady-state portion (Figure 4C) were relatively higher than those observed when the fundamental component was present (Figure 4A), although the overall response (i.e., number of spikes) was about the same. This is reflected in the form of a sharper peak at ca. 5 ms in the ISIH (Figure 4D).

3.2 AM and Inverted AM Stimuli

With AM stimuli, the model responses were synchronized to 200 Hz, the modulation frequency of the signals (Figure 5A). The inverted AM stimulus has exactly the same frequency components, but its temporal waveform is different: it has two major peaks instead of one in the positive direction. However, the model neuron seems to more or less ignore the difference: the peak in the ISIH becomes only slightly wider for the inverted AM stimulus (Figure 5D), and there is little difference in the PSTHs.

3.3 Frequency-Shifted Three-Component Complex

When the frequency-shifted, three-component complex was used as the stimulus, the model produced a less synchronized steady-state response, as shown in the PSTH and in the broader ISIH peak (Figure 6) as compared to responses to the harmonic three-component stimuli (cf. Figure 4C and D). Such responses are expected, since the stimulus is “not harmonic” with regard to 200 Hz. Psychophysical studies have shown that this stimulus has a slightly higher pitch than the harmonic three-component complex. Computation of the average interspike intervals in the first peak in the ISIHs (Figs. 4D and 6B) indeed yields a smaller interval (corresponding to 211 Hz) for the frequency-shifted version.

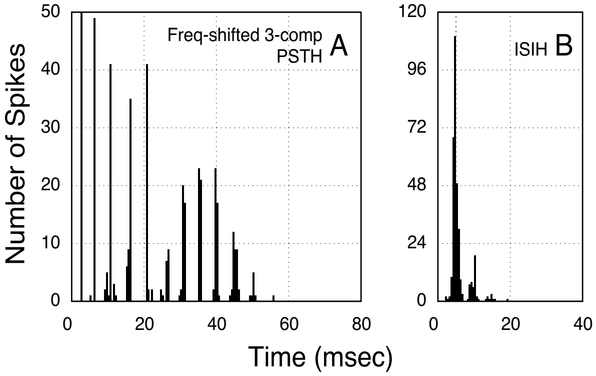


Figure 6 Responses of the model to a frequency-shifted, three-component complex at 70 dB SPL.

3.4 Six-component Complexes

The responses of the model to six-component complexes, with and without the fundamental component, are very similar to those for the three-component complexes (i.e., the one without the fundamental produced sharper peaks in the PSTH and a narrower peak in the ISIH). In Figure 7, we compare responses of the model to two, six-component stimuli, one in which all the components have a starting phase of zero (Figure 7A and B), and one in which component phases are randomized (Figure 7C and D). As can be seen, the model produced fewer discharges and poor synchronization during the steady-state portion when the random-phase stimulus was used, although psychophysical studies suggest that the two stimuli are almost identical perceptually.

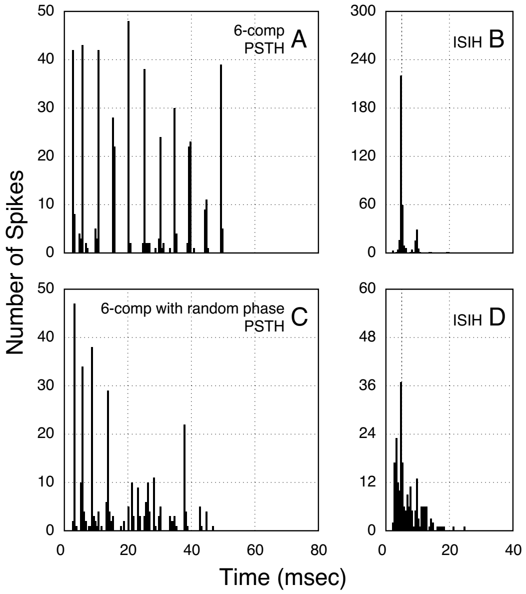


Figure 7 Responses of the model to six-component complexes with (lower panel) and without (upper panel) random phase. The stimulus level was 70 dB SPL.

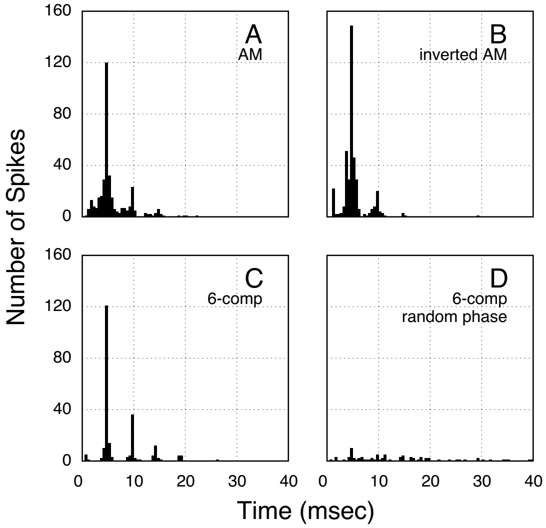


Figure 8 Responses of a gerbil cochlear nucleus neuron (CF = 750 Hz, SR = 0.5 spikes/s, threshold = 26 dB SPL). ISIHs are shown. A and B. AM and inverted AM stimuli, respectively (compare to Figure 5). C and D. six-component complexes with and without random phase (compare to Figure 7). The stimuli were presented at 73 dB SPL.

3.5 Responses of Cochlear Nucleus Onset Neurons

Responses of cochlear nucleus neurons from adult gerbils were compared with simulation results, using an identical set of complex stimuli as were used in the computer simulation. An example of a response from a cochlear nucleus neuron is shown in Figure 8. Spike trains of this low-CF (750 Hz) neuron entrained to tones below 500 Hz. For tone bursts at or above CF the neuron produces an O_I pattern (the steady-state rate was 0.9 spikes/s). The responses of this cell were similar to model outputs (Figs. 5 and 7); both exhibit similar ISIHs in response to the AM and inverted stimuli (Figure 8A and B, cf. Figure 5B and D) and different response patterns when the six-component complexes (with either fixed starting phase or with randomized starting phase) were used as stimuli (Figure 8C and D, cf. Figure 7B and D). When the six-component complex with randomized phases was used, the neuron generated a poor steady-state response (not shown) and exhibited poor synchrony to the fundamental component (200 Hz).

Although more data are required to validate the utility of the model, we are encouraged by the similarities between the model's output and the actual neuronal responses shown here.

4. Discussion

Overall, the model neuron studied here produces sharply defined PSTH peaks that correspond with peaks in the stimuli. Additionally, synchrony to the fundamental component is greatly enhanced in model responses when compared with responses of AN fibers. Using interspike interval estimates, responses of the model to the frequency-shifted three-component complex (Figure 6) correspond to a pitch of 211 Hz, a value very close to that obtained from psychophysical studies. Estimates for other stimuli yield values ranging from 202.5 to 208.9 Hz. Thus, the reciprocal of the average interspike interval in the model responses roughly corresponds to the pitch of the stimuli. These results are consistent with results from experimental studies of CN neurons using complex stimuli [12][13], and suggest that octopus cells are able to process pitch information in complex stimuli.

4.1 Responses to AM Stimuli

Interestingly, the model produced similar responses to the AM and inverted AM stimuli (Figure 5). Such model behavior corresponds well with responses observed in onset cells *in vivo* (Figure 8). The results are also consistent with the psychophysical observation that the two stimuli are perceptually similar. According to Brugge et al. [1], auditory-nerve fibers respond to the envelope of multiple component stimuli. Consistent with that observation, AN fiber responses (inputs to the model neuron) to AM stimuli appear to "follow" the envelope waveform. Envelope following was not observed in responses of the model neuron, suggesting that octopus cells extract critical information about the fundamental periodicity of multiple component signals despite a difference in the temporal waveforms (Figure 2C and D) (i.e., octopus cells do not simply function as a "peak picker" as a simple stimulus "fine, structure" theory suggests [18]).

4.2 Responses to Harmonic Complexes With and Without Random Phase

The model exhibited less robust responses to six-component complex stimuli with randomized starting phases during the steady-state than it did to stimuli in which all components had a starting phase of zero (Figure 7, cf. Figure 2 for waveforms). This was true regardless of the fact that in psychophysical studies these stimuli produce the same pitch. The neuronal responses recorded from the gerbil are similar to the results produced by the model (Figure 8). Although similar results have been reported by Evans and Zhao [4] for high-frequency onset neurons, our results demonstrate that this response property holds for lower frequency neurons which, presumably, play a more important role in the perception of pitch in complex stimuli [15].

Since AN fiber responses are known to "follow the stimulus waveform" faithfully in the frequency range of interest, these differences in responses can be explained by

(1) differences in inputs to the model,
(2) the relatively broad tuning, and
(3) the onset response characteristics of octopus cells.

Broad tuning ensures that octopus cells receive inputs from all harmonic components of the complex stimuli studied here. Consequently, the overall input to the octopus cell would be less synchronized for the random-phase condition than for the zero-starting-phase condition. The onset response character of octopus cells ensures that the cell will respond better to stimuli that are more synchronized (zero starting phases) than to those that are less well synchronized (random-phase version).

4.3 Cellular Basis for Octopus Cell's Processing of Pitch Information

The ability of octopus cells to process pitch information from complex stimuli appears to be derived from its basic onset response characteristics. At low frequencies onset responding neurons entrain to the stimulus. Such onset characteristics are mainly due to a low membrane impedance [3][6][8] and a low threshold K^+ channel of the octopus cell [3]. It would be interesting to study how changes in channel kinetics affect response characteristics of the model to complex stimuli, especially the seemingly contradictory behavior observed when the random-phase and AM stimuli are considered (Figs. 5 and 7). For example, a direct relationship between the kinetics of K_{LT} channels and spike frequency during the steady-state has been observed [3]. If the kinetic characteristics of the K_{LT} channel are faster than in our model, would the response pattern to the inverted AM stimulus change (Figure 2D), and would average interspike intervals decrease? Our preliminary results suggest that the model's response to complex stimuli is less sensitive to changes of model parameters than are its responses to tone bursts.

4.4 Limitations of the Octopus Cell's Pitch Processing Ability

Although octopus cells are able to process pitch information contained in complex stimuli, the results presented in this chapter suggest that such an ability might be limited. First, only moderate and high-SPL stimuli produce substantial steady-state responses in octopus cells, at least in the current model. At low levels the model only responds at stimulus onset, and it is thought that pure onset spikes convey little information other than the onset of the stimulus. Second, as demonstrated by both model output and experimental data (Figs. 7 and 8), octopus cells respond poorly to harmonic complex stimuli with randomized starting phases for each of their components. Third, octopus cells primarily receive inputs from AN fibers with high-SRs. However, AN fibers with lower SRs also carry pitch-related information. These limitations suggest important roles for other neurons within the cochlear nucleus as well as those located at later stages of the central auditory pathway in processing pitch information.

Based on the simulations in this study, neurons with onset-like response characteristics have an advantage over other types of neurons in processing pitch-related information. This is because onset neurons respond well to synchronized inputs. Low-threshold K^+ channels, which are the main cellular component contributing to the onset responses of octopus cells, have been found in both the CN and other nuclei of the central auditory pathway. Those found in higher centers also respond with onset-like discharge behavior and may contribute to pitch perception.

Responses to harmonic complex stimuli with randomized starting phases were as poorly synchronized in the model as in the onset neuronal model. In all likelihood, this reflects the broad tuning of octopus cells and is not characteristic of neurons with relatively narrow tuning characteristics. Fewer inputs converge onto sharply tuned neurons and as a result, inputs are more synchronized, increasing the probability that spikes will occur. Almost all neurons in the cochlear nucleus are more sharply tuned than the octopus cells (the exception being the multipolar stellate cells, a.k.a. onset choppers) and are thus more likely to respond in a more synchronized fashion than octopus cells to harmonic signals with random phase. Generally, these neurons have lower response thresholds than octopus cells and presumably play an important role in the processing of pitch information at low intensities.

Among the different response types associated with octopus cells and large multipolar stellate cells (both have relatively broad tuning characteristics), O_L and O_C types are probably more important in pitch perception than the O_I type, due to their higher discharge rates during the steady-state portion of the response. Psychophysical studies have shown that the lower frequency region plays a dominant role in the perception of pitch [15]. Although all O_I, O_L and O_C units show entrainment at low frequencies (some up to 2 kHz [5][14]), the O_{LF} (low-frequency onset) units should have an advantage due to lower thresholds in this frequency region.

5. Summary and Conclusions

An octopus-cell model neuron was used to study how neurons with broad tuning and transient temporal response properties extract pitch information from complex stimuli. Harmonic and inharmonic complex stimuli, similar to those used in psychophysical studies, were used to collect auditory-nerve data from anesthetized cats, and the resulting spike trains served as input to the neuronal model. The model produced sharply defined peaks in PSTHs at every cycle of the fundamental component in response to three- or six-component harmonic complex stimuli, regardless of the presence or absence of the fundamental component. In response to a frequency-shifted three-component complex, average interspike intervals decreased slightly, corresponding to an upward pitch shift. The model produced very similar responses to an AM stimulus and its inverted version, in conformity with psychophysical data and responses from *in vivo*. In all cases, synchrony to the fundamental frequency of the stimulus was enhanced in the model neuron when compared to the responses of auditory-nerve fibers. These results, consistent with experimental studies of cochlear nucleus neurons, demonstrate that octopus cells are capable of processing pitch information in stimuli through the action of converging inputs from auditory-nerve fibers that originate over a wide frequency range. This result supports the hypothesis that interspike interval information is a correlate of pitch. However, the poor responses of the model and CN neurons to six-component, random-phase stimuli do not agree with the psychophysical finding that phase is unimportant in pitch perception. This suggests that other neurons in the auditory system play an important role in pitch perception as well.

Acknowledgments

This work was supported by NIDCD DC01007 and DC00982. We wish to thank Lei Song and Yiming Zhang for their help during the collection of auditory nerve and cochlear nucleus data.

References

[1] Brugge, J. F., Anderson, D. J., Hind, J. E. and Rose, J. E. “Time structure of discharges in single auditory fibers of the squirrel monkey in response to complex periodic sounds.” *J. Neurophysiol.* 32: 386–401, 1969.

[2] Cai, Y., Walsh, E. J. and McGee, J. “A simple program for simulating responses of neurons with arbitrarily structured dendritic trees.” *J. Neurosci. Methods* 74: 27–35, 1997.

[3] Cai, Y., Walsh, E. J. and McGee, J. “Mechanisms of onset responses in octopus cells: Implications of a model.” *J. Neurophysiol.* 78: 872–833, 1997.

[4] Evans, E. F. and Zhao, W. “Periodicity coding of the fundamental frequency of harmonic complexes: Physiological and pharmacological study of onset units in the ventral cochlear nucleus.” In *Psychophysical and Physiological Advances in Hearing,* A. R. Palmer, A. Rees, A. Q. Summerfield and R. Meddis (eds.), London: Whurr Publishers Ltd., 1998, pp. 186–194.

[5] Godfrey, D. A., Kiang, N. Y. S., and Norris, B. E. “Single unit activity in the posteroventral cochlear nucleus of the cat.” *J. Comp. Neurol.* 162: 247–268, 1975.

[6] Golding, N. L., Robertson, D. and Oertel, D. “Recordings from slices indicate that octopus cells of the cochlear nucleus detect coincident firing of auditory nerve fibers with temporal precision.” *J. Neurosci.* 15: 3138–3153, 1995.

[7] Goldstein, J. L. “An optimum processor theory for the central formation of the pitch of complex tones.” *J. Acoust. Soc. Am.* 54: 1496–1515, 1973.

[8] Levy, K. L. and Kipke, D. R. “A computational model of the cochlear nucleus octopus cell.” *J. Acoust. Soc. Am.* 102: 391–402, 1997.

[9] Meddis, R. and Hewitt, M. J. “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Phase identification.” *J. Acoust. Soc. Am.* 89: 2866–2882, 1991.

[10] Oertel, D. “The role of intrinsic neuronal properties in the encoding of auditory information in the cochlear nuclei.” *Current Opinion Neurobiol.* 1: 221–228, 1991.

[11] Patterson, R. D. “The effects of relative phase and the number of components on residue pitch.” *J. Acoust. Soc. Am.* 65: 1565–1572, 1973.

[12] Rhode, W. S. “Interspike intervals as a correlate of periodicity pitch in cat cochlear nucleus.” *J. Acoust. Soc. Am.* 97: 2414–2429, 1995.

[13] Rhode, W. S. and Greenberg, S. “Encoding of amplitude modulation in the cochlear nucleus of the cat.” *J. Neurophysiol.* 71: 1797–1825, 1994.

[14] Rhode, W. S. and Smith, P. H. “Encoding timing and intensity in the ventral cochlear nucleus of the cat.” *J. Neurophysiol.* 56: 261–286, 1986.

[15] Ritsma, R. J. “Existence region of the tonal residue. I.” *J. Acoust. Soc. Am.* 34: 1224–1229, 1962.

[16] Schouten, J. F., Ritsma, R. J. and Cardozo, B. L. “Pitch of the residue.” *J. Acoust. Soc. Am.* 34: 1418–1424, 1962.

[17] Srulovicz, P. and Goldstein, J. L. “A central spectrum model: A synthesis of auditory-timing and place cues in monaural communication of frequency spectrum.” *J. Acoust. Soc. Am.* 73: 1266–1276, 1983

[18] Wightman, F. L. “Pitch and stimulus fine structure.” *J. Acoust. Soc. Am.* 54: 397–406, 1973.

[19] Wightman, F. L. “The pattern-transformation model of pitch.” *J. Acoust. Soc. Am.* 54: 407–416, 1973.

[20] Wightman, F. L. and Green, D. M. “The perception of pitch.” *Am. Scientist* 62: 208–215, 1974.

TEMPORAL PROCESSING AND PERIODICITY ANALYSIS

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

TEMPORAL PROCESSING AND PERIODICITY ANALYSIS

Roy D. Patterson

Centre for the Neural Basis of Hearing,
Physiology Department
Cambridge University
Cambridge, CB2 3EG, UK

This section contains four papers concerned with modeling auditory temporal processing as it is observed in the responses of single cells to sinusoids and in the perception of complex sounds by humans.

The paper by Peter Heil is concerned with "Aspects of envelope coding in the auditory system." The work is motivated in the introduction by the importance of envelope onsets to the perception of the timbre of musical instruments; the experimental stimuli are, nevertheless, bursts of sinusoids. Single-unit responses are recorded from the primary auditory cortex in cats and they are very interesting inasmuch as these units typically fire only once in response to a tone burst; this is reliably at the onset of the stimulus, no matter what the best frequency of the unit or its bias towards laterality. The paper is interesting because, despite the very low firing rate of the cells, the paper summarizes over 32,000 firings from over 1800 units. Moreover, the firing of the units is extremely precise which enables Heil, with exemplary data analyses, to show that the units respond, not to a fixed threshold level, but rather to the maximum acceleration of peak pressure in the onset. The second half of the paper describes an intriguing envelope-tracking mechanism that could be assembled from a population of these onset units with a range of acceleration thresholds. The system has the potential to represent the shape of the onset of the envelope of a sound in a way that is largely independent of SPL, and to represent it with temporal accuracy that far exceeds that suggested by the phase-locking of the individual units. What is not specified, however, is how the system would respond to the falling portion of the envelope — either the long slow decay of the gross envelope at the end of a piano note or a vowel, or the more rapid decay within the period of the note or vowel which contains timbre information about the width of resonances in the source. The units in this study are essentially onset detectors whose response to continuous tones is actually inhibited for a lengthy period after the initial response, and so they seem likely to coast over the falling portion of the envelope.

The paper by S. Bleeck and G. Langner is concerned with the "Functional significance of latencies in the auditory system," and it is based on the latency data reported by Heil in response to sinusoids with linear and cosine-squared onset ramps. The paper is particularly intriguing as a companion to the Heil paper because of the contrasting approaches that the authors take to modeling their data. In both papers, the authors note that the quasi-hyperbolic form of the data suggests that the neuron is simply firing at a fixed threshold, and they fit fixed-threshold lines to the data. They both report that the fit is reasonably good and note that the data diverge systematically from the fixed threshold fit in several places. Heil then concludes that the divergence between fit and data is sufficient to reject the fixed threshold model entirely, and he goes on to develop his hypothesis that the system is concerned to detect the maximum acceleration of peak pressure. This leads to a power function fit to the data, which when modified to start from an empirical minimum, provides an excellent fit to the data. In contrast, Bleeck and Langner take the goodness of the fit as sufficient evidence of a fixed-threshold process to develop a modified simple-threshold model in the form of a "leaky integrate and fire neuron." In the course of the fitting process, they modify the neuron and in this way they also produce a better fit to the data than a simple threshold model. The fit does not appear to be quite as good as Heil's (neither provide measures of the rms error) but it has the obvious advantage of providing a physiological description of neural responses.

What these papers do not explain, however, is why these neurons only fire once to sinusoids as long as 170 ms in duration. Heil's power function and Bleeck and Langner's leaky integrate neuron have no temporal parameters that extend to any significant duration and there is no lockout-and-reset mechanism in either case. This raises the interesting question as to how long the sinusoid would have to be to produce a second firing and what would the firing rate be if the sinusoid were left on indefinitely? Is the lockout mechanism in the circuit before the cortical neuron, or in the neuron itself, or in a feedback circuit beyond the neuron?

The paper by M. Unoki and M. Akagi describes "A computational model of Co-modulation Masking Release (CMR)" — the fact that a sinusoid presented in a band of modulated noise is more detectable when the flanking noise bands beside the signal band are co-modulated. The computational model is a combination of a Power-Spectrum (PS) model of masking like those proposed by Patterson and Moore [4], and a model of stream segregation based on Auditory Scene Analysis (ASA) [1]. The models are combined by the simple expedient of running them in parallel and then selecting the result of the model that gives the lower estimate of signal threshold. The power-spectrum model gives the lower values when the noise bandwidth is less than the width of the auditory filter centered on the tone (130 Hz at 1.0 kHz). The segregation model gives the lower values when the noise bandwidth is greater than the width of the auditory filter. The function produced by the two together is quite similar to the original results presented by Hall et al.[2]. There is one aspect of the architecture of the computational model that may be puzzling to auditory readers and which is perhaps worth explaining. This is the fact that the component models are run in parallel, both operating separately on the input. This is non-auditory. Moreover, whereas the spectral analysis in the ASA model is performed by a state-of-the-art wavelet transform with a gammatone kernel, the spectral analysis in the PS model is performed by a single auditory filter centered at the signal frequency (1 kHz). The first thing to note is that in this version of the PS model, the calculation is performed in the time domain with a gammatone auditory filter. So the calculations in the PS model are based on the same filtered wave as that flowing from the wavelet filter centered at 1 kHz in the ASA model. This raises the question as to why they did not simply use the wavelet filter as the basis for the PS model calculations. In this case, both models would use the same initial spectral analysis and the model would be much more plausible auditorily; the parallel processing would be in the "central processor." The answer to this question is, of course, simple; the model is intended to prove the concept of a CMR model with parallel processing and for this purpose it does not matter where in the system the parallelism arises. It is important, however, to note these differences in modeling style in order to understand the purpose of the model.

The paper by Peter Cariani is entitled "Neural timing nets for auditory computation." The topic is motivated in the first three sections of the paper by a) drawing attention to the correspondence between 'all-order, interspike interval histograms' and perceptions of pitch and timbre, b) noting that no delay lines of the sort that might convert temporal phase locking into place information through a process like autocorrelation, have been found to date, and c) noting that reasonably good temporal information persists in one way or another up to the thalamus. There are many contentious statements in these sections which should not distract the reader from the main point of the paper which is the intriguing neural timing nets presented in Sections 4 and 8.

The feed-forward net in Section 4 has an initial stage that is like the coincidence mechanism suggested by Jeffress [3] for the detection of interaural time differences. Note, however, that this net is not a binaural processing; the binaural delay lines found in the medial superior olive are only a millisecond long and so there are never multiple pulses in either of the input lines The full feed-forward net is essentially a time-interval sorter; it is interesting because it is simple yet it operates like an autocorrelator, and so it can extract the periodicities associated with the pitch and timbre of vowels (Sections 5 and 6). Do not pay too much attention to the details of the net in Section 4 and autocorrelation functions in Sections 5 and 6; the latter were not calculated with the former. The net would operate on unipolar neural pulses and produce entirely positive autocorrelations; the population autocorrelations were calculated from the wave and so contain negative as well as positive values. Note also that the coincidence-array input lines for pitch would have to be on the order of 33 ms long if the model were to account for the lower limit of temporal pitch (30 Hz).

The recurrent-timing nets in Section 8 are a form of reverberation network in which there is an array of neural loops of different lengths — one for each periodicity. Although it is not specified, the delay line is usually assumed to be a cascade of neurons, in which case it is easy to imagine extending the length of the chain to the long delays required to explain the lower limit of pitch. It would be interesting to see how the number of cells in the chain and the cumulative timing variability along such a chain would relate to the resolution of pitch perception and the variability of pitch matching. In summary, the networks are interesting as they show how simple neural arrays could perform many of the time-interval calculations required for pitch extraction. As the author says at the end of the discussion, however, "This present treatment of timing networks barely ventures beyond an outline of the idea and what kinds of operations might potentially be carried out."

References

[1] Bregman, A. S. "Auditory Scene Analysis: hearing in complex environments." In *Thinking in Sound: The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand (eds.), New York: Oxford University Press, pp. 10–36, 1993.

[2] Hall, J. W. and Fernandes, M. A. "The role of monaural frequency selectivity in binaural analysis." *J. Acoust. Soc. Am.* 76: 435–439, 1984.

[3] Jeffress, L. A. "A place theory of sound localization." *J. Comp. Physiol. Psychol.,* 41: 35–39, 1948.

[4] Patterson, R. D. and Moore, B. C. J. "Auditory filters and excitation patterns as representations of frequency resolution." In *Frequency Selectivity in Hearing,* B. C. J. Moore (ed.), London: Academic Press, pp. 123-178, 1986.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

ASPECTS OF ENVELOPE CODING IN THE AUDITORY SYSTEM

Peter Heil

Leibniz Institute for Neurobiology
39118 Magdeburg, Germany

1. Introduction

Sound consists of very rapid pressure fluctuations (the “fine structure” or “carrier”) with superimposed overall changes of the amplitude of the carrier on a slower time scale (the “envelope”). The envelope, or boundary of the stimulus waveform, is a construct and not part of a signal’s frequency spectrum. Yet envelope characteristics have perceptual correlates, such as timbre. For example, a signal whose spectrum is composed of frequencies which are integer multiples of a fundamental frequency has a temporal envelope that fluctuates periodically, and a human listener perceives a harmonic sound whose pitch corresponds to the fundamental frequency (“periodicity pitch” or a percept of the “missing fundamental”) [31]. The number of spectral components and their relative amplitudes and phases affect the shape of the periodic envelope fluctuations and as well as the sound´s timbre [18][29][30], not its pitch which is largely determined by the fine structure [1][31][37].

In complex signals such as speech, animal communication sounds or music, the spectral composition changes as a function of time (i.e., such signals are composed of independently varying envelopes in different spectral bands). The significance of the details of these envelopes for recognition of speech sounds has been convincingly demonstrated. For example, when speech signals are split up into a number of frequency bands and the envelope of each band is then low-pass filtered (“smeared”), severe reductions in sentence intelligibility and in phoneme identification can result, depending on the number of bands and cut-off frequencies [2]. Conversely, when temporal envelopes from broad frequency bands are extracted and used to modulate noises of the same bandwidths, nearly perfect speech recognition can result when temporal envelope cues are preserved in only a small number of contiguous spectral bands [35]. Thus, the specifics of the changes in temporal envelopes in narrow frequency bands and the relationships of temporal envelopes across different frequency bands are of prime importance for the information conveyed by complex acoustic signals.

In music the temporal envelope contributes significantly to the identification of instruments. When asymmetrically shaped temporal envelopes of sounds produced by instruments [21] or of sinusoids [22][23] are played backwards, they sound quite different making it difficult for listeners to recognize the instrument, although the longer-term spectra of original and time-reversed versions are identical. Thus, the details of the onset seem crucial. For example, it is possible to transform the perception of a trumpet into that of a violin, and vice versa, by manipulating the properties of the signal during the initial 50 ms or so [7]. Generally, rapid changes in a signal’s envelope, particularly at onset (“attack”), as well as changes in harmonic structure during onset are primary sources of acoustic information that contribute to the timbre of sounds [18][29].

The auditory system's representation of the details of temporal envelopes, either at onset or afterwards, is not well understood. This is due to the fact that most studies of envelope coding have focussed on the modulation frequency of periodically amplitude-modulated signals [19]. The potentially confounding effects of concomitant changes in the shape of temporal envelopes have largely been neglected. Only very few studies have compared the neuronal responses to signals amplitude-modulated by different envelopes (i.e., in essence have examined effects of timbre [3][6][20][33]). There are only few studies which have addressed the effects of manipulation of stimulus onset on neuronal responses [6][8][17][24][28][36].

This chapter summarizes and elaborates on some recent ideas of how the details of the onset (i.e., the time course of the initial segment of the envelope) may be encoded by neuronal populations [9][10][12][13][14]. The basic concept is derived from a detailed analysis of the responses of neurons in the cat's auditory cortex to tone bursts and differs from previous approaches in two important respects. First, the analysis of response properties focusses on stimulus properties at stimulus onset. This focus is essential, because nearly all neurons of the auditory system respond vigorously to a tone's onset, but paradoxically onset response properties have previously been analyzed with respect to parameters characteristic of the steady-state of the stimulus, such as the sound pressure level (SPL). Consequently, the changes in neuronal onset responses observed with changes in stimulus SPL have been interpreted as significant for intensity coding [16][25][32][34]. However, dynamic properties such as the rate of change and the acceleration of amplitude (peak pressure) have been inadvertently co-varied with SPL when the rise time is held constant. As this is routinely the case, standard experimental conditions are ambiguous with respect to the relevant stimulus parameter(s). To address this issue I have used tones of different steady-state SPL, rise time and rise function (linear and cosine-squared). The second novel aspect of the concept presented here is that the proposed high-fidelity representation of the initial envelope incorporates the response latency, the precision of response timing and the response magnitude (all of which are stimulus-dependent) of a population of neurons. I will argue that these response properties should be considered jointly.

2. Sound Onset Parameters Shaping Onset Responses

2.1 Response Latency and Precision of Response Timing

2.1.1 Stimulus Parameters Determining First-Spike Latency

Figure 1 shows, for a single neuron from the primary auditory cortex (AI) of a barbiturate-anesthetized cat, the dependence of first-spike latency on various parameters associated with sinusoidal signals presented at the cell's characteristic frequency (i.e., "CF" tones). Figures 1A and B plot the mean first-spike latency (computed from responses to 20 repetitions of each stimulus) as a function of SPL for different rise times of tones shaped with either linear (A) or cosine-squared rise functions (B). For each rise function and rise time, latency decreases with increasing SPL. However, for any given SPL latency also decreases with decreasing rise time. For linear-rise-function tones, the rate of change of peak pressure (RCPP) co-varies with SPL (upper inset in C) and with rise time. When the latency data of Figure 1A are plotted against this derived parameter, they form an invariant function of

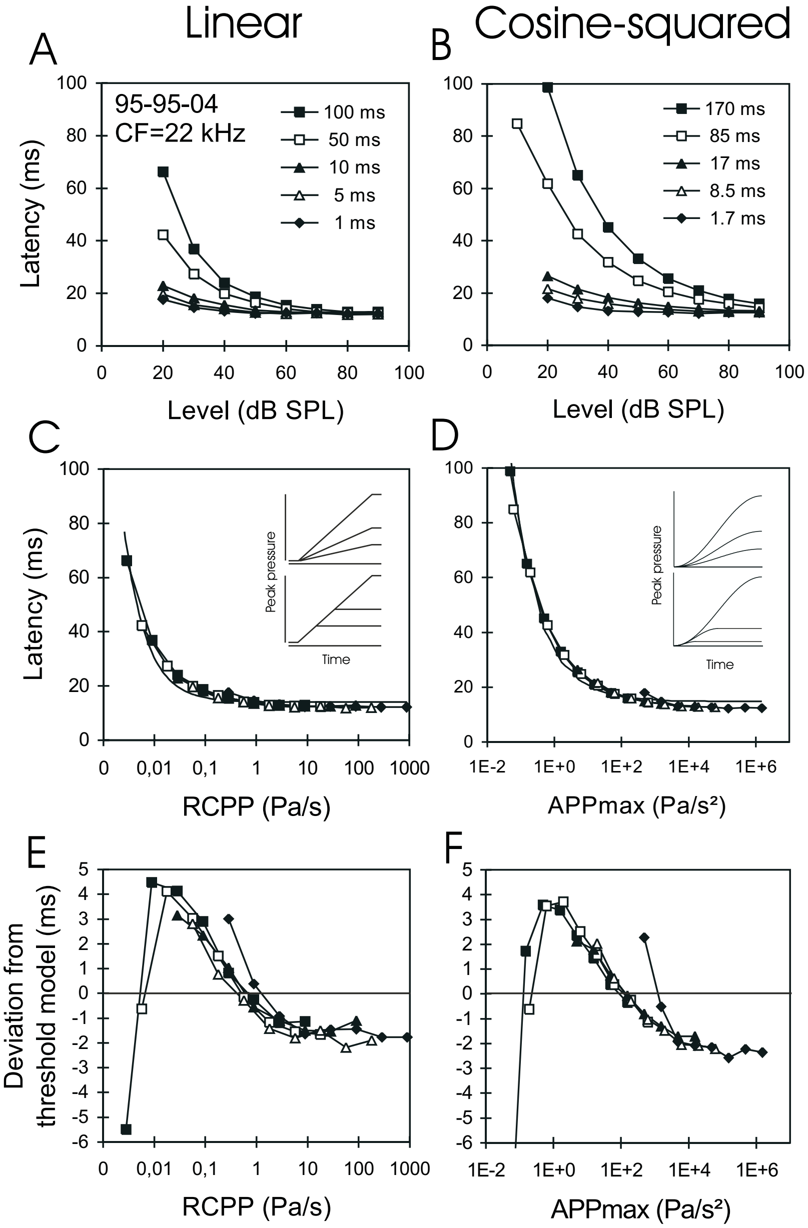


Figure 1 Mean first-spike latency of a single neuron to CF tones shaped with linear (left column) and with cosine-squared rise functions (right column). Tones differ in SPL and rise time (legends in A and B). In A and B, latency is plotted as a function of the steady-state level, and in C and D as a function of the rate of change of peak pressure (RCPP; C) and against the maximum acceleration of peak pressure ($\mathrm{APP}_{\mathrm{max}}$; D) occurring at tone onset. These parameters covary with SPL and rise time. Note the close alignment of the latency functions for different rise times. Upper insets in C and D show the envelopes of signals of different SPL and identical rise time. Lower insets show the envelopes of signals with identical RCPP (C) or $\mathrm{APP}_{\mathrm{max}}$ (D). The continuous lines, without symbols in C and D, represent the best fit of the fixed threshold model to the data. E and F show the deviations of the data from this model. Note the systematic nature of these deviations.

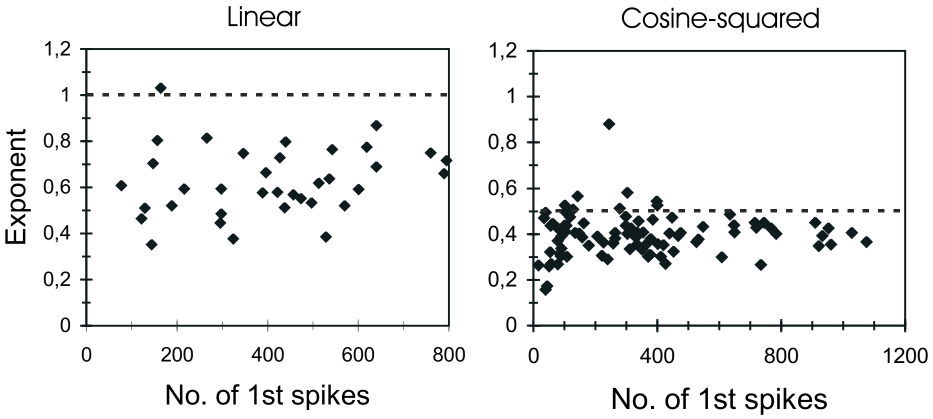


Figure 2 Scatter plots of the exponent c, obtained from the fits of equations (1a) and (1b) to the latency versus RCPP and versus APP_{max} functions with linear and cosine-squared rise functions, respectively against the number of initial spikes having contributed to each fit. Dashed horizontal lines mark the exponents which are expected if a fixed threshold model would explain the latency data.

RCPP (Figure 1C). For cosine-squared rise function tones the latency data of Figure 1B form an invariant function of the maximum acceleration of peak pressure (APP_{max}) at stimulus onset, a feature derived from the second derivative of the envelope, which co-varies with SPL (upper inset in D) and rise time (Figure 1D) [9][11][12][13]. These findings reveal that the latency of the first spike is determined by dynamic properties of the stimuli at their very onset.

2.1.2 Rejection of the Threshold Model

Linear and cosine-squared rise function tones, sharing a common RCPP and APP_{max}, respectively, differ in rise time and in SPL. Therefore, response latency is independent of rise time or SPL per se. However, the initial time course of the envelopes of such tones are in close register (lower insets in C and D). It may therefore be argued that the response of a given neuron is triggered whenever the signal reaches a fixed threshold amplitude during the rise time. Changes in latency co-occurring with changes in stimulus parameters would then reflect the fact that this amplitude is reached at different times, depending on SPL, rise time, RCPP, or APP_{max}. The lines without symbols in Figure 1C and D, which are somewhat difficult to see, represent the best fit of this fixed-threshold model to the data. At first sight the model appears to provide a reasonably good description of the data. However, a closer look reveals systematic, rather than erratic, deviations of the data from the model. Figures 1E and F show that at low and at high values of RCPP and APP_{max} latencies are consistently shorter, and at intermediate values consistently longer than predicted by the model. This is due to the fact that the curvatures of the latency versus RCPP or APP_{max} functions are shallower than predicted by the model (cf. Figure 1C and D).

The general validity of this mismatch between a fixed-threshold model and the data is illustrated in Figure 2. The latency versus RCPP functions of 36 AI neurons were fitted with the equation:

$$L\text{-}L_{min} = (RCPP/RCPP_{(0)})^{-c} \qquad (1a)$$

and the 93 latency-versus-APP_{max} functions obtained from 65 AI neurons were fitted accordingly —

$$L\text{-}L_{min} = (APP_{max}/APP_{max(0)})^{-c} \qquad (1b)$$

where L_{min} represents a constant "transmission delay" (i.e. the minimum against which the latency L converges asymptotically at high values of RCPP and APP_{max} (cf. Figures 1C, D)). $RCPP_{(0)}$ and $APP_{max(0)}$ identify the rate (in Pa/s) and maximum acceleration of peak pressure (in Pa/s^2), respectively, for which the difference between the latency and L_{min} is 1 ms. Thus, a low value of $RCPP_{(0)}$ or $APP_{max(0)}$ indicates that a neuron´s response to signals of any given RCPP or APP_{max} is triggered early during the onset. In other words, a low value of $RCPP_{(0)}$ or $APP_{max(0)}$ identifies a high sensitivity of a neuron to such transients and a high value indicates a low sensitivity.

If the fixed-threshold model were correct then the exponent, *c*, of equations (1a) and (1b) should be 1 and 0.5 for linear and cosine-squared rise functions, respectively. Figure 2 plots the exponents obtained from the fits against the number of initial spikes having contributed to each fit. It is readily seen that the vast majority of exponents is smaller than predicted by the fixed threshold model (dashed lines). Only those fits based on a small number of initial spikes (and hence being less reliable) yield exponents greater than or equal to those predicted by that model.

In summary, the observed changes in response latency with changes in the rate or maximum acceleration of peak pressure are incompatible with a fixed threshold model of latency. The incompatibility is even more pronounced when membrane accommodation and adaptation are taken into account [11][12]. Phrased differently, these results show that the first spike is triggered at an instantaneous amplitude that varies systematically with a variety of stimulus parameters.

2.1.3 A Common Shape of Latency versus APP_{max} Functions

Figure 2 also reveals that the distribution of the exponents are rather narrow. This reflects the fact that the latency-versus-RCPP and latency-versus-APP_{max} functions for various neurons are of strikingly similar shape. This is illustrated for the latter functions in Figure 3. Figure 3A shows mean latency versus APP_{max} functions of five different neurons, obtained from four different cats with tones of very different frequencies (CFs range from 2.3 to 30 kHz) and with different laterality of presentation. The functions differ in their positions within the coordinate system of latency and APP_{max}. These positional differences reflect differences in L_{min} and $APP_{max(0)}$. The functions also differ in extent, but their shapes are virtually identical. This is more clearly illustrated in Figure 3B. Each latency versus APP_{max} function was again fitted with equation (1b) but this time the exponent *c* was fixed at 0.4. This number represents the distribution's average when each individual exponent is weighted by the number of first spikes having contributed to the fit (cf. Figure 2). Then, the difference between the measured response latency and the neuronal L_{min}, as obtained from this fit, is plotted against the ratio of APP_{max} of the stimulus and the neuronal $APP_{max(0)}$, also obtained from the fit. In this plot, the fitted function traverses the coordinate point (1,1). Next, the data from all 65 neurons tested with cosine-squared rise functions were plotted in this way. Figure 3B shows the result for means of first-spike latencies based on a response probability of ≥ 0.5 (i.e., on ≥10 initial spikes out of 20 trials). The approximately 1800 data

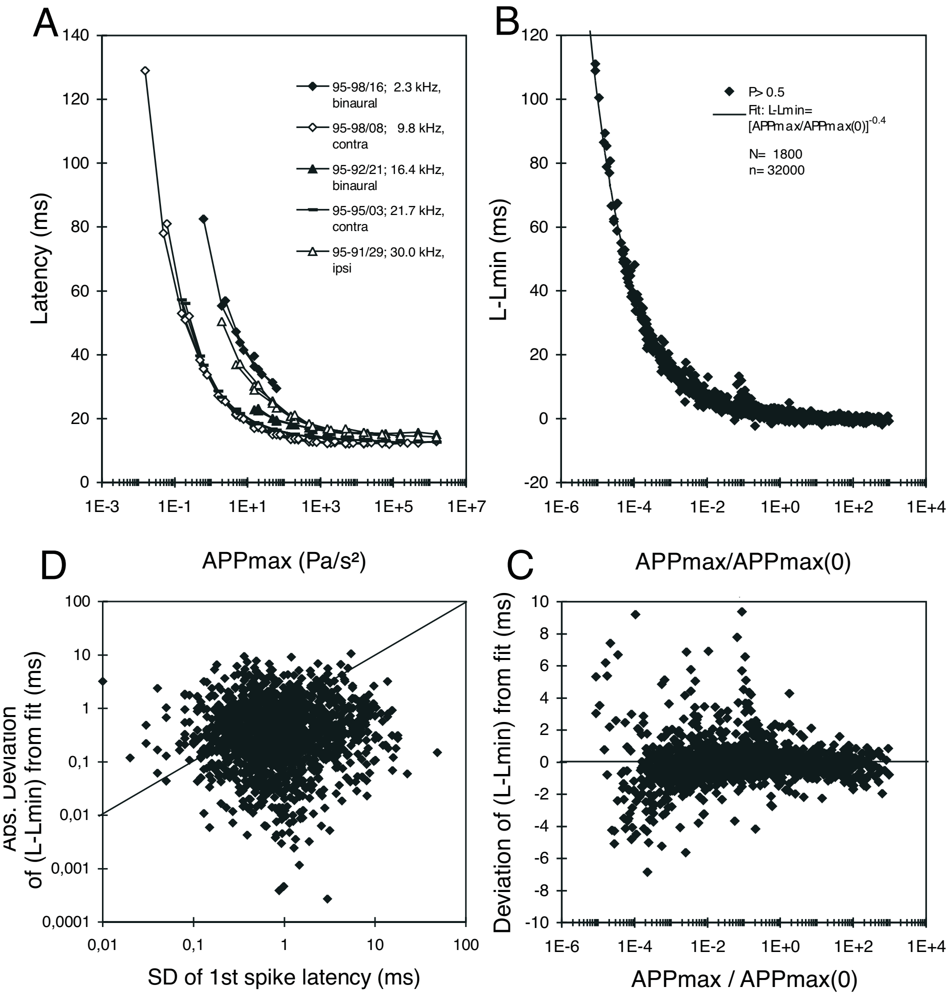


Figure 3 Latency versus APP_{max} functions have identical shape. A. Latency-versus-APP_{max} functions of five auditory-cortex neurons obtained from four different cats with tones of frequencies between 2.3 and 30 kHz, and with different laterality of presentation. Data obtained from a given neuron with tones of different SPL, but of the same cosine-squared rise time, are connected. Note the common shape of the latency functions. Differences in the positions of functions along the ordinate reflect differences in transmission delay (L_{min}) and along the abscissa differences in transient sensitivity $APP_{max(0)}$, with a low value of $APP_{max(0)}$ defining a function in a more leftward position. B. Plots of L-L_{min} against $APP_{max}/APP_{max(0)}$ for some 1800 mean first-spike latencies, each based on a response probability ≥ 0.5. L_{min} and $APP_{max(0)}$ were obtained from the fits of equation (1b) with $c = 0.4$ (the weighted average of the distribution) to the data from each neuron and stimulus condition. Note that points form a narrow band, and a fit of these data with Equation 1b yield an exponent, *c*, again of 0.4 (thin solid line). C. Deviations of the 1800 means from this line of best fit. Note that the points scatter unsystematically around this line. D. Scatter plot of the size of the deviations of the data in B and C from the line of best fit against the standard deviation (SD) of first-spike latency. The continuous line represents the diagonal.

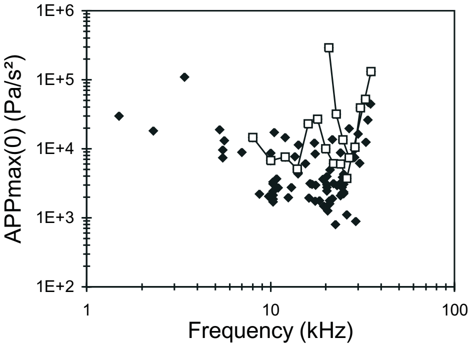


Figure 4 Scatterplot of the transient sensitivity ($APP_{max(0)}$) against tone frequency. Solid symbols identify measures obtained at CF and open symbols connected by lines identify data from two neurons obtained over a range of frequencies.

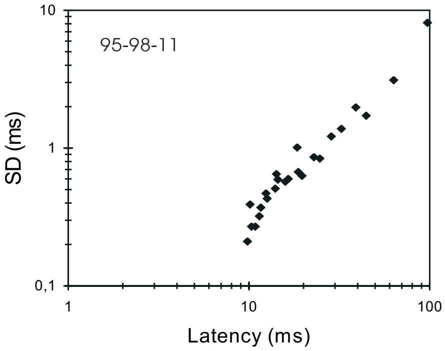


Figure 5 Plot of the standard deviation against the mean first-spike latency for one neuron. Note that the precision of first-spike timing increases with decreasing latency.

points, based on about 32,000 initial spikes, form a very narrow band, and a fit of these data with equation 1b yielded an exponent, c, of again 0.4 (thin solid line in Figure 3B).

Figure 3C shows the deviations of the 1800 means from this line of best fit. The points scatter unsystematically around this line, in stark contrast to the results obtained with the fixed threshold model (cf. Figure 1F). More than 92% of the points fall within ± 2 ms and 81% within ± 1 ms of the fit, while with the threshold model this applies to only about 56% and 25%, respectively. Figure 3D shows that the absolute magnitudes of the deviations of the mean data in Figure 3C from the fit are mostly less than one standard deviation of the mean.

In summary, the latency versus APP_{max} functions of different neurons obtained under different stimulus conditions have essentially the same shape, but differ in extent and position within the coordinate system. The shape of the function is well-described by the power function given by equation 1b and an exponent of 0.4.

2.1.4 Frequency-Dependence of Transmission Delay and Transient Sensitivity

Across different neurons and stimulus conditions L_{min} decreases with increasing CF, but for a given CF, L_{min} differs widely [9]. $APP_{max(0)}$, a measure of the neuron´s transient sensitivity, also varies with CF (Figure 4). The transient sensitivity is highest (i.e., $APP_{max(0)}$ is smallest) around 20 kHz and decreases steeply towards higher frequencies and less steeply towards lower frequencies. For a given neuron, the transient sensitivity varies as a function of stimulus frequency in a similar though not identical manner to threshold tuning curves. Data from two neurons are shown by interconnected open symbols in Figure 4.

2.1.5 Precision of First-spike Timing

The standard deviation of first-spike latency also decreases with increasing RCPP or APP_{max} for linear and cosine-squared rise functions, respectively - roughly proportional to the slope of the corresponding latency functions. Hence, the standard deviation increases with mean latency. Data from a representative neuron are shown in Figure 5. In other words, the precision of spike timing is high when the latency is short and vice versa [9][27]. Over-

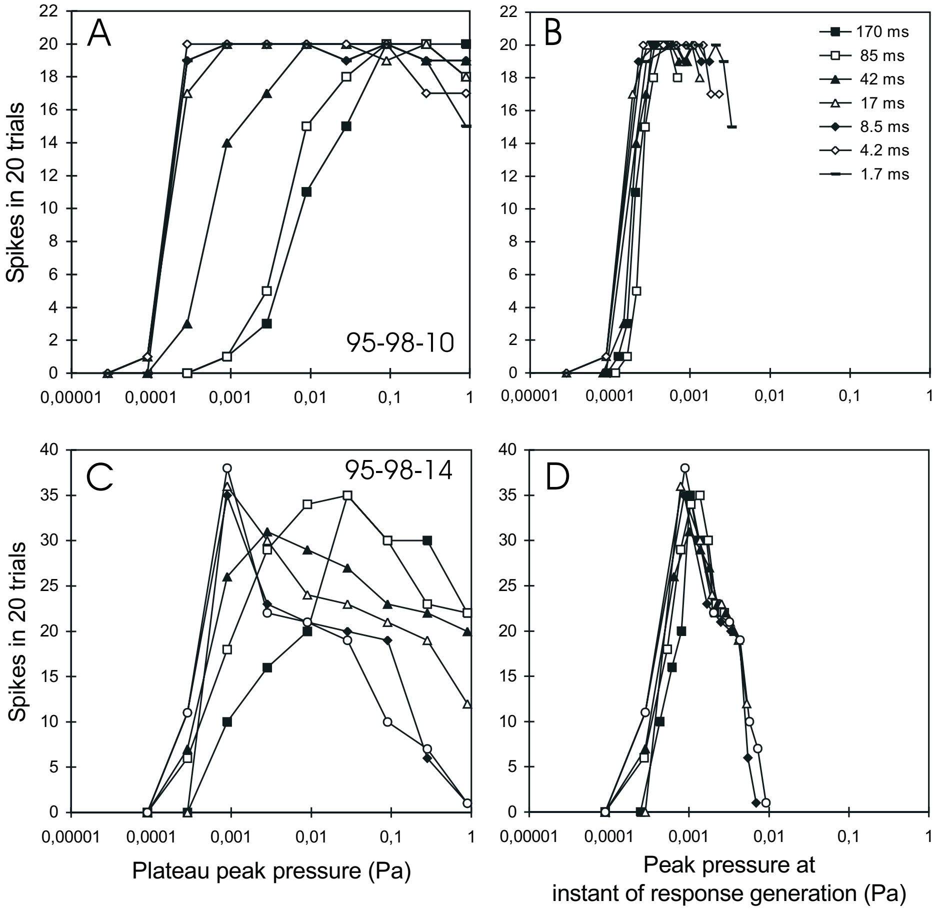


Figure 6 Responses of two neurons to CF tones of different SPL and rise time. In A and C, spike counts are plotted against the steady-state SPL (expressed in Pa; the abscissae correspond to a range from -10 to 90 dB SPL), and in B and D against the instantaneous peak pressure at the time of response generation. Note the close alignment of the response functions for different rise times in B and D.

all, the temporal precision of first-spike timing in AI was found to be as good as [26], or even better than, in the auditory nerve [13], depending on stimulus conditions.

2.2 Response Magnitude

For tones of a given frequency and rise function, the number of spikes discharged by any given neuron varies with SPL. However, for both rise functions, rise time has manifold, and in many cases profound, effects on all properties of such conventional response functions [10][14][24][28][36]. Data, all obtained with cosine-squared rise function tones, from two neurons are shown in Figure 6A and C. In general, threshold SPL, dynamic range and the lowest SPL at which monotonic spike count functions saturate, all increase with lengthening of rise time (e.g. 95-98/10; Figure 6A). In neurons with mostly non-monotonic response functions, “best SPL” increases and the descending high-SPL arm flattens, so that functions obtained with shorter rise times may be highly non-monotonic whereas those obtained with long rise times may be monotonic, and the “tuning” to SPL is less sharp for longer-rise-time tones (e.g. 95-98/14; Figure 6C). Systematic effects of rise time persist when spike counts are plotted as a function of the rate of change or as a function of the maximum acceleration of peak pressure (not shown here, cf. [10]). However, the spike count functions obtained with different rise times, and even with different rise functions, are in close register, when spike counts are plotted as a function of peak pressure at the instant of response generation, rather than as a function of steady-state SPL (Figure 6B,D) [10][14]. The instant of response generation is, of course, given by the difference between the response latency, L, and the transmission delay L_{min}. This suggests that the stimulus-dependent component of the latency, viz. $L\text{-}L_{min} = [APP_{max}/APP_{max(0)}]^{-0.4}$ (ms), can be viewed as a window during which the rate of change of peak pressure is integrated. The window commences with tone onset and its duration is inversely related to the RCPP or APP_{max} for linear and cosine-squared rise functions, respectively, and to the neuron’s transient sensitivity to that stimulus, quantified by either $RCPP_{(0)}$ or $APP_{max(0)}$.

In a sense the peak pressure at the instant of response generation is the threshold peak pressure of the neuron to a given stimulus. The fact that this threshold peak pressure varies with stimulus parameters (see Figure 6B,D) is therefore closely related to the finding that a fixed threshold model is not sufficient to explain the change in latency with stimulus parameters (cf. Figures 1C-F, 2). If the fixed-threshold model were correct, then spike count data would, of course, all fall on a vertical line when plotted against the peak pressure at the instant of response generation. The position of this line along the abscissa would mark the fixed threshold peak pressure.

When compared to the conventional spike-count-versus-level functions, particularly to those obtained with tones of longer rise times, the functions relating spike counts to the peak pressure at the instant of response generation are often characterized by relatively narrow dynamic ranges and, in case of non-monotonic functions, relatively sharp tuning (cf. Fig. 6B with A and D with C). These properties might be useful for the coding of instantaneous peak pressure with high resolution.

3. An Envelope-Tracking Mechanism

The joint consideration of the neuronal response properties and of their stimulus dependencies outlined above suggests a high-resolution envelope-tracking mechanism as described in the following and illustrated in Figures 7 and 8.

The response functions of a given neuron obtained with tones of a given frequency but of different rise times and rise functions are in close register when plotted as a function of peak pressure at the instant of response generation (rather than as a function of steady-state SPL) [10][14] (Figure 6). Therefore, a neuron’s onset response represents a sample of the tone’s envelope taken at a particular time, viz., the instant of response generation. For a given neuron, that specific instant in time depends on the dynamics of the amplitude at onset (Section 2.1). The magnitude of the neuron’s response depends on the amplitude of the tone at that instant (Section 2.2.). The response of an individual neuron may be ambiguous with respect to the instantaneous peak pressure. For example, a neuron with non-monotonic spike count functions will give the same response to two different peak pressures on either side of the one producing the maximum response. Unequivocal coding of the instantaneous amplitude could be achieved by computing the ratio of the responses of a sub-population of neurons whose responses are generated at the same instant (i.e., neurons with identical transient sen-

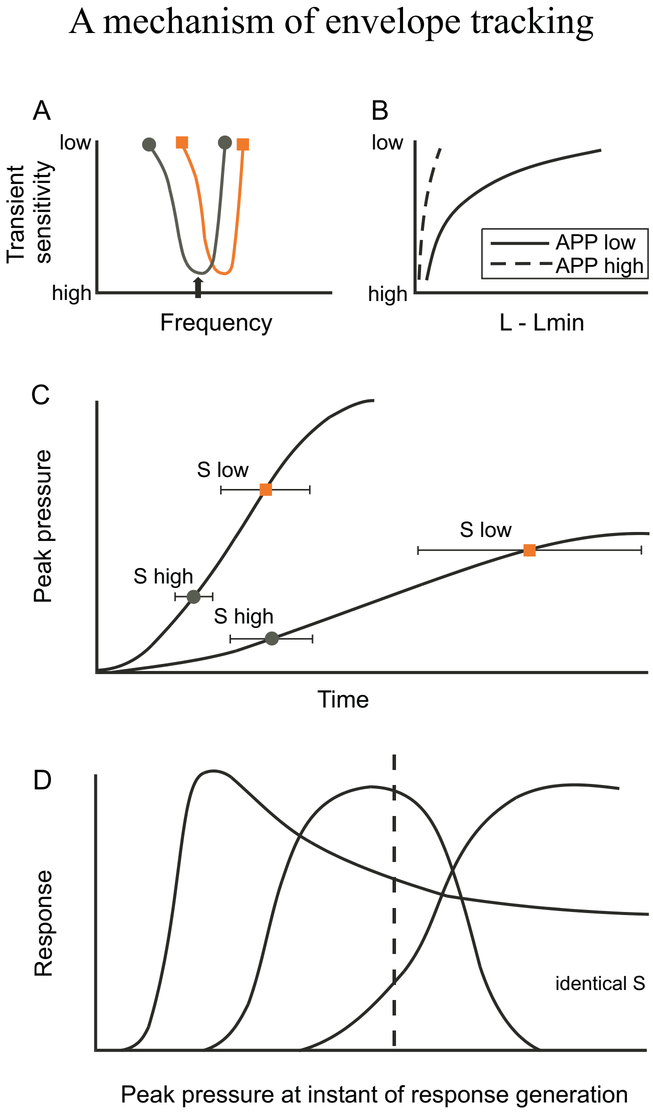


Figure 7 Proposed mechanism for envelope tracking. A. For a given neuron, the transient sensitivity, S (1/$APP_{max(0)}$ or 1/$RCPP_{(0)}$) varies with frequency or spectral content, and for a given frequency, S varies across neurons (cf. Figure 4). B. Integration times (i.e. L-L_{min}) of different neurons to a given signal depend on S, and for different signals on maximum acceleration of peak pressure (APP). C. A given envelope is sampled by neurons with different S. D. Each sample can be represented by neurons with the same S, but different response functions. Note that the sampling rate (i.e., the intervals between the initial spikes of different neurons, and the precision of timing (horizontal bars in C) are adjusted.

sitivity to that particular stimulus) and whose response functions partially overlap (Figure 7D). In the auditory cortex such neurons are topographically organized along the isofrequency contour of the tonotopic map [16][32]. For a given stimulus, the time of response generation varies for different neurons, depending on their transient sensitivity (i.e. $\mathrm{RCPP}_{(0)}$ or $\mathrm{APP}_{\max(0)}$) to that stimulus (see Sections 2.1.3 and 2.1.4 as well as Figure 7A,B). Thus, sub-populations with somewhat lower transient sensitivities to that stimulus generate their responses at successively later instances (Figure 7C) and could represent the instantaneous size of the envelope at these later instances in an analogous fashion. For a tone of a given frequency, this would involve neurons with CFs increasingly distant from the stimulus frequency, because for a given neuron the transient sensitivity varies with stimulus frequency (Figures 4 and 7A). In this way a population of neurons with different transient sensitivities could sample the changing envelope point by point and so track its time course (Figure 7).

This is illustrated in more detail in Figure 8. Panels A and B illustrate those portions of the transients and initial segments of the steady-state that are sampled by the population, depending on stimulus conditions (SPL and rise time). The continuous lines in A and B show the onset envelopes of tones, all of which have the same frequency but are shaped with cosine-squared rise functions of 20 ms (A) and 4-ms rise time (B) and differ in SPL. Different symbols represent different neurons, each with its own transient sensitivity to the tones, and identical symbols represent the same neuron. For clarity, some neurons are labeled with numbers and their symbols are interconnected. Each symbol indicates when relative to stimulus onset (and consequently at what instantaneous peak pressure) the response of a particular neuron is generated. These instances were calculated using Equation 1b, with c equal to 0.4 and a range of transient sensitivities corresponding approximately to that found around 20 kHz in the population of AI neurons (cf. Figure 4). Transient sensitivities of different neurons are equally spaced on a logarithmic axis, as illustrated in panels C and D, which plot the peak pressure at the instant of response generation as a function of those sensitivities. These panels also show that the ratio of peak pressure at which any two neurons generate their responses is largely invariant as a function of SPL. For example, the response of neuron 7 is generated at an instantaneous peak pressure about three times that of neuron 1, relatively independent of SPL. Near-constant ratios are also obtained for other neuron pairs, as evident from the relatively straight and parallel functions in these double-logarithmic plots. Ratios vary with SPL only at low intensities and when neurons are considered whose transient sensitivities are low (i.e., with high $\mathrm{APP}_{\max(0)}$), because their responses are generated during the initial portion of the tones´ steady state. However, these neurons respond very weakly or not at all to low-SPL stimuli [9]. Thus, in general it seems that the response of each individual neuron is generated at an instantaneous peak pressure that varies with stimulus parameters, such as SPL and rise time (Figures 6 and 8). However, the responses of any two neurons are generated at instantaneous peak pressures with a fixed ratio, largely independent of SPL. Hence, the neuronal population would provide a similar relative resolution of amplitude for stimuli of different SPLs.

The functions of latency versus RCPP or $\mathrm{APP}_{\max}$ of different neurons are of essentially identical shape (Figure 3) [9][13][14]. Consequently, and as long as jitter in spike timing is not taken into account (see below), the temporal sequence with which different neurons will generate their spikes will remain constant when the time course of the envelope is altered (e.g., by a change in the SPL or in the rise time of the signal, or both). Note that in Figure 8 the response of neuron 1 is always generated before that of neuron 2, and that of neuron 2 always before that of neuron 3, etc., independent of the signal's SPL or rise time.

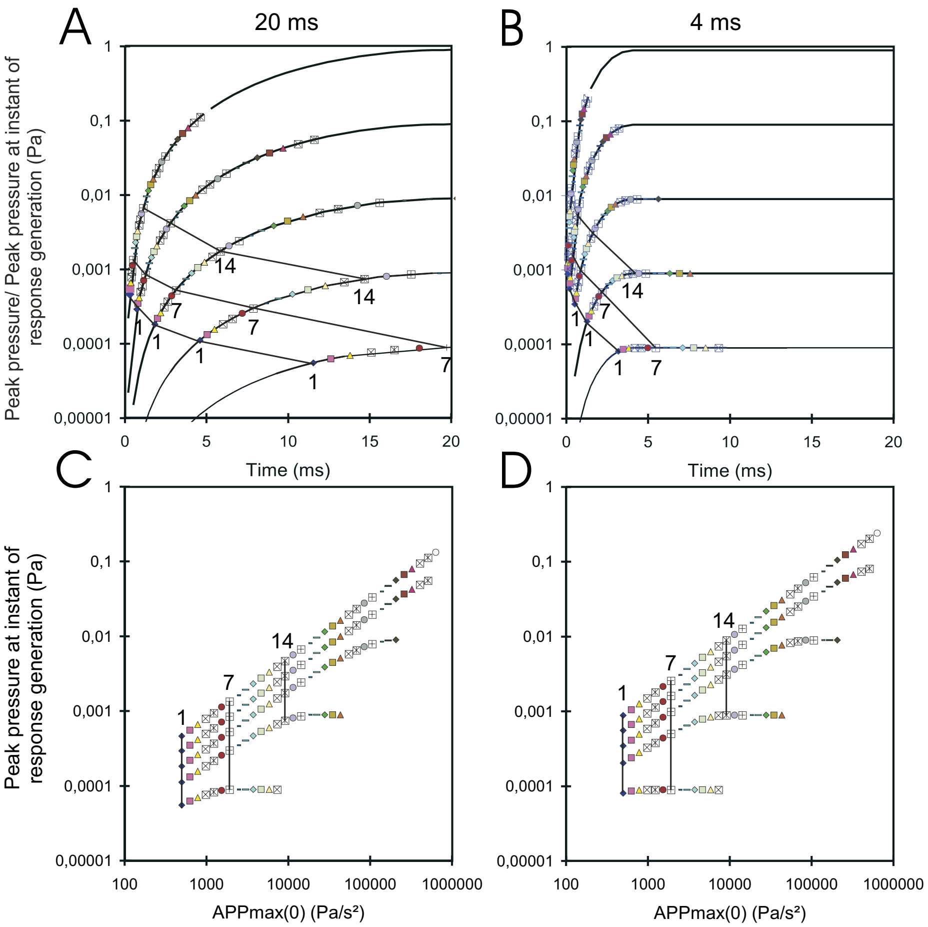


Figure 8 Tracking of different envelopes by a neuronal population. Curved lines in A and B show the initial envelopes of tones of 10, 30, 50, 70, and 90 dB SPL (note the logarithmic axis of peak pressure in Pa), shaped with cosine-squared rise functions of 20 ms (A) and 4-ms rise time (B). Different symbols represent different neurons, each with its own transient sensitivity ($\text{APP}_{\max(0)}$) to the tones, and identical symbols represent the same neuron. Some neurons are labeled with numbers and their symbols are interconnected. Each symbol indicates when relative to stimulus onset and at what instantaneous peak pressure, the response of a particular neuron is generated. These instances were calculated using Equation 1b, with c = 0.4 and a range of transient sensitivities corresponding approximately to that found around 20 kHz in the population of AI neurons (cf. Figure 4). C and D. Plots of the peak pressure at the instant of response generation against those transient sensitivities ($\text{APP}_{\max(0)}$). Transient sensitivities of different neurons are equally spaced on a logarithmic axis. In A and B, note that the sampling rate is adjusted to the rapidity of the envelope changes and that each envelope produces a particular spatio-temporal pattern of first spikes. In C and D, note that the ratio of peak pressures at which any two neurons generate their responses is largely invariant with SPL, except at low SPLs and for neurons with low sensitivity (high $\text{APP}_{\max(0)}$).

However, due to the (common) shape of the latency functions, the intervals between the first spikes of different neurons (i.e., the sampling rate) will vary. When the envelope has a shallow slope or a low acceleration (and hence has a slowly changing time course) the intervals between the onset responses of different neurons are long (i.e., the sampling rate is low). When the envelope has a steep slope or a high acceleration, and hence has a more rapidly varying time course, the sampling rate is high (Figures 7B, C and 8). This certainly holds with respect to different signals. Whether this also applies to the changing time course of any given signal, as Figures 8A and B suggest, depends on details of the distribution of transient sensitivities to the signal and also depends on whether the changing amplitude is viewed on a linear or on a logarithmic scale. In any event, the degree of temporal dispersion of the responses of different neurons would contain information about the time course of the envelope, even though such responses might appear largely synchronous [4][5].

This proposed tracking mechanism relies on the orderly temporal sequence of initial spikes in the responsive population of neurons. Consequently, the representation of fine stimulus details would be limited by the precision of spike timing relative to the sampling rate. The precision of spike timing is high when the latency is short and vice versa (Figure 5) [9][13][14][26]. This is important, because the change in the precision of spike timing counteracts an increase in the temporal overlap of response initiation among the successively activated neurons that occurs with an increase in the rapidity of the transient (e.g., by an increase in a signal's SPL (Figure 7)). It is remarkable that neurons with poorly timed onset responses (prevalent in the cat's posterior auditory field) often do not respond to rapid transients, such as the onsets of tones of high SPL and short rise time [14] or rapid frequency-modulations [15], a correlation which is meaningful in the context of an envelope tracking mechanism.

4. Summary and Conclusions

The joint consideration of response latency, precision of response timing and response magnitude of neurons in the cat's auditory cortex as well as the elaboration of the stimulus and neuronal properties on which they depend, suggest the operation of an envelope-tracking mechanism. This mechanism is characterized by automatic adjustments in sampling rate and precision of spike timing, as well as roughly constant ratios of instantaneous amplitudes at which the responses of any two neurons are generated. These properties make this envelope coding mechanism rather robust against variations in the rapidity of the envelope, brought about by changes in a signal's SPL. The spatio–temporal response patterns produced by various onset envelopes involve both the tonotopic and the isofrequency axes of cortical maps. It appears that such a mechanism could be of a temporal resolution that is orders of magnitude greater than those inferred from the phase-locking capabilities of neurons to amplitude-modulated or other periodic signals, and thus may contribute to the instantaneous coding of transients thought to underlie the categorical perception of speech and certain nonlinguistic sounds. Future research will show whether, and to what extent, the mechanism proposed in this chapter applies to the coding of various periodically amplitude-modulated sounds and thus might contribute to mechanisms mediating pitch and timbre.

Acknowledgements

I am grateful to Professor Dexter R. F. Irvine for help with the experiments and to members of the general staff in the Department of Psychology, Monash University, Clayton, Australia, where the experiments were conducted. H. Neubauer suggested Equation 1 and Dr. Michael Brosch commented on the manuscript. The study was supported by a project grant from the National Health and Medical Research Council of Australia.

References

[1] Boer, E. de "Pitch of inharmonic signals." *Nature*, 178: 535–536, 1956.

[2] Drullman, R., Festen, J. M., and Plomp, R. "Effect of temporal envelope smearing on speech reception." *J. Acoust. Soc. Am.*, 95: 1053–1064, 1994.

[3] Eggermont, J. J. "Differential effects of age on click-rate and amplitude modulation-frequency coding in primary auditory cortex of the cat." *Hear. Res.*, 65: 175–192, 1993.

[4] Eggermont, J. J. "Functional aspects of synchrony and correlation in the auditory nervous system." *Concepts Neurosci.*, 4: 105–129, 1993.

[5] Eggermont, J. J. "Firing rate and firing synchrony distinguish dynamic from steady state sound." *Neuroreport*, 8: 2709–2713, 1997.

[6] Gooler, D. M. and Feng, A. S. "Temporal coding in the frog auditory midbrain: the influence of duration and rise-fall time on the processing of complex amplitude-modulated stimuli." *J. Neurophysiol.*, 67: 1–22, 1992.

[7] Grey, J. and Gordon, J. "Perceptual effects of spectral modifications on musical timbres." *J. Acoust. Soc. Am.*, 63: 1493–1500, 1978.

[8] Hall, J. C. and Feng, A. S. "Temporal processing in the dorsal medullary nucleus of the northern leopard frog (*Rana pipiens pipiens*)." *J. Neurophysiol.*, 66: 955-–973, 1991.

[9] Heil, P. "Auditory cortical onset responses revisited: I. First-spike timing." *J. Neurophysiol.*, 77: 2616–2641, 1997.

[10] Heil, P. "Auditory cortical onset responses revisited: II. Response strength." *J. Neurophysiol.*, 77: 2642–2660, 1997.

[11] Heil, P. "Further observations on the threshold model of latency for auditory neurons." *Behav. Brain Res.*, 95: 233–236, 1998.

[12] Heil, P. and Irvine, D. R. F. "On determinants of first-spike latency in auditory cortex." *Neuroreport*, 7: 3073–3076, 1996.

[13] Heil, P. and Irvine, D. R. F. "First-spike timing of auditory-nerve fibers and comparison with auditory cortex." *J. Neurophysiol.*, 78: 2438–2454, 1997.

[14] Heil, P. and Irvine, D. R. F. "The posterior field P of cat auditory cortex: coding of envelope transients." *Cerebral Cortex*, 8:125–141, 1998.

[15] Heil, P. and Irvine, D. R. F. "Functional specialization in auditory cortex: responses to frequency-modulated stimuli in the cat's posterior auditory field." *J. Neurophysiol.*, 79: 3041–3059, 1998.

[16] Heil, P., Rajan, R., and Irvine, D. R. F. "Topographic representation of tone intensity along the isofrequency axis of cat primary auditory cortex." *Hear. Res.*, 76: 188–202, 1994.

[17] Krahe, R. and Ronacher, B. "Long rise times of sound pulses in grasshopper songs improve the directionality cues received by the CNS from the auditory receptors." *J. Comp. Physiol. A*, 173: 425–434, 1993.

[18] Krumhansl, C. L. and Iverson, P. "Perceptual interactions between musical pitch and timbre." *J. Exp. Psychol.*, 18: 739–751, 1992.

[19] Langner, G. "Periodicity coding in the auditory system." *Hear. Res.*, 60: 115–142, 1992.

[20] Langner, G., Sams, M., Heil, P., and Schulze, H. "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from magnetoencephalography." *J. Comp. Physiol. A*, 181: 665-–676, 1997.

[21] Paquette, C. and Peretz, I. "Role of familiarity in auditory discrimination of musical instruments: A laterality study." *Cortex*, 33: 689-–696, 1997.

[22] Patterson, R. D. "The sound of a sinusoid: spectral models." *J. Acoust. Soc. Am.*, 96: 1409–1418, 1994.

[23] Patterson, R. D. "The sound of a sinusoid: time interval models." *J. Acoust. Soc. Am.*, 96: 1419–1428, 1994.

[24] Phillips, D. P. "Effect of tone-pulse rise time on rate-level functions of cat auditory cortex neurons: excitatory and inhibitory processes shaping responses to tone onset." *J. Neurophysiol.*, 59: 1524–1539, 1988.

[25] Phillips, D. P. "Neural representation of sound amplitude in the auditory cortex: Effects of noise masking." *Behav. Brain Res.*, 37:197–214, 1990.

[26] Phillips, D. P. "Neural representation of stimulus times in the primary auditory cortex." *Ann. NY Acad. Sci.*, 682: 104–118, 1993.

[27] Phillips, D. P. and Hall, S. E. "Response timing constraints on the cortical representation of sound time structure." *J. Acoust. Soc. Am.*, 88: 1403–1411, 1990.

[28] Phillips, D. P., Semple, M. N. and Kitzes, L. M. "Factors shaping the tone level sensitivity of single neurons in posterior field of cat auditory cortex." *J. Neurophysiol.*, 73: 674–686, 1995.

[29] Pitt, M. A. and Crowder, R. G. “The role of spectral and dynamic cues in imagery for musical timbre.” *J. Exp. Psychol.*, 18: 728–738, 1992.

[30] Plomp R. and Steeneken H. J. M. “Pitch versus timbre.” *Seventh Intern. Congr. Acoust.*, Budapest, pp. 387–390, 1971.

[31] Schouten, J. F. “The residue revisited,” *Frequency Analysis and Periodicity Detection in Hearing*. R. Plomp. and G. F. Smoorenburg (eds.), Leiden: Sijthoff, pp. 41–54. 1970.

[32] Schreiner, C. E., Mendelson, J. R., and Sutter, M. L. “Functional topography of cat auditory cortex: Representation of tone intensity.” *Exp. Brain Res.*, 92: 105–122, 1992.

[33] Schreiner, C. E. and Urbas, J. V. “Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields.” *Hear. Res.*, 32: 49–64, 1988.

[34] Semple, M. N. and Kitzes, L. M. “Binaural processing of sound pressure level in cat primary auditory cortex: evidence for representation based on absolute levels rather than interaural level differences.” *J. Neurophysiol.*, 69: 449–461, 1993.

[35] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. “Speech recognition with primarily temporal cues.” *Science*, 270: 303–304, 1995.

[36] Suga, N. “Responses of inferior colliculus neurons of bats to tone bursts with different rise times.” *J. Physiol. London,* 217: 159–177, 1971.

[37] Yost, W. A., Patterson, R., and Sheft, S. “The role of the envelope in processing iterated rippled noise.” *J. Acoust. Soc. Am.*, 104: 2349–2361, 1998.

FUNCTIONAL SIGNIFICANCE OF LATENCIES IN THE AUDITORY SYSTEM

Stefan Bleeck[1] and Gerald Langner[2]

[1]*Center for the Neural Basis of Hearing*
Physiology Department,University of Cambridge
CB2 3EG Cambridge, United Kingdom

[2]*Institute of Zoology, Schnittspahnstrasse, 3*
Technical University of Darmstadt
D-64287 Darmstadt, Germany

1. Introduction

Temporal processing became an important topic in auditory research when it was discovered that nerve fibers can code signals by firing action potentials at precise points during a stimulus. It soon became evident that temporal delays in the range of a few milliseconds play an important role in auditory processing. It appears that delays and latencies are not disturbing physical restrictions but may contain information useful for the analysis of sound. It is therefore interesting to study latencies at every stage of auditory processing. Latencies are defined here as the time between the start of a stimulus and when the neuron fires. Possible sources for latencies are propagation delays on dendrites and axons, physical restrictions of the motion on the basilar membrane, inhibitory effects from other neurons and intrinsic neuronal properties like integration times, intrinsic oscillations or secondary messenger cascades.

The following examples speak for the importance of latencies in the auditory system.

1.1 Direction

The superior olivary complex is known to be responsive to interaural disparities in intensity and timing [21]. The direction of a sound source can be detected by correlating the timing information from both ears. The underlying mechanism is a network of neurons that receives input from both ears with varying latencies. These latencies are produced by delays between cochlea and olivary neurons. When a signal arrives at one ear with a certain time delay relative to the arrival at the other ear, a neuron with interaural delays appropriate for this condition is activated. Such neurons are said to code the direction of a sound in space on the basis of its different input latencies. The task of these delays is to compensate for the different arrival times of signals from both ears, so that coincidence detection can be performed. The temporal precision for this mechanism is on the order of microseconds for barn owls [6].

1.2 Periodicity

A map of latency was described in the inferior colliculus of the cat [16][22]. Latency was defined as the time difference between tone onset and the first spikes of the response. The latency axis was orthogonal to the tonotopic axis in the IC [23]. A correlation between latency and best modulation frequency indicates that latency is important for periodicity analysis [15]. Moreover, the range of latency between 7 and 18 ms, and the orientation of the latency maps, are comparable to the periodicity map found in the same region.

1.3 Echolocation

In a similar fashion, delay-tuned neurons in the midbrain of the big brown bat are also arranged in a map in the IC [9]. Delays are defined here as the time between two particular FM components of a pulse-echo pair [5]. Delay tuning means that neurons are characterized by a specific “best delay” response that is facilitated by a pair of stimuli. Each stimulus was characterized by a pair of particular frequencies and amplitudes and separated by a specific delay. The map of delay was arranged orthogonal to the map of frequencies. Such delays are in the range from 0.8-50 ms, which corresponds to a echo range of 0.3 to 16 m.

1.4 Autocorrelation

Autocorrelation is a powerful tool to analyze acoustic stimuli as well as neuronal responses to acoustic stimuli. Moreover, some authors believe that autocorrelation holds the key to deciphering auditory processing [1][2].

Latencies are an essential part of every correlation function in general and autocorrelation in particular. The autocorrelation function

$$y(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(t) f(t+\tau) dt$$

describes the similarity of the function *f(t)* (spike train, sound, etc.) with itself after a time $t+\tau$. When for a delay τ the result of the correlation is high, then the function is similar to itself with a period of τ. If *f(t)* is a sound, the correlation is sufficiently strong and τ is in the proper range, we perceive a pitch corresponding to $1/\tau$.

The parameter τ describes a temporal delay. This delay is the significant variable in the correlation. When we assume that such a correlation of spike trains is performed by neurons, then delays must be present somewhere in the neuronal system.

Many theories of acoustic perception use the (auto)correlation function to determine the perception of the pitch of a complex tone [17][1][20] or of other auditory sensations like rhythms [24] or binaural space location [5]. It is striking that temporal delays in all these examples range from 0.8 ms to 50 ms. This corresponds to the range of the repetition rate of periodicity pitch between 20 and 1250 Hz. Likewise it is the important range of echo delays for bats to detect prey in the distance between 0.3 and 16 meters.

2. Latencies Measured in Auditory-Nerve Fibres

Auditory-nerve fibers are capable of coding various temporal and static aspects of the stimulus. Well-known features are frequency (by spike intervals) and volume (by spike rate). In addition, more complex stimuli properties like periodicity are coded temporally by auditory nerve fibers [1]. Even complex perceptional features such as the pitch shift are coded in the spike train of auditory-nerve fibers and if analyzed by autocorrelation, may be explained by a correlation analysis in the auditory system [17][2][19].

Heil proposed recently [11][12][also this volume], that a hitherto ignored feature is coded in the temporal course of the spike train: the temporal derivatives of the stimulus envelope.

In the following, we discuss measurements of latencies from neurons in the auditory cortex of the cat. Figure 1 shows a plot of the data we will model (from [12][14]). All data are

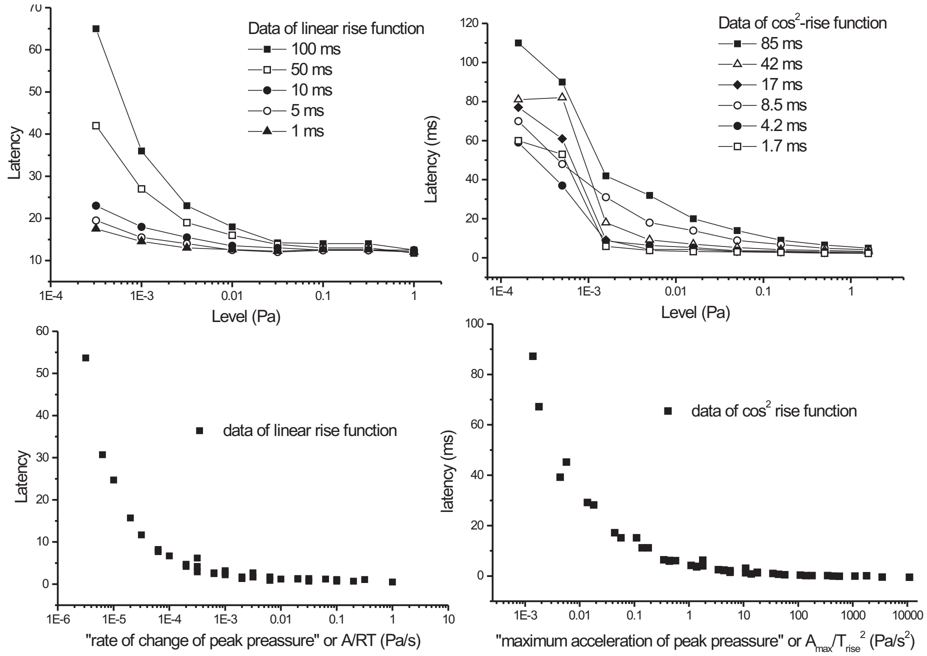


Figure 1 Original data from a cat with (A) a linear rise function [12], neuron 95-95/04) and (B) cos^2-rise function [14] neuron 97-107/35). Top: First-spike latency obtained with tones of different rise times plotted as a function of amplitude. Bottom: First-spike latency plotted as a function of the velocity of the envelope (linear) and maximum acceleration (cos^2).

taken from [12]. The stimuli in this experiment were pure tones with different onset envelopes. Because we are only interested in the beginning of the signal, the only important parameters for the present discussion are maximum amplitude, A_{max}, risetime, T_{rise} and the shape of the envelope during the rise time. Two different envelope shapes were used by Heil: linear and cos^2-rise functions (see Figures 2 and 3). The left side shows the results of an experiment with linear rise functions. The right side shows experiments with cos^2-rise rise times. Two different neurons are shown here. These data were chosen because they were pre-

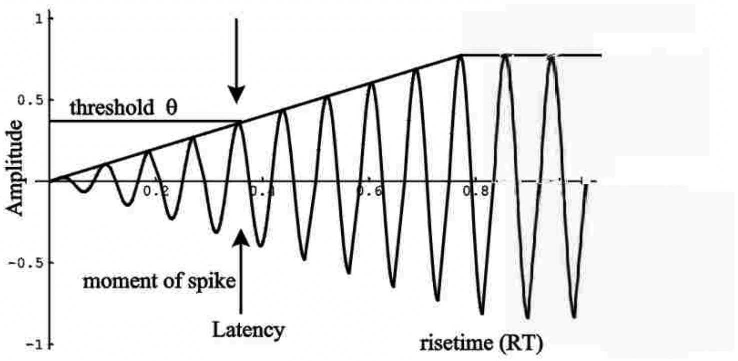


Figure 2 Sketch of the situation for the linear rise time paradigm. The amplitude rises to the final amplitude A_{max} during its rise time. The threshold Q is reached after a time, which is defined as the latency L.

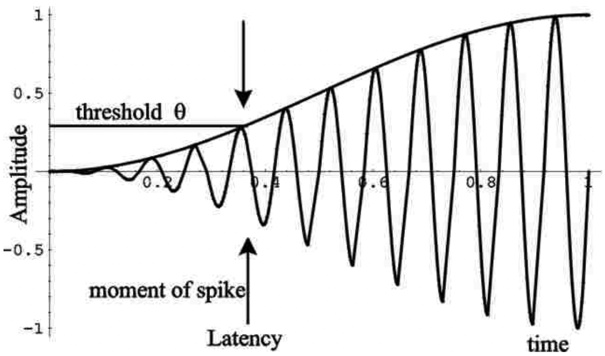


Figure 3 A sketch of the situation for the $\cos^2$-rise-time paradigm as in in Figure 1. The threshold, θ, is reached after a certain time which is the latency L.

sented in great detail in the original papers. Nevertheless, by selecting the data inaccurately, a small deviation from the original data might exist, especially for small latencies.

In the upper part of Figure 1 the latency is plotted versus the signal's amplitude. Different rise times are indicated by different symbols. The stimuli on the left side consisted of sinusoidal signals with $\cos^2$-rise functions and rise times between 1.7 ms and 85 ms. On the right side are shown the sinusoidal signals with a linear rise function between 1 and 100 ms. As can be seen, the latency of the first spike depends not only on the amplitude of the stimulus, but also on its rise time. Latency decreases with increasing amplitude, but increases with rise time even for the same maximum amplitude. In the lower picture the data are plotted in a different way — latency is now plotted against (in Heil's terms) "maximum acceleration of peak pressure" ($\cos^2$) or "rate of change of peak pressure" (linear), instead of amplitude as in the upper pictures. Now all the data overlap in a single curve. The upper and lower figures show the same data; only the x-values are different.

The maximum acceleration of a $\cos^2$-function is reached at the beginning of the stimulus. As shown in Section 3.2, it can be computed from rise time (T_{rise}) and maximum amplitude (A_{max}) by A_{max}/T_{rise}^2 It therefore seems reasonable to say that acceleration (or velocity, depending on the actual rise function) is a better parameter for latency than the signal amplitude and risetime. The primary aim of this chapter is to discuss this assumption in detail.

In the case of $\cos^2$-rise time signals, the curves for different rise times overlap when plotted versus the maximum of the second temporal derivative of the stimulus envelope (the acceleration). In the case of linear rise time, there is no acceleration but now the curves overlap when plotted versus the first temporal derivative, which is the velocity of the envelope. The velocity of the envelope of the linear rise function is a constant during the rise time and is equal to A_{max}/T_{rise}

In the following section we will show why a simple threshold neuron is able to explain the measured data with high fidelity.

3. A Simple Threshold Model

We consider a very simple model of a neuron which fires always at the monent when it reaches its threshold. We assume that the current envelope of the stimulus is somehow translated instantaneously to the membrane potential of the model neuron. When the membrane potential reaches the neuron's threshold, the model neuron immediately fires an action potential. No further latencies and no adaptation mechanisms are needed. The threshold is fixed to a constant value. Of course, this is a very simple model of a neuron, but we will see how much of the data we can explain. In the following sections we will explore several features of this model by testing on the data obtained with linear- and $\cos^2$-rise functions.

3.1 Envelope, Threshold and Temporal Derivatives

3.1.1 Linear Rise Functions

For linear rise times the time course of the envelope $A_{lin}(t)$ is

$$A_{lin}(t) = \begin{cases} \frac{A_{max}}{T_{rise}} t & t \le T_{rise} \\ A_{max} & t > T_{rise} \end{cases}$$

Here, and in the following, $A_{lin}(t)$ means the temporal course of the *linear* envelope. A_{max} is the maximal amplitude of the envelope, $A(t)$, that is reached (and held) after the risetime. T_{rise} is the risetime, that is, the time from $A(0)=0$ to $A(t=T_{rise}) = A_{max}$. Our simple neuron fires when the envelope reaches its threshold, Θ:

$$A_{lin}(t) = \frac{A_{max}}{T_{rise}} t = \Theta .$$

Solving for t, which is, by definition, the latency L_{lin}, gives

$$L_{lin} = \frac{\Theta T_{rise}}{A_{max}}$$

The latency describes the time between tone onset ($A(t) = 0$) and the time of the spike. The threshold, Θ, must be smaller than A, otherwise the neuron would never react. In reality, spontaneous activity sets an upper border to the latency.

The first temporal derivative of the linear rising envelope during the rise time is

$$\frac{d}{dt} A_{lin}(t) = \frac{A_{max}}{T_{rise}}$$

which is constant in time.

3.1.2 Cos^2-Rise Functions

In the $\cos^2$-case, the stimulus envelope is described by

$$A_{\cos^2}(t) = \begin{cases} A_{max} \cdot cos^2\left(\frac{\pi}{2T_{rise}} t\right) \quad or \quad A_{max} \cdot sin^2\left(\frac{\pi}{2T_{rise}} t\right) & t \le T_{rise} \\ A_{max} & t > T_{rise} \end{cases}$$

Again the simple threshold neuron always fires when it reaches its fixed threshold, Θ:

$$\Theta = A \cdot \sin^2\left(\frac{\pi}{2T_{rise}} t\right)$$

or solved for $t = L_{cos}{}^2$:

$$L_{\cos^2} = \frac{2}{\pi} T_{rise} a \sin\left(\sqrt{\frac{\Theta}{A_{max}}}\right).$$

Figure 3 shows a sketch of the $\cos^2$-rise function. Assume that the maximum amplitude is much higher than threshold ($\Theta << A$). Then the term can be approximated by

$$L_{\cos^2} \approx \frac{2}{\pi} T_{rise} \sqrt{\frac{\Theta}{A_{max}}}$$

Figure 1 shows the latency for $\cos^2$-rise functions versus the maximum acceleration of peak pressure (or the second temporal derivative of the envelope).

The second temporal derivative of the $\cos^2$-function is equal to

$$\frac{d^2}{dt^2} A_{\cos^2}(t) = \frac{A_{max}\pi^2}{2T_{rise}^2} \cos\left(\frac{\pi t}{T_{rise}}\right)$$

The maximum acceleration is reached when $\cos\left(\frac{\pi t}{T_{rise}}\right) = 1$. This is true already at the beginning of the stimulus ($t=0$). Therefore, the maximum acceleration simplifies to

$$\ddot{A}_{\cos^2} = \frac{d^2}{dt^2} A_{\cos^2}(0) = \frac{A_{max}\pi^2}{2T_{rise}^2}$$

3.2 Transforming the Data

The amazing property of the data in Figure 1 is that they converge onto the same curve when plotted versus the appropriate variable. For the $\cos^2$-case this is maximum acceleration and in the linear case it is the velocity of the envelope. In the following, we will show that this is a simple result of the threshold model.

3.2.1 Linear Rise Function

If we plot the latency achieved from the simple threshold model over maximum amplitude we obtain hyperbolas of the form $L_{lin} = \Theta \frac{T_{rise}}{A_{max}}$. From the first temporal derivative of the envelope $\dot{A}_{lin} = \Theta \frac{A_{max}}{T_{rise}}$ we see, that $L_{lin} = \Theta \frac{1}{\dot{A}_{lin}}$.

In other words, when data from an experiment with a linear rise function are plotted versus the first temporal derivative of the envelope, they follow a single 1/x function. In other words, when we use a transformed x-axis for plotting, all of the data converge onto a single curve. The resulting curve is a 1/x function, which is no longer an explicit function of A_{max} and T_{rise}. The only free parameter in this equation is the threshold, Θ, which in the threshold model is a constant. Therefore, all measurements of latencies using different combinations of A_{max} and T_{rise} result in the same curve.

Note that the velocity of the envelope is not an explicit parameter of the threshold model. Therefore, the only information signaled by a spike of such a neuron is that its threshold was reached at a certain time, but not which velocity was present.

Access to the velocity requires prior knowledge, in this instance, that the stimulus is a linear-rise function.

3.2.2 Cos^2-Ramp

Above, we found the latency of the threshold model for a cos^2-rise function is equal to

$$L_{\cos^2} = T_{rise}\sqrt{\Theta/A_{max}}$$

and the second temporal derivative is equal to

$$\ddot{A}_{\cos^2} = (A_{max}\pi^2)/2T_{rise}^2$$

We can therefore write the neuron's latency as

$$L_{\cos^2} = \pi\sqrt{\Theta/2}\cdot 1/(\sqrt{\ddot{A}_{lin}})\,.$$

This means that latency can be written as function of acceleration. Neither amplitude nor rise time is an explicit parameter anymore. The resulting function is of the form $1/(\sqrt{x})$. In other words, the threshold model predicts that when data from an experiment with a cos^2-rise function are plotted over the maximum second temporal derivative of the envelope they are modeled by a single $1/(\sqrt{x})$ curve. The resulting function is no longer an explicit function of A_{max} or T_{rise}. All other parameters in the equation are constant. Therefore, all combinations of A_{max} and T_{rise} leads to the same curve. The response of the neuron does not indicate that a certain acceleration was present, but only, that its threshold was reached.

3.3 Comparison with the Observed Data

We are now able to predict from the threshold model the measured latency of the neuron. For a given rise function, the calculated latency can be written as a function of the temporal derivatives.

The computations above showed that the threshold model predicts that the data for first spike latency can be fitted to a function of A_{max}/T_{rise_2} or the velocity for linear envelopes. The form of the resulting function is *1/x*. For the cos^2-rise function, we expect a $1/(\sqrt{x})$-function to fit the data when plotted over A_{max}/T_{rise}^2 , which corresponds to the acceleration of the envelope at *t = 0*

3.3.1 Linear Rise Function

Figure 4 (replotted from Figure 1) shows the experiment with linear rise functions. The measured latencies are plotted over A_{max}/T_{rise} . Additionally, a fit with a *1/x*-function is plotted. This fit is the prediction from the simple threshold model, assuming a fixed threshold. In this case the assumed threshold was 38.8 dB and the minimum latency L_{min} was 13.6 ms.

Figure 5 shows the measured latencies plotted versus the calculated latencies. Additionally a straight line with a slope of one is plotted. The fit is quite good, but there are systematic errors of this simple model. For very short and very long latencies the calculation slightly overestimates the measured values. For latencies between 1 and 6 ms the predictions from the threshold model underestimates the observed latencies of the neuron.

3.3.2 Cos^2-case

Figure 6 data shows the data from the experiment with cos^2 -rise function, together with the proposed fit from the simple threshold model. The assumed threshold was 31.1 dB and

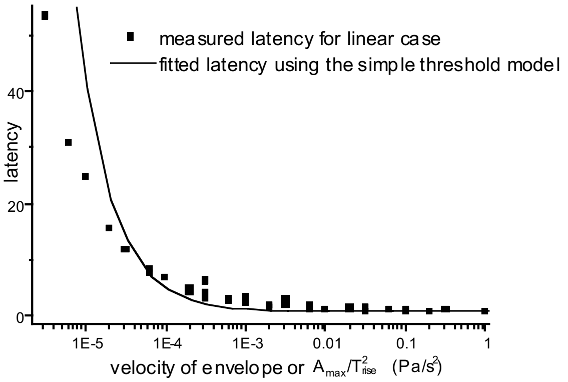


Figure 4 Data from the linear case shown in Figure 1 are plotted against A/T. Additionally, the fit with a 1/x function is plotted.

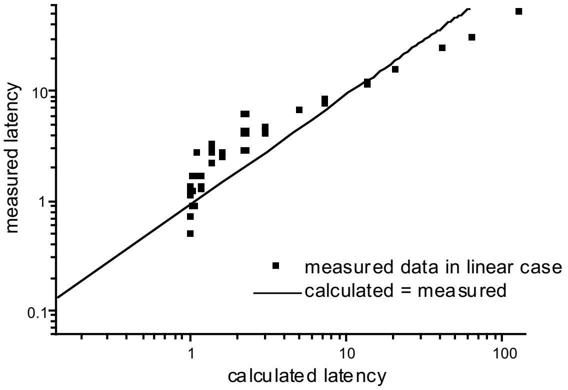


Figure 5 Plot of the measured latencies in the case of a linear function over the calculated latencies. The straight line has a slope of one..

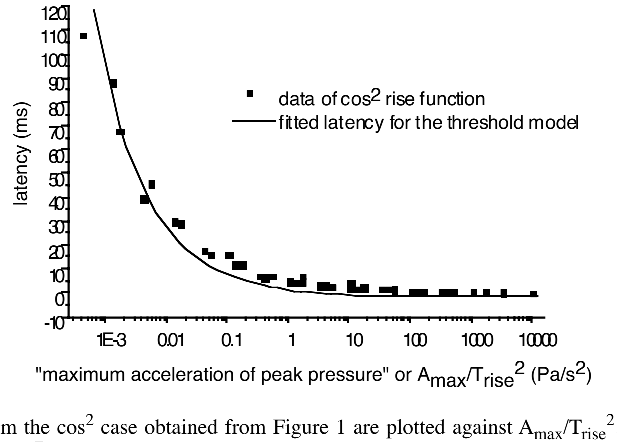


Figure 6 Data from the cos^2 case obtained from Figure 1 are plotted against A_{max}/T_{rise}^2. Additionally a fit with a $1/(\sqrt{x})$ -function is also plotted.

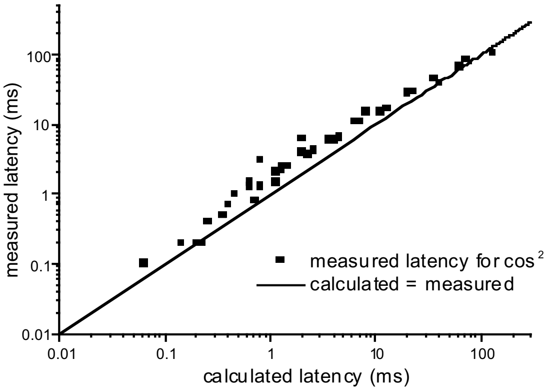


Figure 7 Plot of the measured latencies in the case of the cos^2-rise function over the calculated latencies. The straight line has a slope of one.

the minimum latency 3.16 ms. The fit has the form of a $1/(\sqrt{x})$- function as explained above. Data are plotted against A_{max}/T_{rise}^2.

Again, we can calculate how long the latency should be, assuming a fixed threshold. Figure 7 shows the result of this comparison. The calculated latencies are compared directly with the measured latencies. Additionally a straight line with a slope of one is shown. As can be seen, the prediction from the threshold model is very good. But again the latencies are systematically underestimated for latencies in the medium range.

4. Modification of the Simple Threshold Model

The calculated fits are already quite good, especially for the cos^2-case. They cover most of the characteristics of the measured data. However, the systematic deviations suggest that it might be possible to make a more accurate model. In the following a fit is presented that is derived from a more complex model for the linear-rise data.

The fit for the threshold model chosen for the data in Figure 8 is not the optimal fit shown in Figure 5. Instead, the minimum latency was chosen to be equal to 11 ms, a little bit smaller than before, so that most of the calculated data had a positive deviation from the

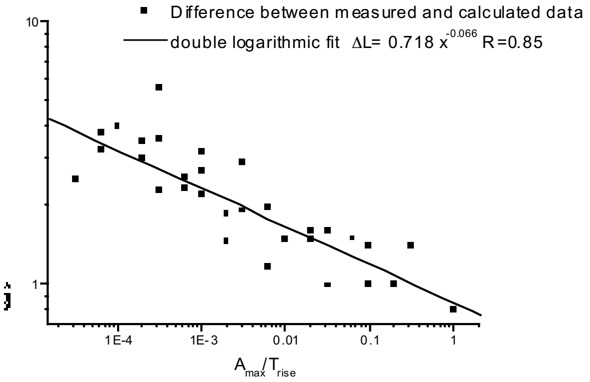


Figure 8 Deviation between data from linear rise function and the fit from the threshold model.

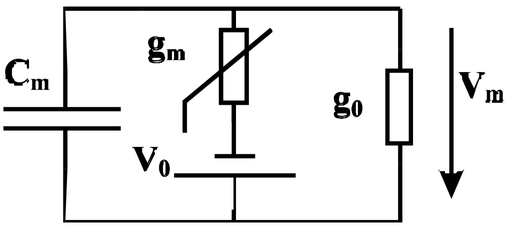


Figure 9 The implementation of a leaky, integrate-and-fire neuron used in the simulation C_m = membrane capacity, g_m = controlled membrane conductance = f(transmitter in cleft), g_0 = leakage conductance, V_0 = steady state potential, V_m = membrane potential.

measurement. Only 3 points with very high A_{max}/T_{rise}-relation, which are not plotted, had a negative calculated difference.

As we can see, the calculated data differ slightly but systematically from the measured data. It is therefore straightforward to include a time constant in the simple model. Such a time constant should lead to delayed responses for slower rise functions.

In the following we present a leaky, integrate-and-fire model with fixed threshold (Figure 9) that has several time constants that might explain the prolonged delays. The new model has the disadvantage of increased complexity in comparison to the simple threshold model, but the advantage that all its parts are associated with parts of real neurons.

Because we are interested in the precise timing of firings, it is necessary to use a neural model that is dynamically coupled to the transmitter concentration in the synaptic cleft. The spike generating part is designed as a leaky, integrate-and-fire unit. The main parts of the neuron membrane model are the membrane capacity, C_m, and the leaky conductance, g_0. To simulate the transmitter input we use the controlled membrane conductance, g_m, which is coupled with the steady-state potential, V_0. The membrane conductance depends on the amount of transmitter in the synaptic cleft.

Assume that the amount of transmitter released per time in the synaptic cleft is proportional to the amplitude of the stimulus $\dot{T}(t)_{gain} \propto A(t)$, where *T(t)* is the amount of transmitter in the synaptic cleft and *A(t)* is the stimulus amplitude. We further assume that transmitter is lost over time at a rate proportional to the amount of the transmitter present. The temporal change of the amount of transmitter in the cleft is therefore

$$\dot{T}(t) = A(t) \cdot \kappa T(t)$$

For the case of linear rise functions, solving for *T(t)* in this differential equation gives the amount of transmitter in the cleft time *t:*

$$T(t) = \frac{A_{max}}{\kappa^2 T_{rise}} \frac{1 - e^{\kappa t} + \kappa t e^{\kappa t}}{e^{\kappa t}}$$

The conductance of the membrane g_m is proportional to the amount of transmitter $g_m(t) = \chi \cdot T(t)$, where χ is a constant factor. We can now define the membrane time constant $\tau = C_m/[g_0 + g_m(t)]$ and the conductance value $\gamma = 1 + g_0/g_m$

The potential of the whole membrane is described by the difference equation:

$$\dot{U}(t) \cdot \tau(t) + U(t) = \frac{V_0}{\gamma(t)}$$

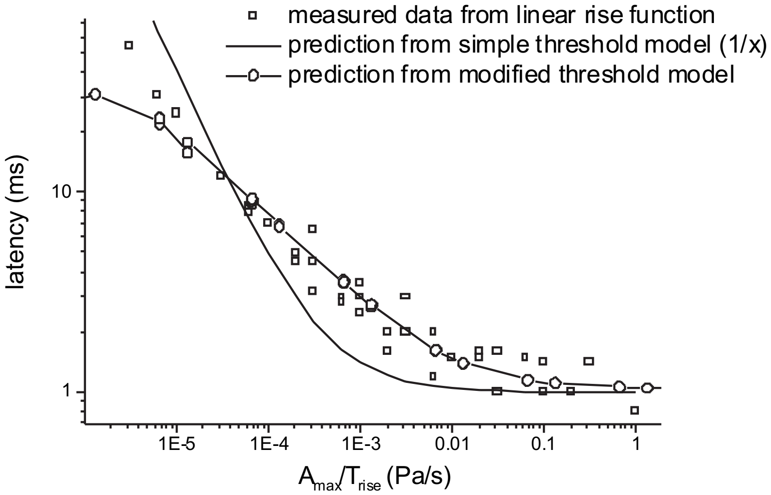


Figure 10 Result obtained from the simulated neuron with a fixed threshold for linear rise times. The open circles represent the simulation data from the leaky, integrate-and-fire neuron. Filled squares are the original data from Figure 1. The line indicates the prediction from the simple threshold model.

where *U(t)* describes the potential of the membrane. The neuron fires again when it reaches a fixed threshold. The task is therefore to find an equation for *U(t)* and to then solve *U(t)=Θ* to find *t*. The resulting time, *t,* is the latency, *L*. Surprisingly, we can solve the equation analytically for *U(t)*, but the result is not shown due to a length limitation on this chapter. Unfortunately, it is not possible to solve the result for the latency *L* analytically. Therefore, we simulated the system numerically. With the constants V_0 = potential of sodium (Na$^+$)=60mV and *SR* = sampling-rate (45.45 μs), C_m = 0.8 μF, g_0 =10 μS the membrane potential is given by discretization of *U(t)*. The result of the discretization is an iteration formula, which describes how large *U(t)* is after one time step *U(t-SR)*.

$$U(t) = \frac{V_0 \cdot SR + \gamma(t)\tau(t)}{\gamma(t) \cdot (\tau(t) + SR)} \cdot U(t - SR)$$

U(t) is 0 at *t* = *0*. The simulated neuron fires every time when its membrane potential exceeds the threshold of the neuron. The “zero”-threshold, Θ_0, was set to 14.8 mV. When modeling onset latencies only the first spike is important. Therefore, in this case the threshold is a constant. For linear-rise functions we simulated rise times from 1 ms to 100 ms over a wide range of amplitudes. Results of the simulation are shown in Figure 10. As can be seen, the fit for the simulated data is indeed better then the fit from the simple threshold model, which is indicated in Figure 10 by a line.

Figure 11 shows the remaining error between a fit of the measured data and the calculation from the modified-threshold model. The first three values with very small velocities are not present in the plot, because their deviation is more then 30 ms and would require a different scaling. But all data with higher velocity have a deviation smaller than 1 ms. By choosing different parameters for the leaky, integrate-and-fire model the deviation might become even smaller.

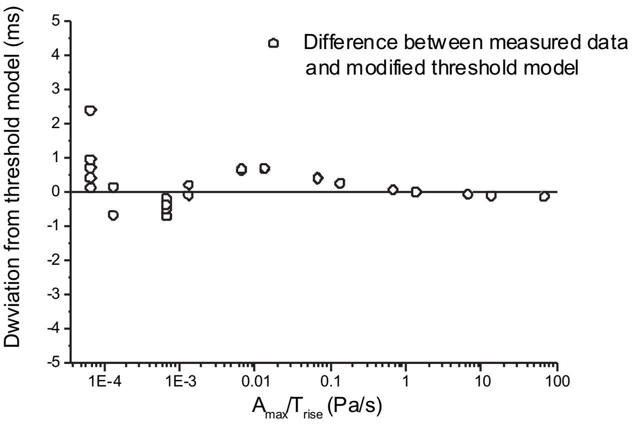


Figure 11 Comparison of the simulated data and a fit of the original data.

If we simulate more complex situations, when more than the first spike is important, the threshold change in dependent on spikes in τ seconds in the past according to

$$V_{threshold} = \Theta_0 + \left(2\frac{V_0}{\Theta_0} - 1\right)\exp\left\{\frac{-t - T_{last} - T_{absRef}}{\tau}\right\}$$

where T_{last} is the time of the last spike and T_{absRef} is the absolute refractory period. A small Gaussian noise is then added to the threshold in the simulations. This noisy behavior gives the neuron a small "memory" for the history of what it has done in the past. As mentioned previously, this changing threshold is probably not important for the observed first spike latency. Any model however, that models more then the first spike should be able to deal with changing thresholds.

5. Discussion

5.1 The Threshold Models

The simple threshold model is able to predict the first-spike latency of the observed neurons in general fashion. The result might be satisfying or not, depending on how precise a description of the onset latency is required. A more realistic and therefore more complex simulation of a leaky integrate-and-fire neuron is able to predict the latencies better. However the principal idea of a fixed threshold remains untouched. Consequently, it is possible to predict the latency of a neuron with high accuracy using a model that assumes a fixed threshold. The simplicity of this idea is striking.

It might be possible to provide a better fit to the data with different values for the numerous parameters. The idea however, that a neuron fires with a latency that is determined mainly by a fixed threshold is even more likely. It is not necessary (and conceptually misleading) to use other stimulus parameters such as the temporal derivatives to describe onset latency. It is, of course, not surprising that a model that includes a larger number of (and superior) selected parameters, as in the simple threshold model, leads to a better prediction for latency.

The temporal derivatives of the envelope rise functions are not an explicit feature of the integrate-and-fire model. It is a coincidence that they supply a convenient way of plotting latency data. The latency of the first spike is therefore determined only by the threshold and not by any derived envelope features.

It is not satisfying that specific forms of envelope rise functions should have any influence on the model parameters determining latency. The leaky integrate-and-fire neuron uses the same set of parameters for all rise functions, while the general interface between the neuron and the stimulus is given by the equation $\dot{T}_{gain}(t) \propto A(t)$.

Several interesting questions arise if the hypothesis that latency is determined by maximum acceleration of the envelope is true. Because the maximum acceleration is reached at the beginning of the stimulus, neurons would need to calculate their precise latencies during the first few milliseconds of the stimulus without knowledge of the signals time course in the future. Furthermore, at that moment the signal amplitude is zero. So, actually, no signal is present at all. This is a puzzling contradiction.

5.2 The Role of Delays for Information Processing

Intensity, frequency, and periodicity of a stimulus are coded by the rate and statistics of spikes. It is, however, possible to code temporal features such as the envelope waveform of the signal by using additional delays. Let us assume that every neuron has a fixed threshold. Every time the neuron reaches its threshold, it fires an action potential, as long as it is not in any absolute or relative refractory period. Furthermore, it takes some time for the spike to travel to the succeeding postsynaptic neuron. These two assumptions are realistic for real neuronal networks. The two free parameters, threshold and minimum latency, in this model correspond to the two free parameters in the threshold model.

Figure 12 shows how such an information coding could be performed. Several neurons with different thresholds fire at different times during the time course of the stimulus. In this example, thresholds increase from neuron *A* to neuron *E*. Neuron *A* reaches its threshold

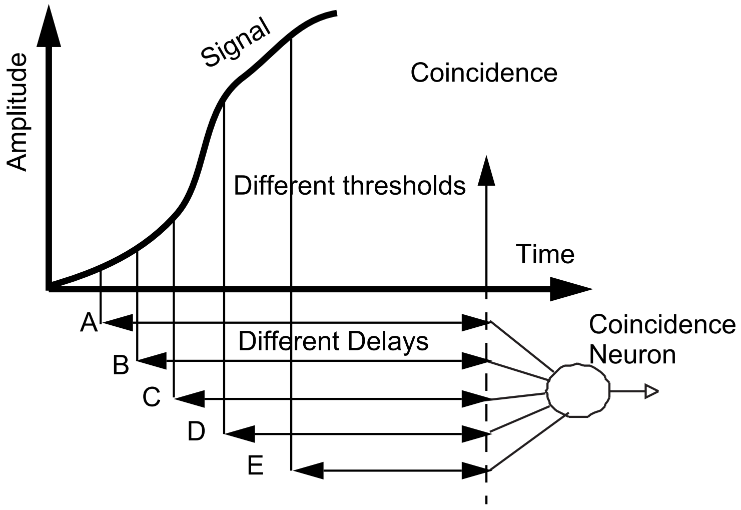


Figure 12 Hypothetical model of a "sampling" network. Several neurons (A-E) fire with different thresholds and their spikes arrive at a coincidence neuron with different delays.

first, when the amplitude of the stimulus is low. The delay between the neurons in this example and a coincidence neuron are chosen such that their spikes will arrive at the coincidence neuron at the same time. Neuron *A* has the lowest threshold, but its spikes arrive with the longest delay, while neuron *E* reaches its threshold at last, but its spikes arrives at the coincidence neuron with the shortest delay. Finally, all spikes arrive at the same time. By this mechanism it is possible to take a sample of the temporal waveform. If a certain coincidence neuron is active, the corresponding waveform have activated the neurons in this channel. It is known from physiology that auditory-nerve fibers indeed have different thresholds, filling the hearing range uniformly [8].

Several problems arise with such a system when used for signal detection: detection depends on signal amplitude and the phase of adjacent frequencies. It would be surprising if such a system would be sufficient to detect real, noisy signals. Nevertheless the given coincidence scheme is a simplified description of an auditory processing mechanism. The resulting correlation of input neurons should take place in more or less narrow frequency regions to avoid phase effects. This supposition corresponds to the idea of critical bandwidths. Frequencies outside a critical bandwidth do not show interaction in phase [21].

Simple coincidence neurons are sufficient to detect such temporal coincidences. Such a hypothetical coincidence neuron would fire after a particular complex waveform has activated more peripherally located neurons. Many such neurons together would be able to provide a sample of the whole waveform. But only when they are analyzed together does it become apparent which signal they code. While a record of a single neuron would bear only limited information about the signal. An interesting point about the sampling network is that single neurons do not code extracted signal features like periodicity, intensity or frequency directly. Instead, many neurons together code the whole signal temporally. The quality of the sampling depends only on the number of neurons used. If half of the neurons are removed, the sampling is of lesser quality. The described sampling mechanism might exemplify neuronal behavior at every level of the auditory pathway, since data from the auditory nerve up to auditory cortex show the required temporal behavior in high precision [11]. Every single spike has its own history. It is therefore straightforward to think of a model of auditory short term memory.

5.3 A Neuronal Network that can Handle Delays

Information coded by delays could be extracted from spike trains by neuronal networks that are sensitive to temporal delays. A possible implementation of such a network is shown in Figure 13. This network is a modification of a classical feed-forward network with spiking neurons and delays added between the neurons. When a neuron fires, the spikes arrive at the connected neurons after fixed delays. Weights and latencies are altered by learning rules in a simple fashion. The temporal learning rule in the model could be a variation of the Hebbian rule, “Fire nearby, fire closer,” (i.e., the latency of a connection is altered with a probability that depends on the number of spikes arriving in the (near) past [14]). Due to inherent delays this model is designed to find not only simple temporal features, but is able also to perform a correlation between different signals. Therefore, it is adequate to detect periodicity or echo delays. With the correct learning rule it can learn, in a self-organized way, temporal properties of the incoming signals. At present, the model is able to detect specific temporal features in a continuous spike-train. In the future, more complex stimuli will be presented to our network with the task of localizing stimuli and to detect periodicity. A simple enhancement to the model introduces recursive (inhibitory) connections.

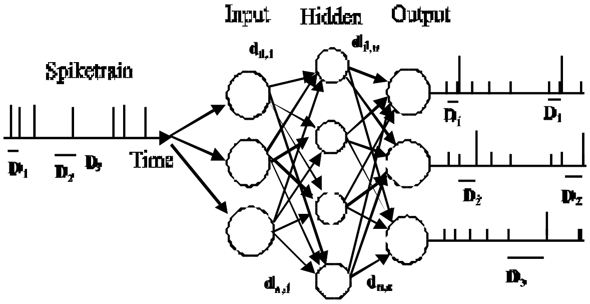


Figure 13 An implementation of the idea, that latencies carry information in an artificial neuronal network. The model is able to detect not only spatial patterns like any other perceptron network, but it can also detect temporal patterns.

The task of the model is to stay as simple as possible and to modulate complex timing behavior of neurons.

Neurons are viewed here as minimal parts of a distributed system in which no part has any information other then the temporal input from other neurons, either inhibitory or excitatory. The only informations are the spikes that travel from one neuron to another, and a learning rule that says what to do with synapse strength and latency. Except for potential information about its own location, this view of real neurons is an adequate description of what a real neuron sees. In any further model many parameters can be influenced by the location of the neurons: threshold, time constants, preferred input and so on. Since the described process is nothing more than a correlation function between several input functions (spike trains), it can be used as a minimum model for all latency-related phenomena described in the introduction.

6. Summary

Latencies are an important feature in auditory neuronal systems. These latencies can be used by neurons to produce delays and to analyze an input stream like sound for temporal features. Data that are obtained from measurements of first spike latency are in good correlation with a simple model of a neuron that has a fixed threshold and always fires when it reaches its threshold. An even better fit is possible when a more complex model is used that includes intrinsic time constants. The resulting neuron model is able to describe the measured data with a good fit. More complex envelope features like the acceleration or the velocity are not necessary to describe latencies.

Features of sound like direction and periodicity can be detected by neurons or networks of neurons that handle delays.

References

[1] Cariani, P. A. and Delgutte, B. "Neural correlates of the pitch of complex tones:1. Pitch and pitch salience." *J. Neurophysiol.* 76:1698–1716, 1996.

[2] Cariani, P. A. and Delgutte, B. "Neural correlates of the pitch of complex tones: 2. Pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch." *J. Neurophysiol.* 76:1717–1734, 1996.

[3] Casseday, J. H. and Covey, E. "Mechanisms for Analysis of Auditory Temporal Patterns in the Brainstem of Echolocating Bats." *Neur. Rep.*, 25–51: 1995.

[4] Dau, T., Puschel, D., and Kohlrausch, A. "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure." *J. Acoust. Soc. Am.* 99: 3615–3622, 1996.

[5] Dear, S. P. and Suga, N. "Delay-tuned neurons in the midbrain of the big brown bat." *J. Neurophysiol.* 73:1084–1100, 1995.

[6] Gerstner, W., Kreiter, A. K., Markram, H., and Herz, A. V. "Neural codes: Firing rates and beyond." *Proc. Natl. Acad. Sci. U.S.A,* 94:12740–12741, 1997.

[7] Glünders, H. and Hünning, H. "Detection of spatio-temporal spike patterns by unsupervised synaptic delay learning." *Göttingen Neruobiologentagung*, 1996

[8] Hartmann, W. M. *Signals, Sound, and Sensation*. New York: Springer-Verlag, 1997.

[9] Hattori, T. and Suga, N. "The inferior colliculus of the mustached bat has the frequency- vs.-latency coordinates." *J. Comp. Physiol.* A 180: 271–284, 1997.

[10] Heil, P. and Irvine, D. R. F. "On determinants of first-spike latency in auditory cortex." *Neur. Rep.* 7: 3073–3076, 1996.

[11] Heil, P. and Irvine, D. R. "First-spike timing of auditory-nerve fibers and comparison with auditory cortex." *J. Neurophysiol.* 78:2438–2454, 1997.

[12] Heil, P. "Auditory cortical onset responses revisited: 1. First-spike timing." *J. Neurophysiol.* 77: 2616–2641, 1997.

[13] Heil, P. "Auditory cortical onset responses revisited: 2. Response strength." *J. Neurophysiol.* 77:2642–2660, 1997.

[14] Heil, P. and Irvine, D. R. "First-spike timing of auditory-nerve fibers and comparison with auditory cortex." *J. Neurophysiol.* 78:2438–2454, 1997.

[15] Langner, G., Schreiner, C. E., and Merzenich, M. M. "Covariation of latency and temporal resolution in the inferior colliculus of the cat." *Hearing Res.* 31:197–202, 1987.

[16] Langner, G. and Schreiner, C. E. "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms." *J. Neurophysiol.* 60: 1799–1822, 1988.

[17] Langner, G. "Periodicity coding in the auditory system." *Hearing Res.* 60:115–142, 1992.

[18] Langner, G. "Neural processing and representation of periodicity pitch." *Acta Otolaryngol.,* 532: 68–76, 1998

[19] Meddis, R. and Hewitt, M. J. "Virtual pitch and phase sensitivity of a computer-model of the auditory periphery. I: Pitch identification." *J. Acoust. Soc. Am.* 89: 2866–2882, 1991.

[20] Meddis, R. and O'Mard, L. "A unitary model of pitch perception." *J. Acoust. Soc. Am.* 102:1811–1820, 1997.

[21] Pickles, J. O. *An Introduction to the Physiology of Hearing*. London: Academic Press, 1988.

[22] Schreiner, C. E. and Langner, G. "Periodicity coding in the inferior colliculus of the cat. II. Topographical organization." *J. Neurophysiol.* 60: 1823–1840, 1988.

[23] Schreiner, C. E. and Langner, G. "Laminar fine structure of frequency organization in auditory midbrain." *Nature* 388: 383–386, 1997.

[24] Todd, N. P. M. "The kinematics of musical expression." *J. Acoust. Soc. Am.* 97: 1940–1949, 1995.

A COMPUTATIONAL MODEL OF CO-MODULATION MASKING RELEASE

Masashi Unoki and Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai Tatsunokuchi Nomi Ishikawa, 923-1292, Japan

1. Introduction

In investigations of the frequency selectivity of the auditory system, the power-spectrum model of masking [6] is widely accepted as an explanation of the phenomenon of masking. This model assumes that when a listener tries to detect a sinusoidal signal amid background noise he makes use of the output of a single auditory filter having its center frequency close to the signal frequency and having the highest signal-to-masker ratio. In addition, it assumes that the stimuli are represented by long-term power spectra, and that the masking threshold for the sinusoidal signal is determined by the amount of noise passing through the auditory filter. With these assumptions, the power spectrum model explains many masking phenomena such as simultaneous masking. However, this model cannot explain all masking phenomena because it ignores the relative phases of the components and the short-term fluctuations in the masker.

In 1984, Hall *et al*. demonstrated that across-filter comparisons can enhance the detection of a sinusoidal signal in a fluctuating noise masker [3]. The crucial feature for achieving this enhancement is that the fluctuations are coherent or correlated across different frequency bands. They called this across-frequency coherence in their demonstrations "co-modulation." Therefore, the enhancement in signal detection obtained using coherent fluctuation, i.e., this reduction in masking threshold, was called "Co-modulation Masking Release" (CMR). Many psychoacoustical experiments were carried out [7][4][9] and the same phenomenon was repeatedly demonstrated. The experiments revealed the condition when CMR can occur. But so far, no computational model has been proposed that takes advantage of across-frequency coherence.

On the other hand, the human auditory system can easily segregate the desired signal in a noisy environment that simultaneously contains speech, noise, and reflections. Recently, this ability of the auditory system has been regarded as a function of an active scene analysis system. Called "Auditory Scene Analysis" (ASA), it has become widely known as a result of Bregman's book [1]. Bregman reported that the human auditory system uses four heuristic regularities related to acoustic events to solve the problem of Auditory Scene Analysis. These regularities are (i) common onset and offset, (ii) gradualness of change, (iii) harmonicity, and (iv) changes occurring in the acoustic event [2].

In this work we tackle the problem of segregating the desired signal from a noisy signal [8] using Bregman's regularities [2]. We stress the need to consider not only the amplitude spectrum but also the phase spectrum when attempting to completely extract the signal from noise, both of which are present in the same frequency region [8]. Based on this approach,

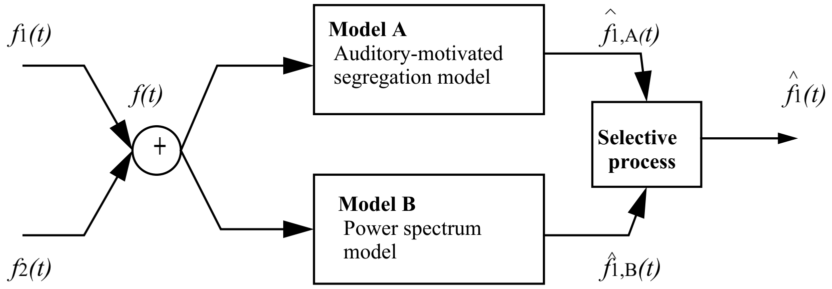


Figure 1 Computational model of CMR. This model consists of two models — our auditory-motivated segregation model (model A) and the power spectrum model of masking (model B) — followed by a selection process that selects one of their results.

we seek to solve the problem of segregating two acoustic sources — the basic problem of acoustic source segregation using regularities (ii) and (iv) of Bregman's principles [2].

This paper proposes a computational framework for CMR that consists of two models — our auditory-motivated segregation model and the power spectrum model of masking proposed by Patterson *et. al.* — followed by a selection process.

2. Computational Model of CMR

Our computational model of CMR is shown in Figure 1. It consists of two models (A and B) and a selection process. In this model, we assume that $f_1(t)$ is a sinusoidal signal and $f_2(t)$ is one of two types of noise masker (bandpassed random noise and AM bandpassed random noise) whose center frequency is the same as the signal frequency. We also assume that the sinusoidal signal $f_1(t)$ is added to $f_2(t)$. Since the proposed model can observe only the mixed signal $f(t)$, it extracts the sinusoidal signal $f_1(t)$ using the two models (A and B). Model A is the auditory-motivated segregation model we proposed earlier [8]. Model B is the power spectrum model of masking [6].

We propose a computational framework for CMR, where these two models work in parallel and extract a sinusoidal signal from the masked signal. Here, let $\hat{f}_{1,\mathrm{A}}(t)$ and $\hat{f}_{1,\mathrm{B}}(t)$ be the sinusoidal signals extracted using models A and B, respectively. The fundamental idea arises from the fact that the masking threshold increases as the masker bandwidth increases, up to the bandwidth of the signal auditory filter (1 ERB) and then it either remains constant or decreases depending on the coherence of the fluctuations. Thus, model B can explain part of CMR by using the output of a single auditory filter when the masker bandwidth increases up to 1 ERB. Model A can explain part of CMR by using the outputs of multiple auditory filters when the masker bandwidth exceeds 1 ERB.

3. Model A: Auditory-Motivated Segregation Model

The auditory-motivated segregation model shown in Figure 2 consists of three parts: (a) an auditory filterbank, (b) separation block, and (c) grouping block. The auditory filterbank is constructed using a gammatone filter as an "analyzing wavelet." The separation block uses physical constraints related to heuristic regularities (ii) and (iv) proposed by Bregman [2]. The grouping block synthesizes each separated parameter and then reconstructs the extracted signal using the inverse wavelet transform.

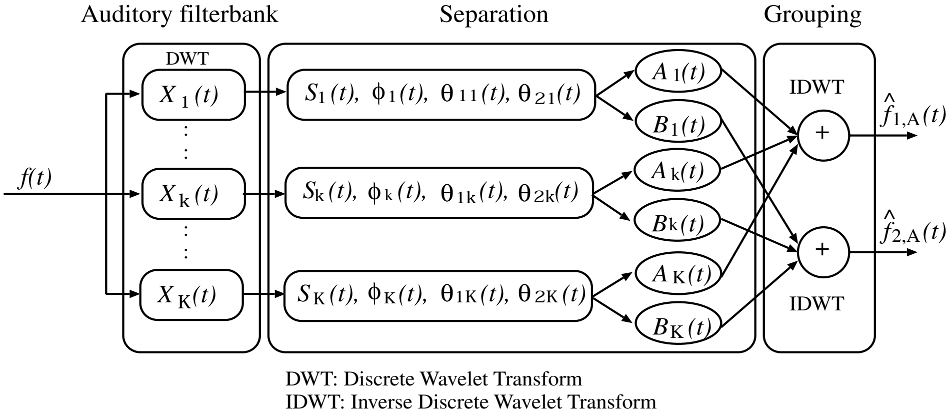

Auditory filterbank
Separation
Grouping
DWT
$f(t)$
$X_1(t)$
$X_k(t)$
$X_K(t)$
$S_1(t),\ \phi_1(t),\ \theta_{11}(t),\ \theta_{21}(t)$
$S_k(t),\ \phi_k(t),\ \theta_{1k}(t),\ \theta_{2k}(t)$
$S_K(t),\ \phi_K(t),\ \theta_{1K}(t),\ \theta_{2K}(t)$
$A_1(t)$
$B_1(t)$
$A_k(t)$
$B_k(t)$
$A_K(t)$
$B_K(t)$
IDWT
+
+
IDWT
$\hat{f}_{1,A}(t)$
$\hat{f}_{2,A}(t)$
DWT: Discrete Wavelet Transform
IDWT: Inverse Discrete Wavelet Transform

Figure 2 Model A: an auditory-motivated segregation model. This model consists of three parts: (a) an auditory filterbank, (b) separation block, and (c) grouping block.

3.1 Auditory Filterbank

An auditory filterbank is constructed using the wavelet transform, where the basic function $\psi(t)$ is the impulse response of the gammatone filter [5] which is represented using the Hilbert transform.

$$\psi(t) = At^{N-1}\exp(j2\pi f_0 t - 2\pi f_b t), \tag{1}$$

where $\mathrm{ERB}(f_0) = 24.7(4.37(f_0/1000)+1)$ and $f_b = 1.1019\ \mathrm{ERB}(f_0)$. This is a constant Q filterbank having a center frequency f_0 of 1 kHz, a bandpass region from 100 Hz to 10 kHz, and 128 channels. The bandwidth of each auditory filter is 1 ERB. In addition, we compensate for the group delay by adjusting the peak in the envelopes of Equation (1) for all scale parameters, which is called “alignment processing,” because a different group delay occurs at each scale.

3.2 Separation and Grouping

First, we can observe only the signal $f(t)$, where $f(t) = f_1(t) + f_2(t)$, $f_1(t)$ is the desired signal and $f_2(t)$ is a noise masker. The observed signal $f(t)$ is decomposed into its frequency components by an auditory filterbank. Second, the output of the k-th channel, corresponding to $f_1(t)$ and $f_2(t)$, are assumed to be narrow-band sinusoids

$$f_1(t):\quad A_k(t)\sin(\omega_k t + \theta_{1k}(t)), \tag{2}$$

and

$$f_2(t):\quad B_k(t)\sin(\omega_k t + \theta_{2k}(t)). \tag{3}$$

Here, ω_k is the center frequency of the auditory filter and $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are the input phases of $f_1(t)$ and $f_2(t)$, respectively. Since the output of the k-th channel $X_k(t)$ is the sum of Equations (2) and (3),

$$f(t) = S_k(t)\sin(\omega_k t + \phi_k(t)). \quad (4)$$

Therefore, the amplitude envelopes of the two signals $A_k(t)$ and $B_k(t)$ is equal to

$$A_k(t) = \frac{S_k(t)\sin(\theta_{2k}(t) - \phi_k(t))}{\sin\theta_k(t)}, \quad (5)$$

and

$$B_k(t) = \frac{S_k(t)\sin(\phi_k(t) - \theta_{1k}(t))}{\sin\theta_k(t)}, \quad (6)$$

where $\theta_k(t) = \theta_{2k}(t) - \theta_{1k}(t)$ and $\theta_k(t) \neq n \times \pi, n \in \mathbf{Z}$. Since the amplitude envelope $S_k(t)$ and the output phase $\phi_k(t)$ are observable, then if $\theta_{1k}(t)$ and $\theta_{2k}(t)$ are determined, $A_k(t)$ and $B_k(t)$ can be determined using the equations above. Finally, all the components are synthesized from Equations (2) and (3) in the grouping block. Then $f_1(t)$ and $f_2(t)$ can be reconstructed by the grouping block using the inverse wavelet transform. Here, $\hat{f}_{1,\mathrm{A}}(t)$ and $\hat{f}_{2,\mathrm{A}}(t)$ are the reconstructed versions of $f_1(t)$ and $f_2(t)$, respectively.

In this paper, we assume that the center frequency of the auditory filter corresponds to the signal frequency. Therefore, we consider the problem of segregating $f_1(t)$ from $f(t)$ when $\theta_{1k}(t) = 0$ and $\theta_k(t) = \theta_{2k}(t)$.

3.3 Calculating the Four Physical Parameters

The amplitude envelope $S_k(t)$ and phase $\phi_k(t)$ of $X_k(t)$ are determined using the amplitude and phase spectra. Since $\theta_{1k}(t) = 0$, we must find the input phase $\theta_{2k}(t)$. It can be determined by applying three physical constraints, derived from regularities (ii) and (iv), as shown below [8].

Constraint 1. Gradualness of change (slowness)

Regularity (ii) means that “a single sound tends to change its properties smoothly and slowly (gradualness of change)” [2]. The first constraint we describe as “slowness,” is $dA_k(t)/dt = C_{k,R}(t)$, where $C_{k,R}(t)$ is an R-th-order differentiable polynomial. By applying this constraint to Equation (5), and solving the resulting linear differential equation, we obtain

$$\theta_{2k}(t) = \arctan\left(\frac{S_k(t)\sin\phi_k(t)}{S_k(t)\cos(\phi_k(t) - C_k(t))}\right), \quad (7)$$

where $C_k(t) = \int C_{k,R}(t)dt + C'_{k,0}$. Here, we assume that in a small segment, Δt , $C_{k,R}(t) = C_{k,0}$.

Constraint 2. Gradualness of change (smoothness)

The second constraint we describe as “smoothness.” At the boundary ($t = T_r$) between the earlier segment ($T_r - \Delta t \leq t < T_r$) and succeeding segment ($T_r \leq t < T_r + \Delta t$),

$$|A_k(T_r + 0) - A_k(T_r - 0)| \leq \Delta A \quad (8)$$

$$|B_k(T_r+0)-B_k(T_r-0)| \le \Delta B \tag{9}$$

$$|\theta_{2k}(T_r+0)-\theta_{2k}(T_r-0)| \le \Delta\theta \tag{10}$$

From the above relationships, we can use this constraint to determine $C_{k,0}$, which must satisfy $C_{k,\alpha} \le C_{k,0} \le C_{k,\beta}$. The variables $C_{k,\alpha}$ and $C_{k,\beta}$ are the upper and lower values of $C_{k,0}$ when $\theta_{2k}(t)$ is determined by substituting any value of $C_{k,0}$ for $C_{k,R}(t)$ and then Equations (8)–(10) are satisfied.

Constraint 3. Changes occurring in an acoustic event (regularity)

Regularity (iv) means that "many changes take place in an acoustic event that affect all the components of the resulting sound in the same way and at the same time" [2]. The third constraint, which we describe as regularity, is

$$\frac{B_k(t)}{\|B_k(t)\|} \approx \frac{B_{k\pm l}(t)}{\|B_{k\pm l}(t)\|}, \qquad l = 1, 2, \ldots, L \ , \tag{11}$$

where L is the number of adjacent auditory filters.

Here, a masker envelope $B_k(t)$ is a function of $C_{k,0}$ from Equations (6) and (7). We consider this constraint to select an optimal coefficient $C_{k,0}$ using

$$\max_{C_{k,\alpha} \le C_{k,0} \le C_{k,\beta}} \frac{\langle \hat{B}, \tilde{B} \rangle}{\|\hat{B}\| \cdot \|\tilde{B}\|} \ , \tag{12}$$

where $\hat{B}_k(t)$ is the masker envelope given by any $C_{k,0}$, and

$$\tilde{B}_k(t) = \frac{1}{2L} \sum_{l=-L,\, l\ne 0}^{L} \frac{\hat{B}_{k+l}(t)}{\|\hat{B}_{k+l}(t)\|} \,. \tag{13}$$

Hence, the above computational process can be summarized as follows: (a) a general solution of $\theta_{2\mathrm{k}}(t)$ is determined using physical constraint 1; (b) candidates of $C_{k,0}$ that can uniquely determine $\theta_{2k}(t)$ are determined using physical constraint 2; (c) an optimal $C_{k,0}$ is determined using physical constraint 3; and (d) $\theta_{2k}(t)$ is uniquely determined by the optimal $C_{k,0}$.

In this chapter, we consider the problem of segregating a masked sinusoidal signal in which the localized signal $f_1(t)$ is added to the noise $f_2(t)$. Therefore, when we solve the above problem using the proposed method, we must know the duration for which two acoustic signals overlap. This can be determined by detecting the onset and offset of $f_1(t)$. By focusing on the temporal deviation of $S_k(t)$ and $\phi_k(t)$, we can determine onset $T_{k,\mathrm{on}}$ and offset $T_{k,\mathrm{off}}$ of $f_1(t)$ as follows:

1. Onset $T_{k,\mathrm{on}}$ is determined by the nearest maximum point of $|d\phi_k(t)/dt|$ (within 25 ms) relative to the maximum point of $dS_k(t)/dt$.
2. Offset $T_{k,\mathrm{off}}$ is determined by the nearest maximum point of $|d\phi_k(t)/dt|$ (within 25 ms) relative to the minimum point of $dS_k(t)/dt$.

The segregated duration is $T_{k,\mathrm{off}} - T_{k,\mathrm{on}}$.

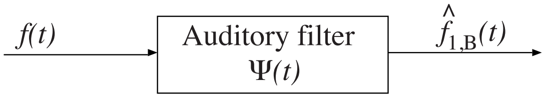


Figure 3 Model B: a power spectrum model of masking.

4. Model B: The Power Spectrum Model of Masking

In the power spectrum model [6], we assume that when a listener is trying to detect a sinusoidal signal with a particular center frequency amid background noise, he uses the output of a single auditory filter whose center frequency is close to the signal frequency, and which has the highest signal-to-masker ratio. Therefore, we assume that only the component passed through a single auditory filter affects masking. In particular the masking threshold for a sinusoidal signal is determined by the amount of noise passing through the auditory filter.

The power spectrum model consists of model B as shown in Figure 3. This filter consists of a gammatone filter whose center frequency is 1 kHz and bandwidth is 1 ERB. In this model, the sinusoidal signal $\hat{f}_{1,B}(t)$ extracted from the masked signal $f(t)$ is the output of the single auditory filter $X_k(t)$.

5. Simulations

5.1 Co-modulation Masking Release

Hall *et al*. measured the masking threshold for a sinusoidal signal in one of their experiments as a function of the bandwidth of a continuous noise masker. They used a center frequency of 1 kHz, a duration of 400 ms and kept the spectrum level constant [3]. They used two types of masker — a random noise masker and an amplitude modulated random noise masker — which were both centered at 1 kHz. The random noise masker had irregular fluctuations in amplitude, and the fluctuations in different frequency regions were independent. The amplitude-modulated masker was a random noise that was amplitude modulated at an irregular, slow rate; a noise that was lowpass filtered at 50 Hz was used as a modulator. Therefore, fluctuations in the amplitude of the noise in different spectral regions were the same.

Figure 4 shows the results of that experiment. For the random noise (denoted by R), the signal threshold increased as the masker bandwidth increased up to ca. 100–200 Hz, and then remained constant. This is exactly as expected from the traditional model of masking. The auditory filter at this center frequency had a bandwidth of ca. 130 Hz. Hence, for noise bandwidths up to about 130 Hz, increasing the bandwidth increased the noise passing through the filter, so the signal threshold increased. In contrast, increasing the bandwidth beyond 130 Hz did not increase the noise passing through the filter, so the threshold did not increase. The pattern for the modulated noise (denoted by M) was quite different. For noise bandwidths greater than 100 Hz, the signal threshold decreased as the bandwidth increased. This indicates that subjects could compare the outputs of different auditory filters to enhance signal detection. The fact that the threshold decreased with increasing bandwidth only with modulated noise indicates that fluctuations in the masker are critical and that the fluctuations need to be correlated across frequency bands. Hence, this phenomenon has been called “co-modulation masking release” (CMR). The amount of CMR in that experiment, defined as the difference in thresholds for random noise and modulated noise, was at most 10 dB [3].

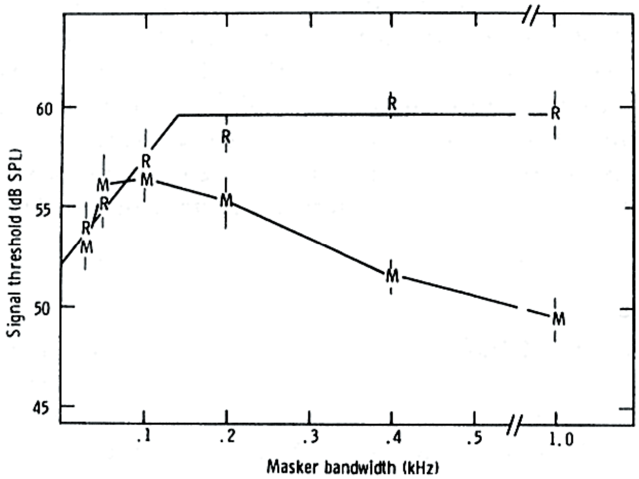


Figure 4 Results for CMR (Hall *et al*, 1984). The points labeled 'R' are thresholds for a 1-kHz signal centered in a band of random noise, plotted as a function of the bandwidth of the noise. The points labeled "M" are the thresholds obtained when the noise was amplitude modulated at an irregular,

5.2 Simulations for Model A

5.2.1 Stimuli and Procedure

We considered conditions equivalent to the experimental ones used by Hall *et al.* In this simulation we assumed that $f_1(t)$ was a sinusoidal signal, where the center frequency was 1 kHz, the duration was 400 ms, and the amplitude envelope was constant, and the masker $f_2(t)$ was two types of bandpassed noise having its center frequency close to the signal frequency. One was a bandpassed random noise $f_{21}(t)$ and other was an AM bandpassed random noise $f_{22}(t)$. The AM masker was calculated by amplitude modulating $f_{21}(t)$, where the modulation frequency was 50 Hz and the modulation rate was 100%. Here, the power of the noise masker $f_2(t)$ was adjusted so that $\sqrt{f_{21}(t)^2/f_{22}(t)^2} = 1$. Moreover the power ratio between $f_1(t)$ and $f_2(t)$, i.e., the SNR (signal-to-noise ratio), was –6.6 dB.

In this simulation, we must determine the number of adjacent auditory filters, *L,* to use in Equation (11). However we don't know this number when CMR occurs. We don't know which channels actually contribute to the CMR effect observed in psychoacoustics. We assume that the number of relevant auditory filters required in this model is simply determined by the total masker bandwidth. To realize the different experimental conditions, the initial bandwidth of the masking noise($f_{21}(t)$ and $f_{22}(t)$) was kept constant at 1 kHz, and only the number of auditory filters to be processed by the model was adjusted. The mixed signals were $f_R(t) = f_1(t) + f_{21}(t)$ and $f_M(t) = f_1(t) + f_{22}(t)$, corresponding to the stimuli in Figure 4 labeled R and M, respectively. Simulation stimuli, consisting of 10 sinusoidal signals, were formed by varying the onset. 30 maskers of the two types were generated by varying the random seeds. Thus, the total number of stimuli was 300. For example, one of the two types of mixed signals is shown in Figure 5. Here, a sinusoidal signal $f_1(t)$ is masked visually in the all-mixed signal, but we can hear the sinusoidal signal from $f_M(t)$ because of the CMR. However, we cannot hear the sinusoidal signal from $f_R(t)$ because of the masking.

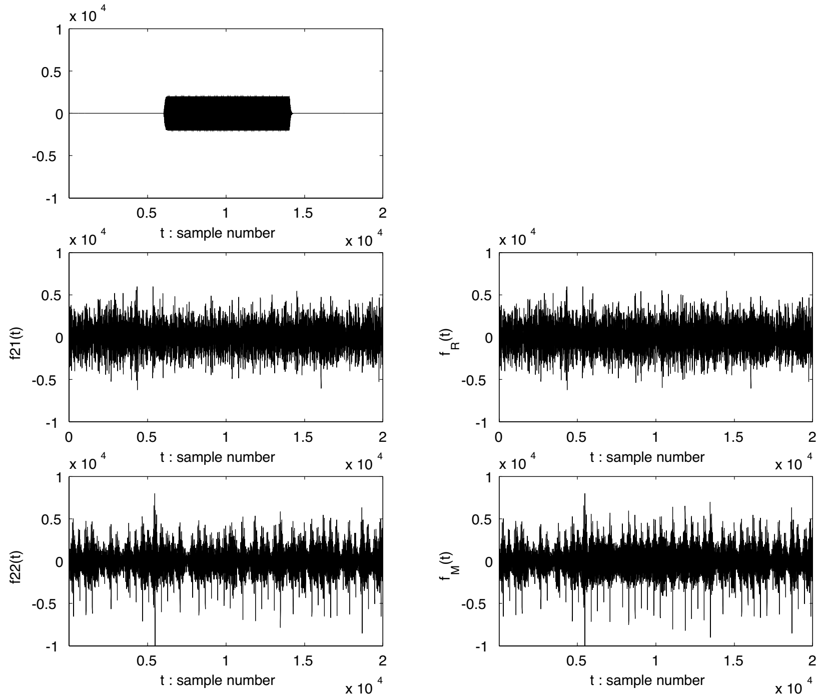


Figure 5 Stimuli: a sinusoidal signal $f_1(t)$ (left-top), a bandpassed random noise $f_{21}(t)$ (left-middle), and an AM bandpassed noise $f_{22}(t)$ (left-bottom). Mixed signals $f_{\mathrm{R}}(t)$ (right-top) and $f_{\mathrm{M}}(t)$ (right-bottom).

In this paper, we set the parameter as $\Delta t = 3/(f_0\alpha^{k-K/2})$, $\Delta A = |A_k(T_r - \Delta t) - A_k(T_r - 2t\Delta)|$, $\Delta B = 0.01S_{\max}$, $\Delta\theta = \pi/20$, and $S_{\max}$ is the maximum of $S_k(t)$. In their demonstration of CMR, Hall *et al* measured the masking threshold as a function of the masker bandwidth.

In their demonstration of CMR, Hall *et al* measured the masking threshold as a function of the masker bandwidth. Our simulation conditions are equivalent since we measured the SNR of the extracted sinusoidal signal $\hat{f}_{1,\mathrm{A}}(t)$ as a function of the number of adjacent auditory filters *L*, which is equivalent to the masker bandwidth, where the masker bandwidth is fixed. Therefore, $\theta_{2k}(t)$ is uniquely determined by the amplitude envelope $\tilde{B}_k(t)$ as a function of *L* from Equations (7), (12), and (13). The bandwidths related to L=1, 3, 5, 7, 9, 11 are 207, 352, 499, 648, 801, 958 Hz, respectively.

5.2.2 Results and Discussion

Simulations were carried out according to the conditions described above. The results are shown in Figure 6, where the vertical and horizontal axes show the improved SNR of the extracted sinusoidal signal $\hat{f}_{1,\mathrm{A}}(t)$ and the bandwidth related to L, respectively. Moreover, the line and the error bars show the mean and standard deviation of the SNR of the signal $\hat{f}_{1,\mathrm{A}}(t)$ extracted from 300 mixed signals, respectively. It was found that for the mixed signal $f_{\mathrm{M}}(t)$, a sinusoidal signal $\hat{f}_{1,\mathrm{A}}(t)$ became detectable as the number of the adjacent auditory filters *L* increased, but for the mixed signal $f_{\mathrm{R}}(t)$, $\hat{f}_{1,\mathrm{A}}(t)$ was not detectable as *L* increased. Therefore, the results show that a sinusoidal signal is more detectable when the

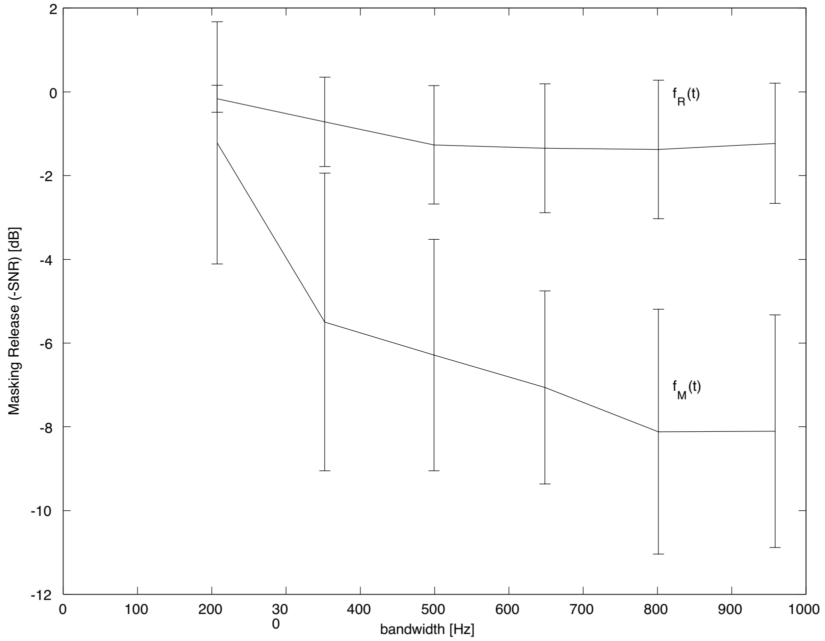


Figure 6 Relationship between the bandwidth related to the number of adjacent auditory filters and the SNR for the extracted signal $\hat{f}_{1,\mathrm{A}}(t)$. The vertical and horizontal axes show the improved SNR of the extracted sinusoidal signal $\hat{f}_{1,\mathrm{A}}(t)$ and the bandwidth related to L, respectively. The real line and the error bars show the mean and standard deviation of the SNR of the signal $\hat{f}_{1,\mathrm{A}}(t)$ extracted from 300 mixed signals, respectively.

components of the masker have the same amplitude modulation pattern in different frequency regions or when the fluctuations in the masker envelopes are coherent. Hence, model A simulates the reduction of masking using the outputs of multiple auditory filters.

5.3 Simulations for Model B

5.3.1 Stimuli and Procedure

These simulations assumed that $f_1(t)$ was the same 10 sinusoidal signals as those used as the stimuli in model A and that $f_2(t)$ was 45 bandpassed random noise maskers of two types formed by varying random seeds (five types) and by varying the bandwidth (nine types). Thus, the total number of stimuli was 450. The masker bandwidths were 33, 67, 133, 207, 352, 499, 648, 801, and 958 Hz because we don't need to determine the number of adjacent filters, L, and we can control the masker bandwidth directly. Three of these bandwidths were related to 1/4, 1/2, and 1 ERB, respectively. The remainder were the same bandwidths used in the simulations for model A.

In model B, in order to measure the masking threshold as a function of the masker bandwidth, we measure the SNR of the extracted sinusoidal signal $\hat{f}_{1,\mathrm{B}}(t)$ from noise-added signal as a function of the masker bandwidth, using the same evaluation measure of masking threshold in model A.

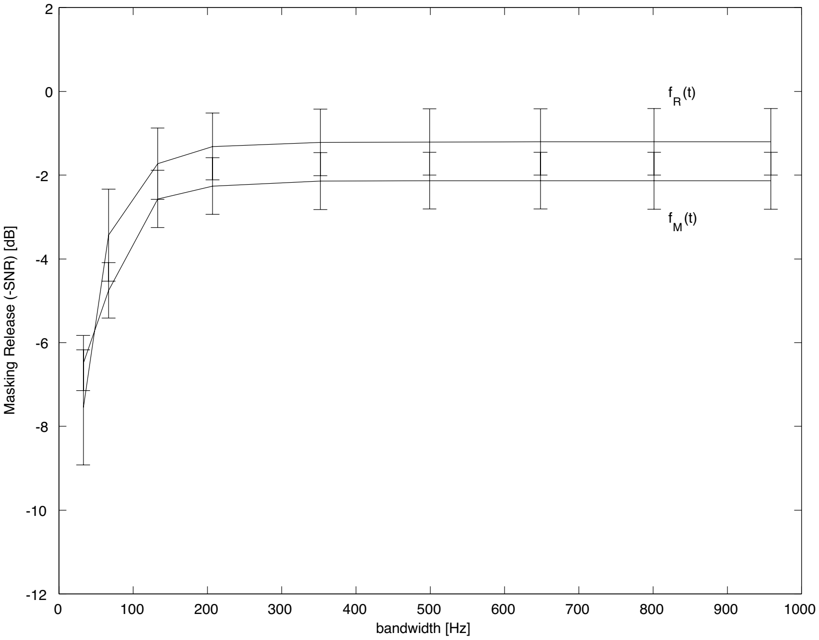


Figure 7 Relationship between the masker bandwidth and the SNR for the extracted signal $\hat{f}_{1,\mathrm{B}}(t)$. The vertical and horizontal axes show the improved SNR of the extracted sinusoidal signal $\hat{f}_{1,\mathrm{B}}(t)$ and the bandwidth related to L, respectively. The real line and the error bars show the mean and standard deviation of the SNR of the signal $\hat{f}_{1,\mathrm{B}}(t)$ extracted from 300 mixed signals, respectively.

5.3.2 Results and Discussion

Simulations were carried out according to the descriptions above. The results are shown in Figure 7, where the vertical and horizontal axes show the improved SNR of the extracted sinusoidal signal $\hat{f}_{1,\mathrm{B}}(t)$ and the masker bandwidth, respectively. Moreover, the line and the error bars show the mean and standard deviation of the SNR, respectively. Figure 7 shows that the SNR for the extracted sinusoidal signal $\hat{f}_{1,\mathrm{B}}(t)$ increased as the masker bandwidth increased, independent on the type of masker. In particular, as the masker bandwidth increased up to 1 ERB the masking threshold (SNR) increased and then remained constant. Hence, model B simulates the phenomenon of simultaneous masking using the output of a single auditory filter.

5.4 Considerations for Computational Model of CMR

The results of simulations for the two models show two types of CMR behavior. Model A simulates the phenomenon of CMR/simultaneous masking by using the coherence of the fluctuations in the amplitude envelope of a masker as the masker bandwidth increases above 1 ERB. By contrast, model B simulates simultaneous masking in which the threshold increases as a function of the masker bandwidth as the masker bandwidth increases up to 1 ERB and then the threshold remains constant. The selection process therefore selects the lowest of these masking thresholds. In other words, it selects the highest SNR of the signal extracted from $\hat{f}_{1,\mathrm{A}}(t)$ and $\hat{f}_{1,\mathrm{B}}(t)$, and then $\hat{f}_1(t)$ is the extracted signal with the highest SNR. Thus, based on the results in Figures 6 and 7, the proposed model has the masking threshold shown in Figure 8. In the selection process, the extracted signal with the lowest

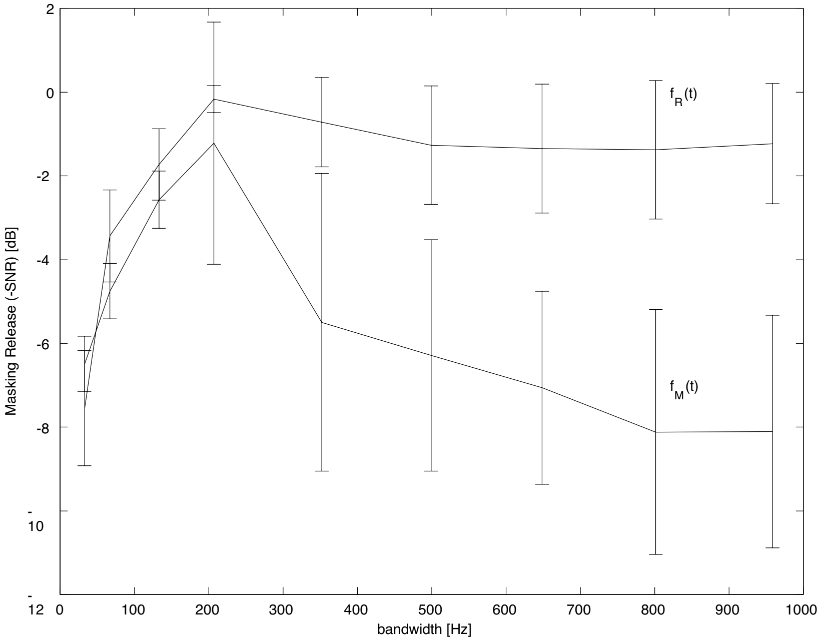


Figure 8 Relationship between the masker bandwidth and the SNR for the extracted signal. This characteristic was obtained from the result of the selection process from Figures 6 and 7.

threshold is selected from the signals extracted using the two models. These characteristics show that the phenomenon of CMR is similar to Hall *et al.*'s results. Hence, the proposed model is a computational model of CMR. The maximum amount of CMR in Hall *et al.*'s demonstrations was about 10 dB, whereas in our model it is about 8 dB.

6. Conclusions

In this paper, we have proposed a computational framework for CMR. This framework consists of two models, our auditory-motivated segregation model (model A) and the power spectrum model of masking (model B), as well as a selection process that selects one of their results. The mechanisms for extracting a sinusoidal signal from a masked signal work as follows: model A uses the outputs of multiple auditory filters and model B uses the output of a single auditory filter.

Simulations of the two models were carried out using two types of noise masker, the same as Hall *et al.*'s demonstration conditions: bandpassed random noise and AM bandpassed random noise. In model A, the signal threshold decreased depending on the type of masker and the masker bandwidth. In the case of bandpassed random noise, the signal threshold did not vary as the masker bandwidth increased. In contrast, for AM bandpassed noise, the signal threshold decreased as the masker bandwidth increased. In model B, the signal threshold increased as the masker bandwidth increased up to 1 ERB and then remained constant for both noise maskers. The selection process then selected the highest SNR from the sinusoidal signals extracted from the two models. As a result, the characteristics of the proposed model show that the phenomenon of CMR closely corresponds to Hall *et al.*'s results. The maximum amount of CMR in the proposed model was about 8 dB.

Hence, the proposed model is a computational model of CMR. We also showed that signal slowness and smoothness — related to regularity (ii) — and the same fluctuation pattern in different frequency regions — related to regularity (iv) — are all important cues to explain CMR.

Acknowledgments

This work was supported by Grant-in-Aid for science research from the Ministry of Education (Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists and 10680374) and by CREST (Core Research for Evolutional Science and Technology) of the Japan Science and Technology Corporation (JST).

References

[1] Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[2] Bregman, A. S. "Auditory Scene Analysis: hearing in complex environments." In *Thinking in Sound: The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand (eds.), New York: Oxford University Press, pp. 10–36, 1993.

[3] Hall, J. W. and Fernandes, M. A. "The role of monaural frequency selectivity in binaural analysis." *J. Acoust. Soc. Am.* 76: 435–439, 1984.

[4] Hall, J. W. and Grose, J. H. "Comodulation masking release: Evidence for multiple cues." *J. Acoust. Soc. Am.* 84: 1669–1675, 1988.

[5] Patterson, R. D. and Holdsworth, J. "A functional model of neural activity patterns and auditory images." In *Advances in Speech, Hearing and Language Processing*, Volume 3, London: JAI Press, 1991.

[6] Patterson, R. D. and Moore, B. C. J. "Auditory filters and excitation patterns as representations of frequency resolution." In *Frequency Selectivity in Hearing,* B. C. J. Moore (ed.), London: Academic Press, pp. 123–178, 1986.

[7] Moore, B. C. J. "Comodulation masking release and modulation discrimination interface." In *The Auditory Processing of Speech, from Sound to Words*, M. E. H. Schouten (ed.), New York: Mouton de Gruyter, pp. 167–183, 1992.

[8] Unoki, M and Akagi, M. *Method of Signal Extraction from Noise-Added Signal, Electronics and Communications in Japan, Part 3, Vol. 80, No. 11, 1997.* Translated from *IEICE, Vol. J80-A*, No. 3, pp. 444–453, March and http://www.jaist.ac.jp/~unoki/index-e.html.

[9] van den Brink, W. A. C., Houtgast, T., and Smoorenburg, G. F. "Effectiveness of comodulation masking release." In *The Auditory Processing of Speech, from Sound to Words*, M. E. H. Schouten (ed.), New York: Mouton de Gruyter, 1992.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

NEURAL TIMING NETS FOR AUDITORY COMPUTATION

Peter Cariani

Eaton Peabody Laboratory, Massachusetts Eye and Ear Infirmary
243 Charles St., Boston, MA 02114, USA

1. Introduction: Temporal Coding of Auditory Qualities

Pitch, timbre, and rhythm are basic auditory qualities that are fundamental to the perception of speech, music and environmental sounds. These perceptual qualities have much in common:

(1) they are very precise (subtle discriminations can be made),
(2) they are largely invariant in the face of large changes in stimulus intensity, location in auditory space and background noise levels, and
(3) they are apprehended by a wide variety of animals.

A central goal for auditory physiology has always been to understand the nature of the neural codes, representations and processing architectures that subserve these auditory form-percepts. Many auditory physiologists and psychoacousticians have recognized the pervasive parallels that exist between auditory percepts on one hand and the temporal discharge patterns of auditory neurons on the other. On many levels, the properties of neural representations based on the stimulus-locked character of neural discharge patterns (spike timings, synchronicity, interspike intervals) mirror those common properties listed above.

The strongest candidate neural codes for pitch at the level of the auditory nerve and brainstem are those based on all-order, interspike-interval distributions of populations of auditory neurons ("population-interval distributions"). Historically a diverse array of models and simulations has pointed to the use of interspike-interval information by the auditory system in explaining the various pitches that are heard [33][34][37][39][43][56], as well as the precision with which they can be discriminated [21][43][55]. While a large number of neurophysiological studies of the auditory nerve have examined the interspike-interval correlates of pitch perception and frequency discrimination, it has only been relatively recently that population-interval distributions have been estimated from auditory-nerve data [10][11][45]. In our own investigations [10][11], we found that features of population-interval distributions estimated from observed responses of 50–100 single auditory-nerve fibers of Dial-anesthetized cats closely parallel those of human pitch perception [10][11]. With very few exceptions, the most frequent interval in the auditory nerve at any given time corresponds to the pitch that is heard. Many complex pitch-related phenomena are readily explained in terms of these population-interval distributions: the pitch of the missing fundamental, pitch equivalence, relative phase and level invariance, non-spectral pitch, pitch shift of inharmonic tones and the dominance region.

We have also observed empirically that patterns of major and minor peaks in population-representations resemble those of their respective stimulus autocorrelation functions [9]. In retrospect, it has become apparent that this similarity is a general consequence of the phase-locking of neural discharges. Because phase-locked responses are found in many other sensory systems, such as vision, mechanoreception, and electroreception, this finding has broad implications outside of the auditory system [4][7][47]. To the extent that a receptor system produces neural discharges whose timings are highly correlated with stimulus time structure, distributions of all-order, interspike intervals resemble the stimulus autocorrelation function. In the auditory system, by virtue of the phase-locking abilities of auditory neurons, population-interval distributions provide very general autocorrelation-like representations for stimulus periodicities up to the limits of phase-locking.

Population-interval distributions representations are also capable of representing the timbre of stationary sounds, such as vowel quality [4][9][25][37][40][45]. These timbres are associated with shapes of spectral envelopes, which manifest themselves in autocorrelation functions as patterns of minor peaks (Figure 3). To the degree that each stimulus component produces phase-locked discharges, it contributes its time structure to the population interval distribution. Consequently, in the auditory nerve, different vowels, with different sets of dominant frequency components, produce population-interval distributions with characteristic patterns of short intervals that reflect their respective formant structures. Changes in these population-interval patterns closely follow vowel-identification boundaries [25].

Population-interval distributions thus appear to be capable of subserving a wide variety of auditory qualities associated with pitch and timbre. These strong psycho-neural correspondences beget questions of whether the central auditory system, in fact, utilizes this interval-based information, and if it does, how does it use it. Related to these questions are still others that concern the fate of neural timing information as one ascends the auditory pathway. Is the neural timing information that is so precise and robust, and in such abundance at the level of the auditory nerve, converted to across-neuron patterns of activation in higher, central auditory stations? Or is the temporal structure preserved in some way, perhaps in less synchronous and more spatially distributed form than is found in lower stations? If temporal information is in fact available in central auditory stations at the level of the midbrain, thalamus, and/or cortex, what kinds of neural processing architectures would be needed to make use of it?

This paper explores some possible means by which neural networks might analyze distributed, population-based temporal representations of auditory qualities. For the most part we will put aside for the present questions of where these neural networks might be concretely located, in favor of more functionally oriented ones devoted to exploring their potential information-processing capabilities. Whether these kinds of neural computations are in fact carried out in central auditory structures are empirical questions that can only be answered through directed neurophysiological experiments. While a detailed understanding of how the auditory portion of the brain works as an information-processing system remain our ultimate goal, we can only direct our neurophysiological lenses effectively if we already have some strong ideas about the kinds of neural computational mechanisms that might be possible.

2. Time-to-Place Conversions

In the past virtually all of the temporal theories of hearing have assumed that the temporal information found at the level of the auditory nerve is converted to spatial patterns of activation somewhere higher in the auditory pathway. Many of the first neural networks that were proposed for auditory computation, such as the Jeffress model for auditory localization [27] and the Licklider duplex model for pitch perception [33], were time-delay neural networks whose purpose was to carry out this conversion. It was generally assumed that the outputs of such networks would then be analyzed via traditional, channel-coded connectionist networks in more central stations. For example, Licklider's time-delay architectures [33][34] converted temporal input patterns to spatialized autocorrelation profiles by means of delay lines and coincidence detectors. However, sharply tuned autocorrelator-like periodicity detectors have yet to be found in the auditory pathway. Likewise, neurophysiological investigations in the auditory cortex have failed to find other kinds of simple pitch-detection units [54]. The most promising evidence for a time-to-place transformation has involved the modulation-tuning properties of central auditory neurons [32][53]. However, modulation-tuning tends to be relatively coarse, and to weaken at higher levels and in background noise [48][49]. Moreover, as one ascends the auditory pathway to auditory midbrain, thalamus, and cortex, best modulation frequencies (BMFs) generally decline, with progressively fewer BMFs covering the periodicity pitch range (50-500 Hz). This shift towards lower BMFs parallels declines in average discharge rates and synchronization indices that are seen. Finally, modulation-based representations, like first-order interval detectors, sometimes diverge from the autocorrelation-like behavior that characterizes pitch judgments (e.g. de Boer's rule for pitch shifts of inharmonic AM tones).

3. Neural Timing Networks: Time-Time Comparisons

A second possible strategy for representing and analyzing auditory forms is to retain temporal information in one form or another, and to perform comparisons between different time patterns by observing their interactions. For example, one can detect extremely subtle differences in frequency by binaural comparisons in which one listens for the presence of binaural beats. A major question for such an approach concerns the availability of temporal information to be analyzed. Unfortunately, the existence limits of neural timing information in the auditory pathway are still not well established. Pitch-related temporal patterns are omnipresent in the auditory nerve and cochlear nucleus [5][51] and are still quite evident in the auditory midbrain [22][32]. Although neural interspike interval information present in single units thins out dramatically as one proceeds from brainstem to thalamus to cortex, it is nevertheless possible that the requisite timing information to support central time codes for pitch and timbre exists in thalamocortical loops. Roughly half of all units encountered in lightly anesthetized auditory thalamus show significant phase-locking (synchronization index > 0.3) to pure tones of 250-500 Hz, while roughly 10% phase-lock to 1-2 kHz tones [20]. Response periodicities of several hundred Hz are observed in unanesthetized primary auditory cortex [20][57]. To the extent that interspike-interval information exists in many of these stations, it remains precise, robust and faithful to the autocorrelation-like behavior of pitch. It is important to remember that the timing information present in the auditory nerve far exceeds that required for human frequency discrimination [21][55]. Accordingly, only a small fraction of the timing information available at the auditory nerve need be faithfully transmitted and preserved for central auditory analyzers in order to realize the perceptual capabilities that are observed for the organism as a whole.

If the interval-based information is indeed available in central auditory stations, what kinds of neural networks are required for its analysis? Alongside traditional connectionist networks and time-delay networks, neural timing networks can be envisioned that operate on time structure in their inputs to produce interpretable temporal patterns in their outputs (time-to-time mappings). Their closest precursors are simple functional models of neural computation for which fine time structure is of primary importance [1][3][12][27][34][35][36][38][46][50][58]. Some of these precursors were themselves inspired by the functional anatomy of cortical structures [3][50][58].

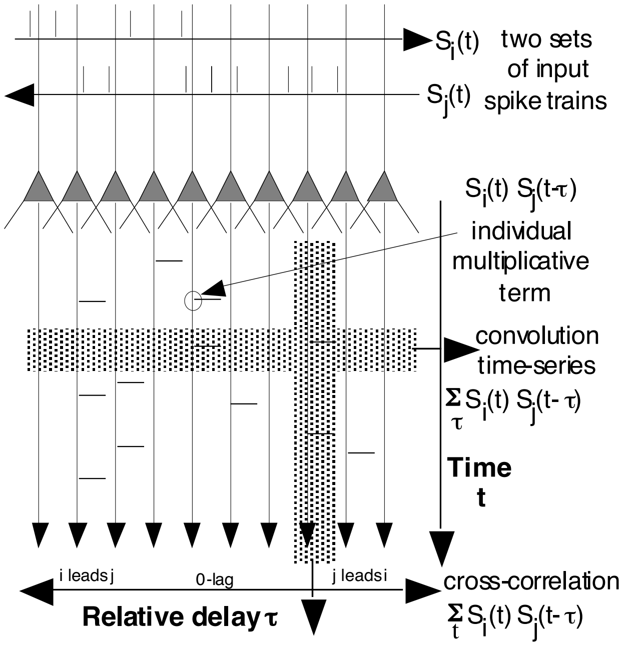

$S_i(t)$
two sets of input spike trains
$S_j(t)$
$S_i(t)\ S_j(t-\tau)$
individual multiplicative term
convolution time-series
$\Sigma_\tau S_i(t)\ S_j(t-\tau)$
Time t
i leads j
0-lag
j leads i
Relative delay τ
cross-correlation
$\Sigma_t S_i(t)\ S_j(t-\tau)$

Figure 1 Simple feedforward timing net consisting of an array of coincidence detectors and two sets of tapped delay lines through which input signals S_i and S_j arrive.

4. Simple Feedforward Timing Nets

Consider an array of coincidence detectors that have inputs from two sets of tapped delay lines arranged in anti-parallel orientation (Figure 1). The configuration is reminiscent of both the Jeffress binaural localization model [27] and the Braitenberg cerebellar timing model [3]. Many relative delays are realized by the slow conduction times across the array such that each position along the tapped delay line corresponds to a particular relative delay between the input signals. Thus, all relative delays are realized up to the conduction time across the array. Each coincidence detector requires nearly simultaneous arrival of a spike in both lines in order to fire. Consequently, each spike in the output of the coincidence array represents the joint occurrence of spike arrivals in the two inputs (or the multiplication of binary inputs, $S_i(t)*S_j(t-\tau)$). A further consequence is that each interspike interval or higher-order spike arrival pattern appearing in a given output channel must also be present in each of the two inputs. Thus the array functions as a temporal sieve, passing those temporal patterns that are common to both sets of inputs. Several basic computations can be carried out. First, the cross-correlation function of the two inputs can be computed by counting the number of spikes in each output channel as a function of relative delay. Their convolution can be computed by summing across relative delay channels for each time step. Similarly, the summary or population-autocorrelation of the outputs can be computed by summing the autocorrelations of each of the output channels.

The conduction time across the array implements a temporal contiguity window; those inputs that arrive within this time window interact, while those arriving at different times do not. All intervals from each set of inputs that arrive within the temporal contiguity window cross their counterparts, such that if one input has M intervals of duration, τ_0, and the other has N such intervals, then $M*N\ \tau_0$ intervals will appear in the outputs. Within the temporal

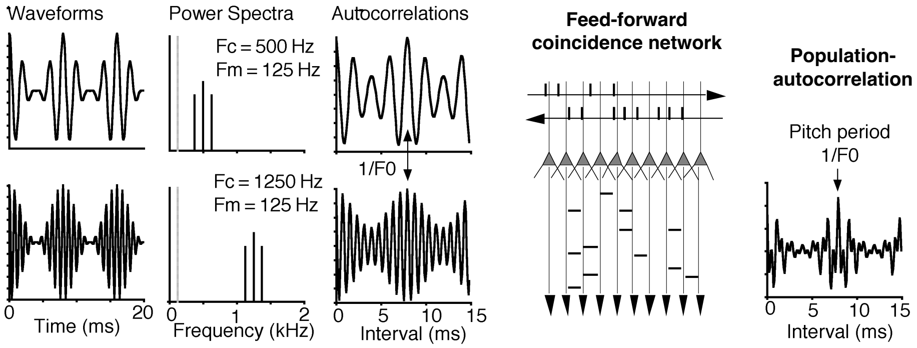


Figure 2 Effect of passing two signals through the coincidence array. The stimuli are two AM tones with different carriers (f_c = 500 Hz, 1250 Hz) but the same modulation frequency (f_m = 125 Hz). The AM tones have no harmonics in common, but they produce a common low pitch at their "missing fundamental ($f_0 = f_m$ = 125 Hz, dotted lines). Right: Population autocorrelation of the output of the coincidence array.

contiguity constraints, the coincidence array therefore performs a multiplication of the autocorrelations of its inputs.

The population-autocorrelation output of such a coincidence array is largely phase-insensitive. Because all of the intervals in the two input lines arriving within the time window cross their counterparts somewhere in the array, the short-term temporal ordering of the intervals within each incoming pulse train signal has little effect on the population-autocorrelation of the output. This behavior is qualitatively similar to the phase-insensitive character of auditory form perception: in general, we have great difficulty distinguishing pitches or timbres of complex tones that differ only in their phase spectra. Temporal contiguity constraints also exist in pitch and timbre perception. Pitches associated with the missing fundamental can be evoked for sets of harmonics that are presented successively, but disappear when brief periods (> 10 ms) of silence are inserted between them [24]. Similarly, two single-formant vowels do not produce a two-formant vowel quality unless the waveforms corresponding to the two formants arrive within a similarly brief time window [13][14]. Provided that their waveforms overlap in time within this window and have the same fundamental, one cannot generally distinguish between combinations of single formant vowels with different relative delays among the vowels. The phase-insensitive nature of this coincidence array means that the mechanism can accommodate a good deal of asynchronous, temporal shifting among its inputs.

5. Recognition of Common Pitch Irrespective of Timbre

Coincidence arrays can extract those periodicities common to their inputs, even if their inputs have no harmonics in common. This is useful for the recognition of common pitches irrespective of differences in timbre (e.g. two different musical instruments playing the same note). As an example, two amplitude-modulated (AM) tones were passed through the coincidence array (Figure 2). The fundamental frequency (f_0) of an AM tone is equal to its modulation frequency (f_m). AM tones produce strong pitches at their fundamental frequencies, despite the lack of any stimulus energy at that frequency (i.e., AM tones produce pitches at various "missing fundamentals"). For this example, the fundamental frequencies of the two signals were both set to 125 Hz, such that the signals produce the same low pitch at that fre-

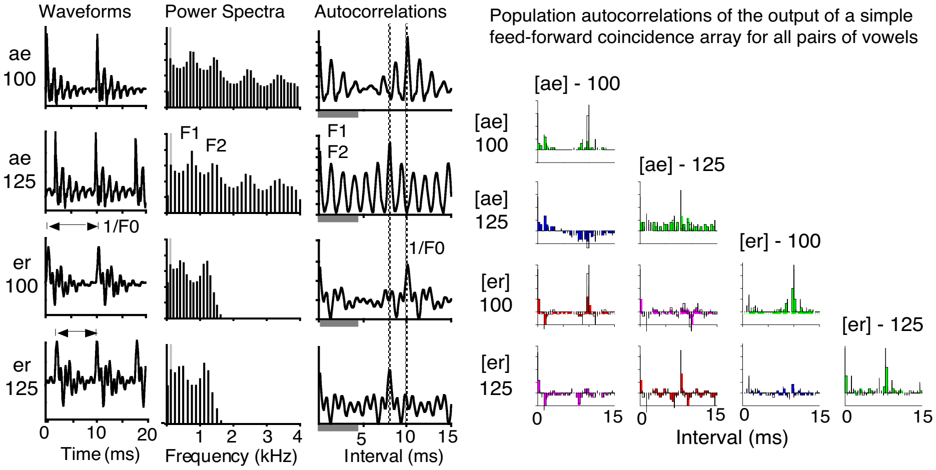


Figure 3 Left: Waveforms, power spectra, and autocorrelation functions for four vowels. The vowel set consists of combinations of two different fundamental frequencies (f_0 = 100, 125 Hz) and two formant structures. Horizontal arrows above waveforms and vertical lines in autocorrelations indicate fundamental periods ($1/f_0$ = 8, 10 ms), which correspond to voice pitch periods. Shaded bars indicate periodicities associated with formant structures that give rise to differences in vowel quality (timbre). Right: Population autocorrelations of the output of the coincidence array for all vowel pairs.

quency. Despite their common fundamental, the two signals have different carrier frequencies (f_c = 500 Hz vs. 1250 Hz) and therefore have different spectral energy distributions. Such signals would produce different timbres. When the two signals are passed through the array, the resulting population autocorrelation is dominated by intervals at the common fundamental period, $1/f_0 = 8$ ms. The array thus extracts those periodicities that are common to the two signals, and the form of those common temporal patterns appears directly in its output.

6. Recognition of Common Timbre Irrespective of Pitch

Coincidence nets can also extract common periodicities that are associated with different timbres or vowel qualities. This is useful for recognizing common timbres irrespective of differences in pitch (e.g. the same musical instrument playing different notes, or two different people speaking the same vowel). Four synthetic vowels consisting of combinations of two fundamental frequencies (f_0s) and two sets of formants (F_1, F_2, F_3, F_4, F_5) were constructed (Figure 3). These signals correspond to the vowels [ae] (as in "hat") and [ε^r] (as in "herd"). Their waveforms, power spectra, and autocorrelation functions are shown in Figure 3 (left). Each vowel evokes a "voice pitch" at its fundamental. Fundamental frequencies (f_0) correspond to spacings between adjacent harmonics in the power spectra; fundamental periods ($1/f_0$) correspond to major peaks in the respective autocorrelation functions. Each vowel also has a characteristic tonal quality ("timbre") which determines whether it will be recognized as an [ae] or an [ε^r] (or some other vowel). The general shape of the power spectrum (spectral envelope) largely determines the timbre of a stationary sound; the spectral envelope, in turn, is largely shaped by positions and magnitudes of spectral peaks (formants). Different combinations of formants produce characteristic patterns of short time intervals in the autocorrelation functions. Similar characteristic patterns corresponding to the fundamental and to formant combinations are observed in population-interval distributions at the level of the auditory nerve [4][8][9][37][45].

All combinations of the four waveforms were passed pairwise through the coincidence net (Figure 3, right panel). Population autocorrelations produced by vowels paired with themselves are equivalent to their own autocorrelations squared. Those vowel pairs that had common fundamental frequencies and similar voice pitches produced large peaks at their common fundamental periods. Those vowel pairs that had common formant structures (common vowel quality or timbre) produced common patterns of short intervals that correspond to their respective formant structures. Those vowel pairs that had neither common fundamental frequency nor common formant structure (different voice pitches and timbres) produced only small peaks associated with overlapping subharmonics.

Thus, a simple, feedforward coincidence array can operate on two sets of temporally coded inputs in order to extract common periodicities underlying common pitches and timbres. This permits a common pitch to be recognized independent of timbre, and a common timbre to be recognized independent of pitch. Further, both operations can be realized using the same, simple mechanism that operates on the interspike-interval statistics of an entire ensemble of neural elements.

7. Binaurally Created Pitches

The feedforward coincidence operations outlined above require the two sets of inputs to be simultaneously present in the network in order to effect pitch and timbral comparisons. The most obvious locations in the auditory system where one has simultaneous phase-locked inputs, tapped delay lines, and arrays of coincidence detectors are structures in the auditory brainstem that receive binaural inputs. Low pitches and rhythmic binaural beats can be created by binaural interactions within these structures [15]. Historically, the existence of “binaurally created pitches” was used to argue against temporal models for pitch that required interaction of neighboring harmonics within the same cochlea (e.g. Schouten’s theory of ‘residue’ pitch [18]). Houtsma and Goldstein [26] showed that binaural combination of two harmonics of a common fundamental could give rise to a binaurally created pitch at the missing fundamental. The existence of these pitches was explained in terms of a spectral-pattern analysis of harmonic structure in a “central spectrum” representation. The feedforward operation outlined above provides a temporal account of the generation of such pitches. Here the two sets of inputs to the coincidence array come from the auditory pathways originating in each ear. As with the two AM tones illustrated above, when two harmonically related pure tones are passed through such a coincidence network, the population autocorrelation function of the output produces a maximum at their common fundamental period. A similar result is obtained if the two harmonics are band-passed filtered, half-wave rectified, and the output of each channel is passed through a similar cross-correlation array [6]. The time-structure of the respective tones are impressed on swaths of frequency channels that overlap and these beat at the fundamental frequency. In those channels, binaural coincidence detectors consequently produce many intervals at the “missing” fundamental period. According to a general temporal autocorrelation theory of pitch, such a population-interval pattern would then be interpreted by central analyzers, much in the same way as monaural pitches, with the result that a binaural interaction pitch at the missing fundamental should be heard. These observations notwithstanding, there are other temporal mechanisms, such as a simple central addition of the monaural population-interval distributions, that would also produce these pitches.

There are also other kinds of pitches that are created through binaural phase differences [2][16] that create troughs in the population autocorrelations of binaural cross-correlation arrays. These troughs correspond to the pitch periods that are heard. Such pitches therefore require cancellation or anti-correlation operations rather than simple coincidence operations [17][19]. Such operations could be incorporated into feed-forward timing nets by adding anti-coincidence detectors that produce output pulses when there is an incoming pulse in only one of the two input lines (an XOR operation). Once both coincidence and anti-coincidences are computed, timing networks attain the means of computing both temporal similarities and differences present in their inputs.

8. Simple, Recurrent Timing Nets

The simple feed-forward networks outlined above carry out comparisons between inputs that are simultaneously presented. In order to perform delayed matching tasks, such networks would require some mechanism for maintaining a working memory representation of what came before. Perhaps the simplest means of storing time patterns, either in the form of post-stimulus-time patterns or interval statistics, is to let the signals themselves circulate in recurrent sets of delay lines (Figure 4). A reverberating memory is thereby created in which the signal itself serves as its own temporal memory trace. Incoming time patterns can then be compared with those that are circulating using the kinds of feed-forward correlational operations outlined above. Matching of pitches or timbres in such a system then involves maximizing the correlation between the stored temporal pattern and the incoming one.

In such a system, recognition operations can be carried out if there are central neural assemblies that can produce temporal patterns that are characteristic of the objects to be recognized (e.g., interval distributions characteristic of particular vowels). Neural responses consistent with this notion have been observed in some neurophysiological conditioning studies [29][44][58], where stimulus-related temporal patterns are "assimilated" by individual neurons and "readout" at different times. If the outputs of an ensemble of such assemblies are cross-correlated with incoming temporal patterns and fed back into the loop, then those incoming patterns that resonate most strongly with those produced by neural assemblies will build-up the fastest. Strongly activated central temporal pattern templates can steer the build up of circulating patterns, such that the resulting resonances resemble the intersection of the incoming pattern with the stored templates, thereby creating "perceptual magnet effects."

A considerable body of psychological evidence exists for mechanisms that build-up, store, and read-out temporal expectations. Studies of conditioning [28][42][58], music perception [30][31], and rhythm production [52] suggest that temporal relationships are explicitly encoded in memory, and that these relationships create sets of temporal expectancies. Recurrent timing nets implement reverberating memories [58] that can dynamically create short-term expectancies and build up temporal patterns that recur over time.

Perhaps the simplest example of a reverberating memory is the recurrent timing net shown in Figure 5. This network cross-correlates incoming time patterns with previous, circulating ones in order to build up those temporal patterns that recur. The network consists of an array of coincidence detectors which all receive the same external signals. Each coincidence detector has an associated delay loop with a different recurrence time. Coincidence

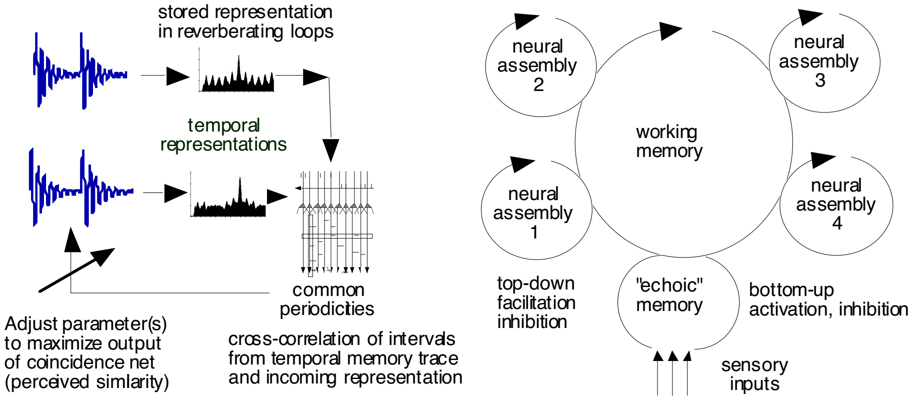


Figure 4 Temporal memory traces, matching tasks, and the build-up of perceptual forms

detectors temporally cross-correlate incoming signals with those that are arriving via the delay loop. As a first step, pulse trains with repeated, randomly selected pulse patterns (e.g. 100101011-100101011-100101011...) are presented to the network. At each time step, the incoming pulse train is multiplied by the circulating pulse train arriving in each respective delay loop and the resultant signal is fed back into the loop. In the absence of pulses arriving through the delay loop, the incoming pulse train is fed into the loop. If there are coincident pulses from both inputs, the amplitude of the output pulses that are propagated through the loop is increased by 5%. Coincidences, therefore, build up the strength of the circulating pattern.

In such a network, periodic pulse patterns invariably build up fastest in the delay loop whose recurrence time matches their repetition time. In their respective loops, rhythmic input patterns create temporal expectancies (when pulses traveling through the loop arrive back at the coincidence detector that generated them) that are reinforced when they are satisfied. Thus, recurrent time patterns are repeatedly correlated with themselves to build up to detection thresholds. In effect, the recurrent cross-correlation loops dynamically create matched filters from repeating temporal patterns in the stimulus. Thus, temporal-pattern invariances are enhanced relative to aperiodic transient activity, such as noise. Similar strategies for periodicity detection were explored in the 1950s [41].

More elaborate recurrent timing nets would also incorporate anticoincidence elements that compute the difference between expectation and the incoming signal. Once both correlation and anti-correlation operations are in place, these networks begin to resemble simplified, time-domain versions of adaptive resonance networks [23]. In place of spatialized input patterns and spatial pattern correlation operations for comparing them, timing nets utilize temporal input patterns, delay lines and coincidence detectors to do the comparisons in the time domain. Temporal correlation and anti-correlation take the place of excitation and inhibition. Both kinds of networks utilize recurrent bottom-up, top-down interactions to build up resonant patterns of activity. When inputs confirm top-down expectations, those expectations are reinforced; when inputs diverge from expectations, their differences form new expectation patterns that can then subsequently be built up.

These simple recurrent timing networks can also separate multiple time patterns with different repetition periods. When two repeating pulse patterns, each with its own repetition period, are summed and presented to the network, the two patterns invariably build up in the two different delay paths that have the corresponding recurrence times. These recurrent timing architectures were inspired by rhythm perception and production (e.g., [30][31][52]),

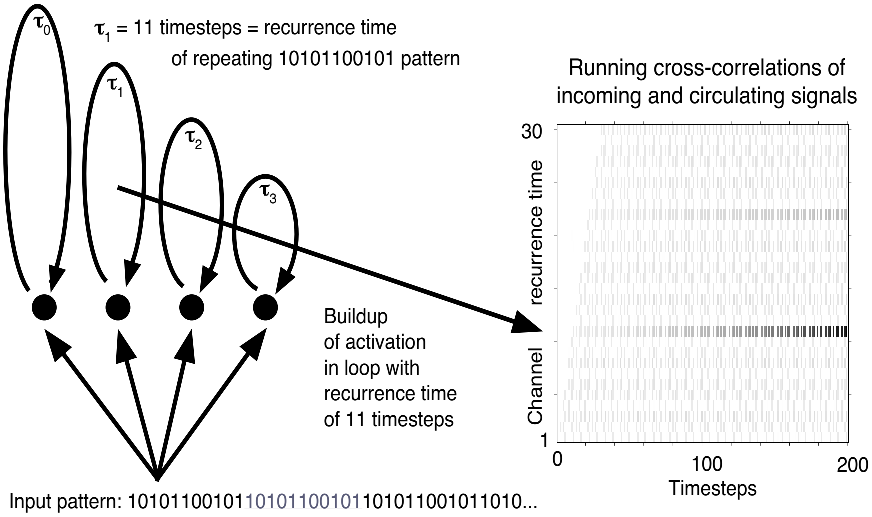

τ_0
τ_1 = 11 timesteps = recurrence time
of repeating 10101100101 pattern
τ_1
τ_2
τ_3
Buildup
of activation
in loop with
recurrence time
of 11 timesteps
Input pattern: 1010110010110101100101101011001011010...
Running cross-correlations of
incoming and circulating signals
30
Channel recurrence time
1
0
100
200
Timesteps

Figure 5 Behavior of a simple recurrent timing net for periodic pulse-train patterns.

and phase-sensitive processes in auditory temporal integration [46]. While they were conceived to operate over longer time windows associated with these phenomena (> 30 ms), many parallels exist between rhythm and pitch, such that these general processing strategies appear to be potentially applicable to pitch-related separations as well.

Two vowels with different fundamental frequencies (f_0 = 100 Hz, 125 Hz) were summed together and presented to the recurrent network (Figure 6). Each period of the two vowels has its own invariant waveform pattern. The internal relations within the vowel periods of each waveform remain constant from period to period, whereas the relation between the two vowel-period waveforms change over time - the vowel periods precess relative to each other, creating “pitch period asynchronies.” Similar precessions and perceptual separations occur when an individual frequency component of a harmonic complex is mistuned. As with pairs of repeating pulse patterns, the two vowels build up their respective waveform patterns in the corresponding delay loops. (A potential problem with this multiplicative [vs. additive] buildup is that successive multiplications alter relative amplitudes of waveform peaks, although zero-crossings remain intact.) Thus, multiple auditory objects with different repetition periods (i.e. fundamentals, rhythms) can be segregated into different delay paths. Fusion is the consequence of recurrent, invariant temporal relations, while segregation is the consequence of changing temporal relations (precession of vocalic periods relative to each other).

Segregation by temporal pattern invariance constitutes an extremely general strategy for the formation and separation of perceptual objects. Traditional strategies for scene analysis are based on channel selection. First, a local feature analysis is carried out on incoming sensory patterns and an attempt is made to select subsets of feature channels that should be grouped together or separated to form different objects. For concurrent vowels, this has meant detecting which frequency channels share common f_0-related modulations and grouping them together (e.g. [40]). The correlational strategy proposed here instead groups patterns of spikes rather than patterns of channels. Here no explicit feature detection is required prior to the formation of auditory objects — the temporal patterns build themselves up and sort themselves out in their respective delay channels.

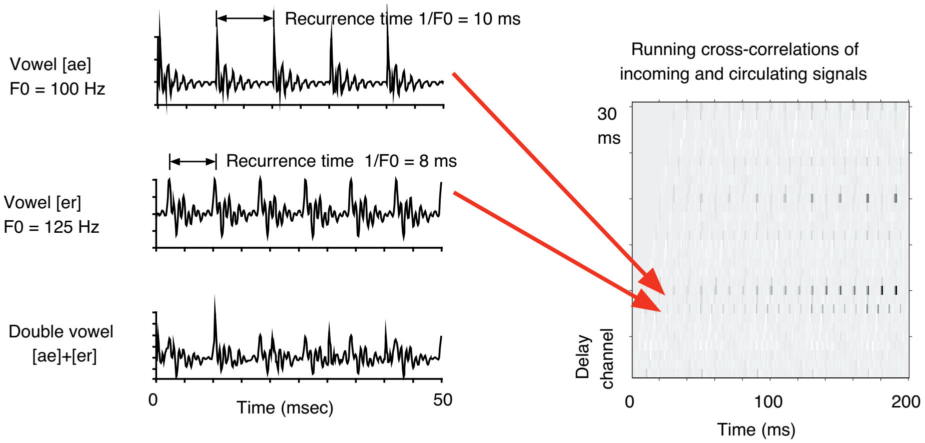


Figure 6 Separation of two auditory objects, with differential fundamental frequencies, in a simple recurrent net.

Both feed-forward and recurrent timing networks share a number of general functional properties that are highly desirable in the context of neural computation in the brain:

1) no highly tuned delay lines, periodicity detectors, or clocks are needed because no explicit time measurements are made,
2) representational precision resides in spike timings instead of in neural activation profiles,
3) harmonic relations implicit in time intervals are preserved (e.g. octave similarities, characteristic musical interval patterns), and
4) population-wide operations that make use of all neural responses, even weak ones, obviate the need to select relevant subpopulations for analysis.

Population-based temporal representations permit information from whole neural populations to be exported *en masse* to other regions. Coincidence networks permit comparisons between activity patterns of neural populations without the necessity of precise point-to-point mappings between them and/or highly regulated synaptic weightings. These properties may greatly simplify the coordination of information processing in large numbers of semi-independent, largely asynchronous populations of neurons.

How such computational strategies might scale up for large numbers of inputs, delay paths and coincidence elements remains to be explored. Simultaneous arrival of incoming pulses in three sets of inputs as a requirement for coincidence leads to higher-order, triple-correlation functions [59] that carry temporal sequence and phase information. Recurrent delay loops can be implemented by multisynaptic pathways, provided that the build-up of jitter can be constrained through general connectivity rules (e.g., fan-in/fan-out factors) or through adjustments of specific connectivities and time delays. If jitter builds up with the average number of synapses traversed and this is, in turn, roughly proportional to the time delay needed to encode a particular duration, then one has a potential explanation for the constant Weber fractions that are observed in discriminations of rhythms and other time intervals [52]. A theory of timing relations in arbitrary conduction networks would clearly be helpful.

This present treatment of timing networks barely ventures beyond an outline of the idea and what kinds of operations might potentially be carried out. Certainly, inhibitory inputs and anticoincidence operations need to be incorporated into such networks, and feedforward and recurrent architectures need to be combined. Once these primitive networks are developed more fully and their behavior understood more deeply, then more realistic psychoneural models can be entertained that point to empirically testable hypotheses that address the real workings of the brain.

9. Conclusions

In the auditory nerve there is an abundance of temporal information that precisely and robustly encodes many perceptually relevant aspects of acoustic stimuli: periodicity, spectral shape, speech modulations, rhythms, and still longer time patterns. Most central models of auditory processing that utilize this timing information have assumed that a time-to-place transformation must occur in the ascending auditory pathway, such that central representations of auditory forms are based on excitation profiles of frequency- or periodicity-tuned units. In these models auditory discrimination and recognition is performed by comparing stored excitation profiles with incoming ones.

However, if neural mechanisms exist by which timing information can be preserved and stored centrally, then purely temporal analyses of similarity and difference can be carried out by temporal-correlation operations. We have outlined two basic processing architectures that could realize such operations. A simple, feedforward neural timing architecture has been presented that utilizes coincidence detectors and tapped delay lines to perform cross-correlation and/or convolution operations on two sets of inputs. Only those periodicities that are common to both inputs appear in the time structure of the outputs. The array functions as a temporal sieve whose summary autocorrelation function is the product of the autocorrelations of its inputs. To the extent that time structure of inputs reflect those of stimuli, such arrays can compute pitch similarity irrespective of timbre and timbral similarity independent of pitch. A simple recurrent timing architecture consisting of an array of many different delay loops is presented that amplifies and separates recurring time patterns.

These purely temporal modes of analysis are carried out on population-wide bases that obviate the need for precise point-to-point connectivities, explicit measurement of local features and/or internal clocks. Timing nets constitute a new and general neural network strategy for performing a host of basic auditory computations: extraction of common periodicities, detection of recurrent time patterns and separation of auditory objects. While the examples considered here are very rudimentary, they nevertheless afford glimpses of the kinds of perceptual computations that might be realized using temporal codes and timing nets.

Acknowledgments

This work was supported by Research Grant DC03054 from the National Institute on Deafness and Other Communicative Disorders, U.S. National Institutes of Health.

References

[1] Abeles, M. *Corticonics*, Cambridge: Cambridge University Press, 1990.

[2] Bilsen, F. A. "Pronounced binaural pitch phenomenon." *J. Acoust. Soc. Am.,* 59: 467–468, 1976.

[3] Braitenberg, V. "Functional interpretation of cerebellar histology." *Nature,* 190: 539–540, 1961.

[4] Cariani, P. "As if time really mattered: temporal strategies for neural coding of sensory information." In *Communication and Cognition — Artificial Intelligence (CC-AI),* 12: 161–229, 1995.

[5] Cariani, P. "Physiological correlates of periodicity pitch in the cochlear nucleus [abstract]." *Assoc. Res. Otolaryn. Abstr.,* 128, 1995.

[6] Cariani, P. "Population-interval models for binaural periodicity pitches." *Soc. Neurosci. Abstr.,* 22: 649, 1996.

[7] Cariani, P. "Temporal coding of sensory information." In *Computational Neuroscience: Trends in Research, 1997*, J. M. Bower (ed.), New York: Plenum, pp. 591–598, 1997.

[8] Cariani, P. and Delgutte, B. "Interspike interval distributions of auditory nerve fibers in response to concurrent vowels with same and different fundamental frequencies [abstract]." *Assoc. Res. Otolaryngology. Abs.,* 373, 1993.

[9] Cariani, P., Delgutte, B. and Tramo, M. "Neural representation of pitch through autocorrelation." *Proc. Audio Eng. Soc.* (Preprint #4583, L-3), 1997.

[10] Cariani, P. A. and Delgutte, B. "Neural correlates of the pitch of complex tones. I. Pitch and pitch salience." *J. Neurophysiol.,* 76: 1698–1716, 1996.

[11] Cariani, P. A. and Delgutte, B. "Neural correlates of the pitch of complex tones. II. Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch." *J. Neurophysiol.,* 76: 1717–1734, 1996.

[12] Cherry, C. "Two ears — but one world." In *Sensory Communication*., W. A. Rosenblith (ed.), New York, MIT Press/John Wiley, pp. 99–117, 1961.

[13] Chistovich, L. A. "Central auditory processing of peripheral vowel spectra." *J. Acoust. Soc. Am.,* 77: 789–805, 1985.

[14] Chistovich, L. A. and Malinnikova, T. G. "Processing and accumulation of spectrum shape information over the vowel duration." *Speech Communication,* 3: 361–370, 1984.

[15] Colburn, S. and Durlach, N. I. "Models of binaural interaction." In *Handbook of Perception*., E. C. Carterette and M. P. Friedman (eds.), New York: Academic Press, pp. 467–518, 1978.

[16] Cramer, E. M. and Huggins, W. H. "Creation of pitch through binaural interaction." *J. Acoust. Soc. Am.,* 30: 413–417, 1958.

[17] Culling, J. F., Summerfield, Q. and Marshall, D. H. "Dichotic pitches as illusions of binaural masking release I: Huggins' pitch and the binaural edge pitch." *J. Acoust. Soc. Am.,* 103: 3509–3526, 1998.

[18] de Boer, E. "On the "residue" and auditory pitch perception." In *Handbook of Sensory Physiology*., W. D. Keidel and W. D. Neff (eds.), Berlin: Springer-Verlag, pp. 479–583, 1976.

[19] de Cheveigné, A. "Cancellation model of pitch perception." *J. Acoust. Soc. Am.,* 103: 1261–1271, 1998.

[20] de Ribaupierre, F. "Acoustical information processing in the auditory thalamus and cerebral cortex." In *The Central Auditory System*., G. Ehret and R. Romand (eds.), New York: Oxford University Press, pp. 317–397, 1997.

[21] Goldstein, J. L. and Srulovicz, P. "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement." In *Psychophysics and Physiology of Hearing*., E. F. Evans and J. P. Wilson (eds.), London, Academic Press, pp. 337–346, 1977.

[22] Greenberg, S. *Neural Temporal Coding of Pitch and Vowel Quality: Human Frequency-Following Response Studies of Complex Signals*. UCLA Working Papers in Phonetics #52, 1980.

[23] Grossberg, S. "Neural dynamics of motion perception, recognition learning, and spatial attention." In *Mind as Motion: Explorations in the Dynamics of Cognition*., R. F. Port and T. van Gelder (eds.), Cambridge: MIT Press, pp. 449–490, 1995.

[24] Hall III, J. W. and Peters, R. W. "Pitch for nonsimultaneous successive harmonics in quiet and noise." *J. Acoust. Soc. Am.,* 69: 509–513, 1981.

[25] Hirahara, T., Cariani, P. and Delgutte, B. "Representation of low-frequency vowel formants in the auditory nerve." *Proc. ESCA Research Tutorial on The Auditory Basis of Speech Perception,* S. Greenberg and W. A. Ainsworth (eds.), Keele University, U. K., pp. 83–86, 1996.

[26] Houtsma, A. J. M. and Goldstein, J. L. "The central origin of the pitch of complex tones: Evidence from musical interval recognition." *J. Acoust. Soc. Am.,* 51: 520–529, 1972.

[27] Jeffress, L. A. "A place theory of sound localization." *J. Comp. Physiol. Psychol.,* 41: 35–39, 1948.

[28] John, E. R. "Switchboard vs. statistical theories of learning and memory." *Science,* 177: 850–864, 1972.

[29] John, E. R. and Schwartz, E. L. "The neurophysiology of information processing and cognition." *Ann. Rev. Psychol.,* 29: 1–29, 1978.

[30] Jones, M. R. "Time, our lost dimension: Toward a new theory of perception, attention, and memory." *Psych. Rev.,* 83: 323–255, 1976.

[31] Jones, M. R. and Yee, W. "Attending to auditory events: The role of temporal organization." In *Thinking in Sound: The Cognitive Psychology of Human Audition*., S. McAdams and E. Bigand (eds.), Oxford: Clarendon Press, pp. 69–112, 1993.

[32] Langner, G. "Periodicity coding in the auditory system." *Hear. Res.,* 60: 115–142, 1992.

[33] Licklider, J. C. R. "A duplex theory of pitch perception." *Experientia,* 7: 128–134, 1951.

[34] Licklider, J. C. R. "Three auditory theories." In *Psychology: A Study of a Science*, S. Koch (ed.), New York, McGraw-Hill, pp. 41–144, 1959.

[35] Longuet-Higgins, H. C. *Mental Processes: Studies in Cognitive Science*. Cambridge, MA: MIT Press, 1987.

[36] Longuet-Higgins, H. C. "A mechanism for the storage of temporal correlations." In *The Computing Neuron*, R. Durbin, C. Miall and G. Mitchison (eds.), Wokingham, MA: Addison-Wesley, pp. 99–104, 1989.

[37] Lyon, R. and Shamma, S. "Auditory representations of timbre and pitch." In *Auditory Computation*., H. Hawkins, T. McMullen, A. N. Popper and R. R. Fay (eds.), New York: Springer Verlag, pp. 221-270, 1995.

[38] MacKay, D. M. "Self-organization in the time domain." In *Self-Organizing Systems*, M. C. Yovitts, G. T. Jacobi and G. D. Goldstein (eds.), Washington, DC: Spartan Books, pp. 37–48, 1962.

[39] Meddis, R. and Hewitt, M. J. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification." *J. Acoust. Soc. Am.,* 89: 2866–2882, 1991.

[40] Meddis, R. and Hewitt, M. J. "Modeling the perception of concurrent vowels with different fundamental frequencies." *J. Acoust. Soc. Am.,* 91: 233–245, 1992.

[41] Meyer-Eppler, W. "Exhaustion methods of selecting signals from noisy backgrounds." In *Communication Theory*., W. Jackson (ed.), London: Butterworths, pp.183–194, 1953.

[42] Miller, R. R. and Barnet, R. C. "The role of time in elementary associations." *Current Directions in Psychological Science,* 2: 106–111, 1993.

[43] Moore, B. C. J. *An Introduction to the Psychology of Hearing* (3rd ed.). London, Academic, 1997.

[44] Morrell, F. "Electrical signs of sensory coding." In *The Neurosciences: A Study Program*., G. C. Quarton, T. Melnechuck and F. O. Schmitt (eds.), New York: Rockefeller University Press, pp. 452–469, 1967.

[45] Palmer, A. R. "Segregation of the responses to paired vowels in the auditory nerve of the guinea pig using autocorrelation." In *The Auditory Processing of Speech*., M. E. H. Schouten (ed.), Berlin: Mouton de Gruyter, pp. 115–124, 1992.

[46] Patterson, R. D., Allerhand, M. H. and Giguere, C. "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform." *J. Acoust. Soc. Am.,* 98: 1890–1894, 1995.

[47] Perkell, D. H. and Bullock, T. H. "Neural Coding." *Neurosci. Res. Prog. Bull.,* 6: 221–348, 1968.

[48] Rees, A. and Møller. A. R. "Stimulus properties influencing the responses of inferior colliculus neurons to amplitude-modulated sounds." *Hearing Res.,* 27: 129–143, 1987.

[49] Rees, A. and Palmer. A. R. "Neuronal responses to amplitude-modulated and pure-tone stimuli in the guinea pig inferior colliculus, and their modification by broadband noise." *J. Acoust. Soc. Am.,* 85: 1978–1994, 1989.

[50] Reitboeck, H. J. "Neural mechanisms of pattern recognition." In *Sensory Processing in the Mammalian Brain*., J. S. Lund (ed.), Oxford: Oxford University Press, pp. 307–330, 1989.

[51] Rhode, W. S. "Interspike intervals as correlates of periodicity pitch in cat cochlear nucleus." *J. Acoust. Soc. Am.,* 97: 2414–2429, 1995.

[52] Rosenbaum, D. A. "Broadcast theory of timing." In *Timing of Behavior: Neural, Psychological and Computational Perspectives*., D. A. Rosenbaum and C. E. Collyer (eds.), Cambridge, MA: MIT Press, pp. 215–235, 1998.

[53] Schreiner, C. E. and Langner, G. "Coding of temporal patterns in the central auditory nervous system." In *Auditory Function: Neurobiological Bases of Hearing*., G. Edelman, W. E. Gall, and W. M Cowan (eds.), New York: Wiley, pp. 337–361, 1988.

[54] Schwartz, D. W. F. and Tomlinson, R. W. W. "Spectral response patterns of auditory cortex neurons to harmonic complex tones in alert monkey *(Macaca mulatta)*." *J. Neurophys.,* 64: 282–298, 1990.

[55] Siebert, W. M. "Frequency discrimination in the auditory system: Place or periodicity mechanisms?" *Proc. IEEE,* 58: 723–730, 1970.

[56] Slaney, M. and Lyon, R. F. "On the importance of time — a temporal representation of sound." In *Visual Representations of Speech Signals*., M. Cooke, S. Beet and M. Crawford (eds.), New York: Wiley, pp. 95–118, 1993.

[57] Steinschneider, M., Reser, D. H., Fishman, Y. I. and Arezzo, J. "Click train encoding in primary auditory cortex of the awake monkey: Evidence for two mechanisms subserving pitch perception." *J. Acoust. Soc. Am.,* 104: 2395–2955, 1998.

[58] Thatcher, R. W. and John, E. R. *Functional Neuroscience, Volume I. Foundations of Cognitive Processes*. Hillsdale, NJ: Lawrence Erlbaum, 1977.

[59] Yellott, J. I. and Iverson, G. J. “Uniqueness properties of higher-order autocorrelation functions.” *J. Optic. Soc. Am. A,* 9: 388–404, 1992.

AUDITORY DYNAMICS

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

AUDITORY DYNAMICS

Shihab Shamma

Center for Auditory and Acoustics Research
Institute for Systems Research
Electrical Engineering Department
University of Maryland, College Park, MD 20742, USA

The auditory cortex is critically involved in complex perceptual tasks such as timbre and pitch processing and the localization of sound sources. Its exact role in accomplishing these tasks, however, remains largely uncertain. A fundamental reason is that physiological responses to standard test paradigms with tones, noise, and other "stationary" well controlled stimuli have not been easily interpretable. For example, single-tone tuning curves and response areas measured in the auditory cortex do not usually resemble the classic V-shapes seen on the auditory-nerve. Instead, cortical cells, if they respond at all, exhibit complex variations on a theme, with tuning curves of different bandwidths, asymmetries, thresholds, and multiple excitatory areas. To add to this complexity, cortical responses are also variable in their temporal coarse, some being very basic (with only a few onset spikes) while others exhibiting sustained but oscillatory response. Similar difficulty in interpreting or eliciting responses has been encountered with other relatively simple stimuli such as noise, AM, FM tones.

Another conceptual difficulty concerns the fundamentally different nature of cortical responses compared to those of the early auditory stages. In the periphery, auditory responses are generally vigorous and phase-locked to noise-like stimuli, and hence are readily suited for generic systems analysis methods such as the reverse-correlation method. These methods have yielded interesting insights into the nature and shape of unit response areas in the auditory-nerve and cochlear nucleus. To adapt the same general methods to cortical physiology, it was necessary to adopt a paradigm shift because cortical units respond relatively poorly to noise stimuli, and rarely phase-lock to their acoustic waveforms. Instead, cortical units respond well to other "higher level" features of the stimulus, and hence our noise waveform must be "noisy in that feature space," and all analysis must abstracted and performed in this feature space.

For example, several investigators have recently observed that cortical units respond well and phase-lock to "modulations" on the spectral envelopes of broadband acoustic stimuli. If one treats these spectral modulations as the "driving stimulus" to the cortical unit, then one can abstract to this "modulation domain" and apply the same standard analysis methods described above. For instance, we can measure the tuning of cortical cells to spectral modulation rates, and use the reverse correlation method to measure the response field of these units.

The paper described in this section provides a nice illustration of these ideas, and their utility in addressing more complex questions such as the origin of cortical response properties, their relationship to pre-cortical stages such as the thalamus, and the transformations that occur at the various auditory stages. The experiments illustrate that without the notion of a stimulus with a dynamic spectrum, it is difficult (or even meaningless) to talk about a cortical response field. Furthermore, they also demonstrate that using an apparently more complex stimulus (the dynamic ripple), one gains an elegant intuitive interpretations of cortical mechanisms that is beyond reach in the responses to simple tones and noise.

A particularly interesting finding of this paper is the relationship between the thalamocortical "intrinsic" (non-stimulus related, and presumably non-functional) oscillations to the responses evoked by complex dynamic stimuli. It seems that when the thalamocortical system is engaged by stimulus driven oscillations, the system suppresses its non-functional synchrony. Such suppression is not observed when simple tonal stimuli are used, reflecting perhaps the relative functional significance of these stimuli to the central auditory system. An exciting extension of this finding is to assess their perceptual relevance, that is, whether suppression of intrinsic oscillations can be related in some reliable way to perception, and in turn to the functional relevance of different thelamocortical structures to these percepts. As is usual with interesting and profound findings, these data spark more ideas and questions than they manage to answer.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

SYNCHRONOUS OSCILLATIONS IN THE THALAMOCORTICAL SYSTEM AND THE EFFECTS OF NATURALISTIC RIPPLE STIMULI

Lee M. Miller, Monty A. Escabí, Christoph E. Schreiner

W. M. Keck Center for Integrative Neuroscience
University of California, San Francisco
San Francisco, CA 94103

1. Introduction

Computational approaches to auditory function are intended to illuminate with modeling and mathematical rigor certain properties of the auditory system that would otherwise have remained obscured. In this chapter, we describe how a modified Wiener-systems approach provides such insight into central auditory representations. Our discussion is two-pronged, alternating between the rationale for using certain stimuli and the interaction such stimuli have with the dynamic state of the system — in this instance, the thalamocortical loop. We begin with the relative merits of conventional, simple stimuli and the effects they have on thalamocortical oscillatory synchrony. Then we motivate the use of parameterized, naturalistic stimuli suitable for Wiener-systems analysis and describe the very different effects these sounds have on the dynamic state.

2. Simple Stimuli and the Thalamocortical Dynamic State

2.1 Conventional Simple Stimuli.

Much central auditory research has traditionally involved probing a neuron's receptive field with rather simple stimuli [3] [14] [24] [27]. For instance, pure tones have been widely used to probe a neuron's spectral response preferences and repetitive clicks have been used to examine temporal response properties. Many variants of these two stimuli, from amplitude-modulated tones to frequency-modulated sweeps and noise bursts of varying duration have also been used. Because of their simplicity and their prevalence in the literature we describe only pure tones and clicks. The loose dissociation between spectral and temporal properties suggested by these stimuli is conceptually appealing, but as described below, central auditory representations do not necessarily maintain this dissociation.

2.1.1 Pure Tones

Pure tones are, in many respects, the simplest possible stimuli for the auditory system. Conceptually (given the transduction of sounds in the cochlea) pure tones of low and moderate intensity should be the only stimuli that excite a relatively local portion of the sensory epithelium with a constant driving force. In exploring a neuron's response properties, tones are often presented over a range of frequencies and intensities, yielding the receptive-field measure known as the frequency response area. This is simply the firing rate of the neuron plotted on an ordered array of frequencies and intensities. Auditory neuroscientists have learned a tremendous amount about frequency representations in the central auditory system by using pure-tone stimuli. Due to the tonotopic nature of the entire lemniscal pathway, tones will continue to be the stimulus of choice for anatomists and physiologists orienting themselves within an auditory neural structure.

In addition to their intuitive application for characterizing frequency responses, tones may also be used to probe temporal response properties. They have been used, for instance, as sinusoidally amplitude-modulated tones [28], by presenting different tones simultaneously and in quick succession [5] [6], or by assessing interaural phase differences (which are important for sound localization) [25]. In central stations of the auditory pathway, however, it is also common to use brief, broadband stimuli to probe temporal responses.

2.1.2 Clicks

Clicks are, in a sense, the simple complement to pure tones. Rather than being spectrally compact, they are spectrally very broad. The energy in an ideal click is, in fact, equally distributed across all frequencies. Thus, clicks should excite the entire sensory epithelium impulsively. They are consequently of little use in probing a neuron's frequency response area, but they are well-suited for studying a neuron's temporal response properties. For example, clicks can help illustrate whether a neuron responds better to certain stimulus repetition frequencies than to others and whether certain temporal patterns tend to facilitate or suppress neural responses [12].

2.1.3 Assumptions in Using Simple Stimuli

Regardless of the intuitive appeal of certain stimuli, when we choose them to probe the workings of a real neural system, we must consider how well our characterizations generalize to other kinds of stimuli or to other states of the system. In other words, how much do we learn about the neural representations of complex, natural sounds by studying the representations of very simple sounds? The underlying assumption is that we learn a considerable amount about how the auditory system responds to more complex stimuli, which are, in theory, a superposition of many simple stimuli. In systems analysis terms, we assume that the system is linear in some important ways. If we assume the system is perfectly linear, then either pure tone stimuli or clicks ought to provide a complete and general description of the system. We will show why this assumption is unwarranted. Another expectation we hold when using simple stimuli is that the neural system has only one possible dynamic state or mode of response. This is similar to an assumption of stationarity, under which the system's basic dynamics or statistics of response remain constant through time. We will also show that even this assumption fails to hold under some conditions.

2.2 Effects of Simple Stimuli on Dynamic State

Beyond the theoretical merits of simple stimuli one may also ask whether they capture a system's dynamics accurately and consistently. The thalamocortical network provides a striking challenge as a neural system since it has more than one basic dynamic mode. These modes, moreover, may change during a given stimulus condition.

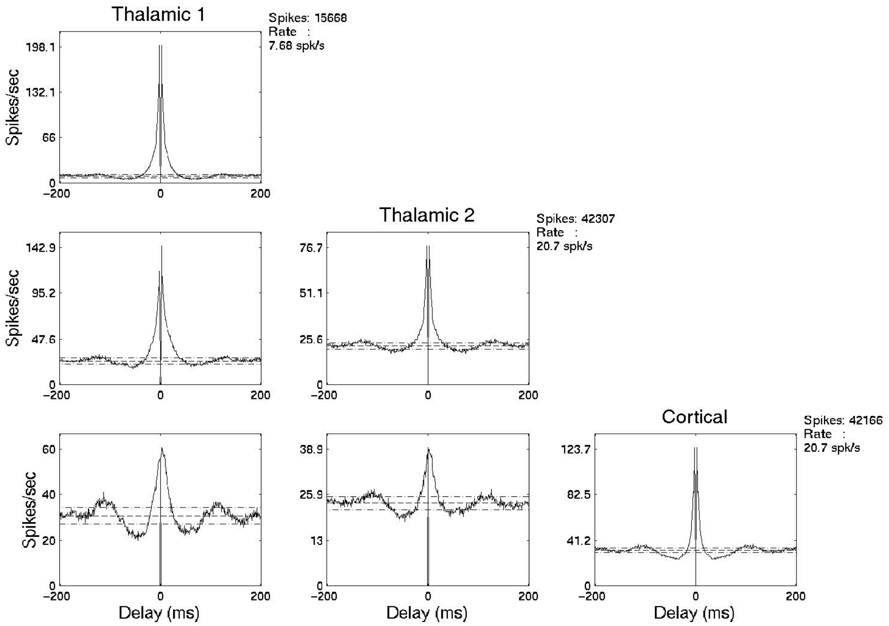


Figure 1 Thalamocortical correlograms under spontaneous conditions. Auto-correlograms for the three cells (two thalamic, one cortical) are plotted on the diagonal; cross-correlograms among the cells are plotted off the diagonal. These correlograms are typical in that their oscillatory frequency falls between 7 and 14 Hz. A similar structure is often seen under pure-tone driven conditions as well. Dashed and dash-dotted lines are expected values and 99% confidence limits, respectively, based on a null hypothesis of independent, Poisson spike trains.

2.2.1 Spontaneous Conditions: Global, Synchronous Oscillations (7–14 Hz)

Global, synchronous oscillations have been observed in about 30% of recordings from auditory cortex [13] and in the thalamocortical system [22] of the ketamine-anesthetized cat under spontaneous conditions. The oscillations usually fall within the 7-14 Hz range and the thalamic and cortical cycles tend to be in phase. Figure 1 shows cross-correlograms of spike trains recorded simultaneously in MGBv and AI in silence. The central peaks' location at or near zero reflects the zero-phase relation within and between thalamus and cortex, and the first side-peaks at ca. 115 ms delay reflect the ca. 9-Hz periodicity of the oscillations.

Shown in Figure 2 are the spectrotemporal receptive fields (STRFs) of the same neurons illustrated in Figure 1. The STRF describes to a first-order approximation a neuron's preferred spectro-temporal stimulus (a complete description of STRFs is provided in Section 3.1.3). In this case, while the two thalamic neurons share many preferences, the thalamic and cortical neurons differ in some marked respects, such as temporal modulation rate and FM sweep speed and direction. Yet, despite this receptive field disparity, strong correlations occur among all of these cells under spontaneous conditions in the form of global, synchronous oscillations, both within and across the thalamus and cortex. It appears that the presence of oscillatory synchrony does not depend on the similarity or disparity of the participant neurons' receptive fields.

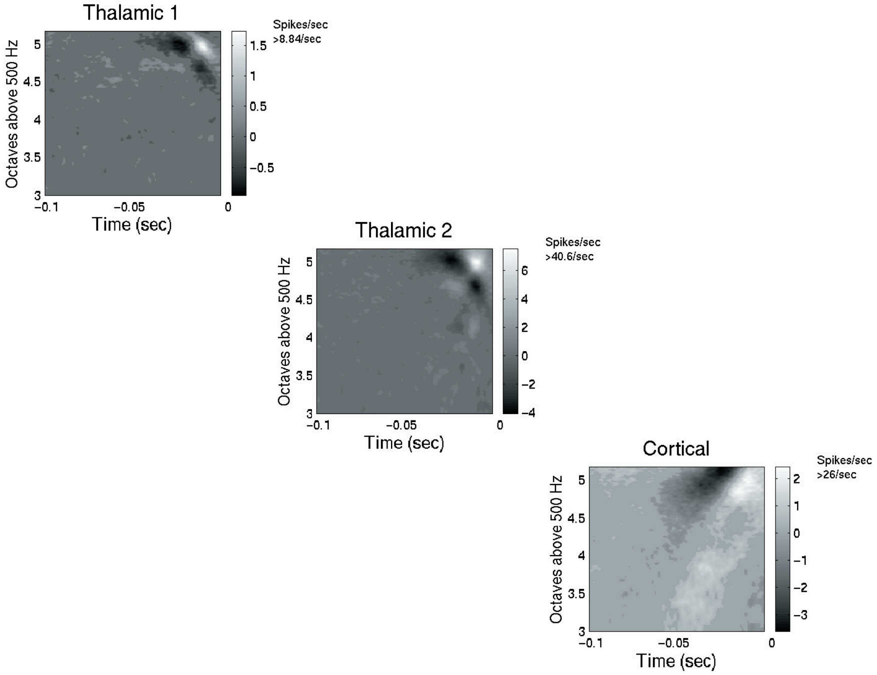


Figure 2 Spectro–temporal receptive fields (STRFs) for the units shown in Figure 1. Frequency, in octaves above 500 Hz, is represented along the ordinate (3-5 octaves corresponds to 4-16 kHz), and the time preceding a spike is represented along the abscissa. The STRFs are expressed in differential spike rates, with respect to the mean rate during the stimulus presentation. Thus, brighter areas denote spectro-temporal features that increase the firing rate above the mean, and darker areas denote those that reduce the firing rate. While the two thalamic units share some spectro-temporal preferences, the thalamic and cortical units differ markedly in many respects, including temporal modulation preference and FM-sweep speed and direction.

2.2.2 Under Stimulation with Simple Stimuli

Synchronous oscillations in the 7-14 Hz range are not only observed under spontaneous conditions. They have also been observed across the primary auditory cortex of anesthetized cats under simple stimulus conditions [13] [22], as well as in the thalamocortical system of the anesthetized rat when driven by tones or clicks [11]. As is the case with spontaneous conditions, however, the oscillations are not *always* present (Eggermont observed them in ca. 60% of recordings [13]). In the next section we explore the implications of a dynamic state that is unpredictable.

2.2.3 Potential Difficulties with Oscillatory Dynamic

It is reasonable to assume that the presence of global, synchronous oscillations in the 7-14 Hz range is indicative of a certain thalamocortical dynamic state [19] [29] [30] and that their absence is indicative of a different state. However, this assumption raises the following difficulty: oscillatory synchrony across the thalamocortical system is commonly but *sporadically* present in the ketamine-anesthetized cat under spontaneous and tone-driven conditions. That is, under the same experimental conditions, the system is apparently not always in the same state. Without knowing which state the system is in, we may unwittingly pool data pertaining to spectral and temporal receptive field properties gathered from what may be considered effectively different neural networks, thereby confounding or washing out any effects that may differ among states. This is the bane of an uncontrolled experiment, where effects cannot be unambiguously assigned to certain experimental causes.

We first consider a principled approach to more complex, naturalistic stimuli. We then revisit the issue of dynamic states with these new methods in hand.

3. Naturalistic Ripple Stimuli and their Effects on Dynamic State

3.1 Dynamic Ripple Stimuli

3.1.1 Rationale for Using Dynamic Ripple Stimuli: A Naturalistic Alternative

In the real world humans and other animals are exposed to a variety of dynamic sounds which contain time-varying spectra. Features such as spectral modulations of formants, temporal modulations, clicks and FM sweeps are commonly observed in natural sounds and are thought to convey much of the content-carrying information. It is well known that the temporal and spectral dimensions of the auditory stimulus are important since neurons at all levels of the auditory pathway can respond selectively to temporal and spectral stimulus features. Such response properties are known to be associated with auditory percepts such as pitch and timbre.

Under a naturalistic stimulus scenario, neurons in the central auditory pathway are exposed to continuous stimulation and are dynamically bombarded with stimulus onsets and offsets which can coexist along the temporal and spectral dimensions of the stimulus. However, neuronal response properties associated with spectral and temporal features of acoustic stimuli are traditionally studied independently using “laboratory type” stimuli such as tone bursts, modulated tones, broad-band noise and clicks. Such stimuli have envelopes that are highly biased and generally fail to jointly excite and probe the physiologically relevant modes of the system along the spectral and temporal dimensions. Given the complexity of the auditory neural network and the highly non-linear stimulus-response relationship observed even at the earliest stages of auditory processing [15] [16], we speculate that using naturalistic stimuli can alter the response dynamic of the central auditory system. The main goal of this section is to focus on the design of an acoustic stimulus that is theoretically sound, so as to retain an unbiased spectral and temporal modulation spectrum allowing us to explore a more naturalistic scenario of auditory processing.

One may argue for using specific natural stimuli directly instead of generic naturalistic stimuli as a means of investigating neuronal dynamics and selectivities. We avoid the use of natural sounds and animal vocalizations directly because neuronal responses to such stimuli are difficult to quantify and interpret (since the envelope features in natural sounds are inherently biased). Our choice of stimulus is motivated by the ripple spectrum noise used to obtain spectral and temporal receptive fields in the ferret and cat auditory cortex [20] [26]. The ripple stimulus is designed so that the stimulus spectrum is a sinusoidal grating on a log-frequency and log-intensity axis. It is analogous to the spatial sinusoidal gratings that are commonly used in visual experiments to investigate neural sensitivity [8] [9]. Figure 3A shows the spectro-temporal envelope of a segment of the dynamic moving ripple stimulus. At any instant of time (Figure 3c) the envelope takes a sinusoidal shape on a log-log axis where the envelope frequency, $\Omega(t)$ (units of cycles per octave), varies dynamically with time. Along the temporal axis (Figure 3B), the envelope is seen to turn on and off dynami-

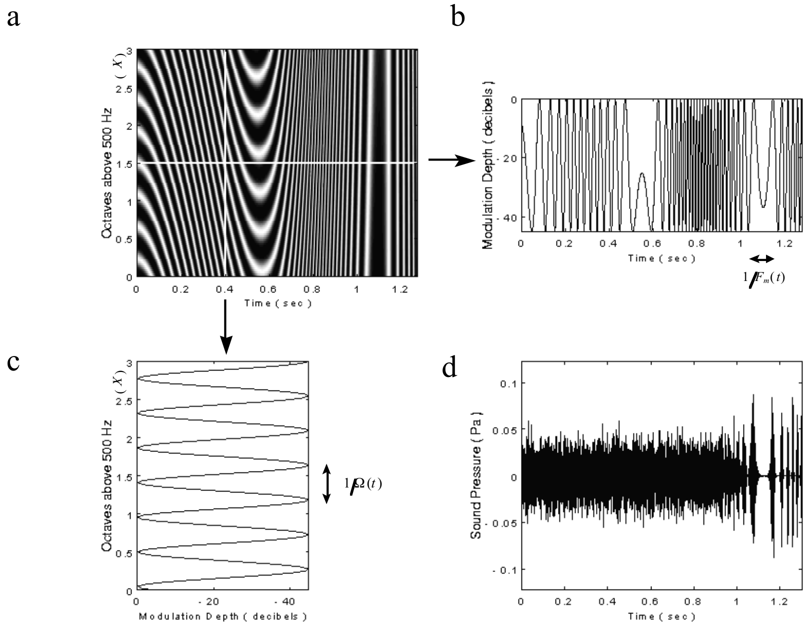


Figure 3 (a) The dynamic ripple spectro–temporal envelope, $S_{DR}(t,X_k)$, showing the ripples (the peaks along the spectral axis) which move downward or upward in time (creating FM sweeps) depending on the sign of the modulation parameter, $F_m(t)$. (b) At a given frequency, the shape of the temporal envelope changes dynamically with time because of the time-varying nature of the parameters $F_m(t)$ and $\Omega(t)$. (c) At any given instant in time the spectral envelope assumes a sinusoidal shape on an octave-frequency versus decibel-amplitude axis as described by equation (2). The spectral separation of adjacent peaks is determined by and is inversely related to the ripple frequency, $\Omega(t)$, at that specific time instant. (d) The sound pressure waveform is obtained by multiplying each carrier frequency by the corresponding temporal envelope and summing across all frequencies (as described by Equation (1)). The acoustic waveform has a "white noise" character since the phase components of each carrier are chosen independently from a random distribution. Each of the representations (b)-(d) is closely related to the spectro-temporal envelope (a). For example, at moments where the ripple frequency parameter is close to zero, as it is around time = 1.1 sec., the short-time spectrum in (a) becomes very broadband. At that moment, the structure in the temporal envelope can be observed as a sequence of clicks in the sound pressure waveform (d).

cally so that the temporal modulation rate, $F_m(t)$ (units of Hz), varies as a function of time. The dynamic ripple stimulus is of particular interest since it mimics the dynamic spectral profiles created by formants (spectral resonance) in speech production and animal vocalizations.

In addition to preserving stimulus features that are common to natural sounds, the dynamic ripple stimulus is designed to retain the basic properties of white noise that are necessary for obtaining reverse-correlation measurements: *(1)* a flat power spectrum and impulsive auto-correlation function, $R(\tau)$, in the vicinity of $\tau = 0$, *(2)* a flat envelope spectrum and impulsive spectro-temporal envelope auto-correlation functions. We distinguish these two constraints and note that property *(1)* is imposed on the signal carriers requiring that they have a white-noise character. To account for the fact that the auditory sensory epithelium is arranged logarithmically in the basilar membrane the acoustic stimulus is designed so that requirement *(1)* is satisfied on a octave-frequency axis. Constraint *(2)*, on the other hand, is imposed on the spectro-temporal stimulus envelope, a second-order property of the stimulus [7] [18]. We require that the stimulus envelope is globally unbiased, so that all spectral envelope and temporal modulation frequencies are equally represented within the physiologically relevant range. In addition, it is required that the stimulus be *globally* uncorrelated along these two dimensions, hence allowing us to perform reverse correlation measurements with respect to the stimulus spectro-temporal envelope. We note that despite this *global* correlation property, the dynamic ripple stimulus is *locally* correlated at any time-frequency instant (as is the case with many natural stimuli [23]), where the localized correlation structure of the stimulus changes dynamically and is determined by the spectral envelope frequency, $\Omega(t)$, and temporal modulation rate, $F_m(t)$, at that given instant.

3.1.2 Design of the Dynamic Ripple Stimulus

We consider the class of acoustic noise stimuli which take the functional form:

$$x(t) = \sum_{k=1}^{L} S_{DR}(t, X_k)\sin(2\pi f_k t + \phi_k) \tag{1}$$

where $S_{DR}(t, X_k)$ is the dynamic ripple spectral profile (spectro-temporal envelope), f_k is the center frequency of the k-th sinusoid carrier component, and $X_k = \log_2(f_k/ f_l)$ is the k-th frequency component defined on an octave frequency axis relative to the first component. To satisfy the flat-power-spectrum criterion *(1),* it is required that the carrier frequencies, f_k, be geometrically spaced (carrier frequencies are separated by 0.0223 octaves and span a total range of 5.32 octaves) so that they obey an equal energy-per-octave rule. The required "white noise" correlation properties of the stimulus are obtained by allowing the carrier phase, ϕ_k, to be randomly chosen for each carrier, f_k, from a uniform distribution in the interval $[0, 2\pi]$. The acoustic noise stimulus, (1), from this point of view is generated via a bank of $L = 240$ chromatically spaced sinusoid carriers of frequency, f_k (ranging from 0.5 to 20 kHz) which are individually amplitude modulated by the dynamic ripple envelope, $S_{DR}(t,X_k)$, and randomly phase shifted by ϕ_k.

The dB-spectral profile for the dynamic ripple stimulus is designed so that it takes the general form

$$S_{DR}^{dB}(t, X_k) = 20\log(S_{DR}(t, X_k)) = \frac{M}{2}\sin(2\pi\Omega(t)X_k + \Phi(t)) - \frac{M}{2} \tag{2}$$

where M is the modulation depth given in decibels ($M = 45$ dB), $\Omega(t)$ is the ripple frequency, or equivalently, the number of spectral peaks (units of cycles per octave) along the octave frequency axis, X_k, and $\Phi(t)$ is a time-varying phase modulation that determines the relative position of the sinusoid spectrum with respect to f_1 and the temporal modulation rate of the temporal envelope. On a linear amplitude scale the spectral profile is given by:

$$S_{DR}(t, X_k) = 10^{\frac{M}{40}\sin(2\pi\Omega(t)X_k + \Phi(t)) - \frac{M}{40}} \tag{3}$$

where $S_{DR}(t, X_k)$ takes a maximum value of one and a minimum value of $10^{-M/20}$ (close to zero). The ripple frequency, $\Omega(t)$, and ripple phase, $\Phi(t)$, are allowed to vary randomly and independently as continuous functions of time over a period of 20 minutes. This allows us to design a spectral profile that is dynamic (as is the case with natural signals) and globally spectro-temporally uncorrelated so that it adheres to criterion *(2)*. It is required that the two-dimensional power spectrum of $S_{DR}(t,X_k)$ be flat and unbiased within the physiologically relevant range. Hence the spectral profile signal is designed to dynamically probe, in an unbiased manner, all physiologically relevant temporal modulation frequencies, $F_m(t)$, and ripple frequencies, $\Omega(t)$.

By definition, the instantaneous rate of change of the ripple spectral envelope along the spectral dimension, X_k, is $\Omega(t)$. This can be verified by differentiating the argument of (2) with respect to X_k and dividing by 2π [7]. This results in

$$\frac{1}{2\pi}\frac{d}{dX_k}(2\pi\Omega(t)X_k + \Phi(t)) = \Omega(t) \; . \tag{4}$$

Likewise the instantaneous temporal modulation frequency, $F_m(t)$, is

$$F_m(t) = \frac{1}{2\pi}\frac{d}{dt}(2\pi\Omega(t)X_k + \Phi(t)) = \Omega'(t)X_k + \frac{1}{2\pi}\Phi', \tag{5}$$

where the derivative is now taken with respect to the time variable. We would like to designate the parameter signals, $F_m(t)$ and $\Omega(t)$, *a priori,* so that they are statistically independent and adhere to a fixed statistical structure that yields an unbiased spectral profile. Since $F_m(t)$ is a function of $\Omega(t)$ this independence criterion is, in theory, violated. We can approximate it, however, by allowing

$$\frac{1}{2\pi}\Phi'(t) \gg \Omega'(t)X_k \tag{6}$$

so that the temporal-modulation-rate parameter, $F_m(t)$, is largely dependent on $\Phi(t)$ with little contribution from $\Omega(t)$. The parameters $F_m(t)$ and $\Omega(t)$ are allowed to vary randomly and independently, where $F_m(t)$ takes uniformly distributed values in the interval [-100, 100] Hz (negative modulation rates indicate that the ripples move from high to low frequencies producing a downward FM sweep) and $\Omega(t)$ assume uniformly distributed values in the interval [0,4] cycles per octave. Using a bandwidth of 3 Hz for $F_m(t)$ and 0.5 Hz for $\Omega(t)$, equation (6) is approximately satisfied and the parameter $F_m(t)$ has a mean-RMS error of 3.5%.

3.1.3 Reverse Correlation and the Spectro–Temporal Receptive Field (STRF)

The reverse correlation method has been successfully applied to investigate neuronal dynamics for numerous sensory systems [4] [8] [9] [10] [21] [31]. In the peripheral auditory system, the first- and second-order Wiener kernels have been widely used to investigate neural tuning and non-linear stimulus-response properties [4] [31]. In central auditory stations such techniques have proven unsuccessful, partly because neuronal sensitivities in these locations are highly non-linear with respect to the stimulus carrier. This implies that an input to a neuron at a given frequency (e.g., 5 kHz) does not produce a neuronal response at 5 kHz. Instead, central auditory neurons respond to envelope modulations which occur over different frequency-tuned channels. To overcome these limitations we employ the spectro-temporal receptive field (STRF) which was first described by Aertsen et. al. [1] [2] [17] to investigate neuronal tuning of the spectro-temporal envelope in the frog midbrain. The STRF is a descriptive functional entity which depicts the envelope features, along time and frequency dimensions, to which a neuron responds.

The STRF is obtained by performing the first-order, reverse correlation of the neuronal

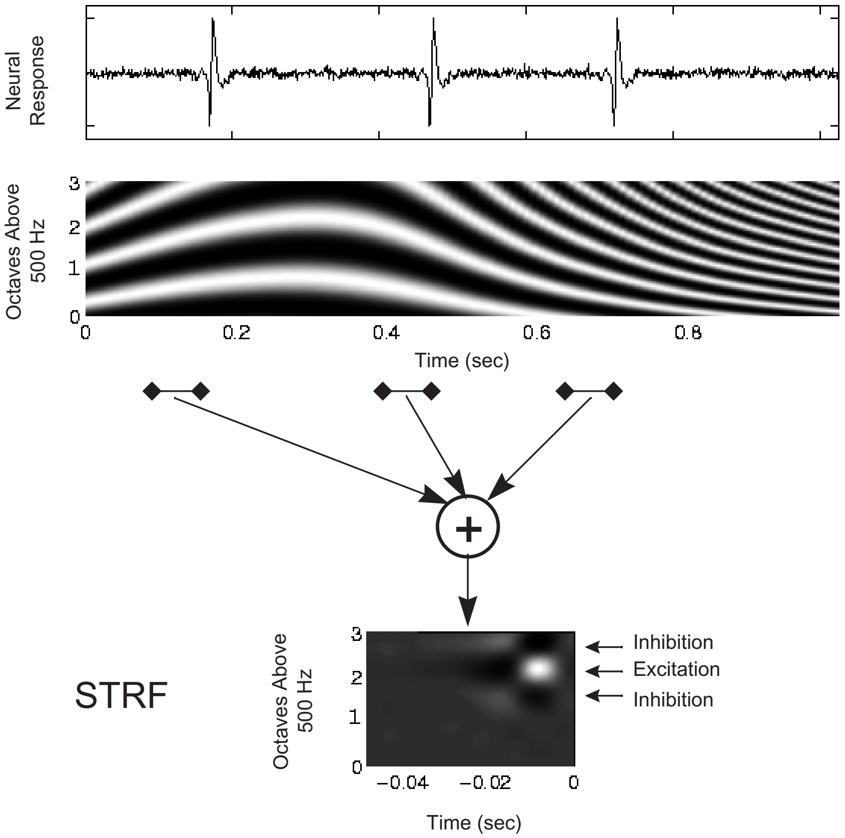


Figure 4 The STRF is obtained by averaging the pre-event, spectro–temporal envelope at all instants where a neural event (spike) occurred. The average pre-event stimulus conveys information about when the stimulus was on or off at a given time-frequency instant. In the example shown, white regions indicate that the stimulus was on whenever a neural response occurred at time zero. Similarly, dark regions indicate that the stimulus tended to be off at that specific time-frequency instant. In addition, the STRF can be interpreted as a transfer function descriptor which depicts the causal relationship between the stimulus and response. Using this interpretation, the "on" regions of the average pre-event stimulus are interpreted as excitation whereas "off" regions in the average stimulus are taken to reflect inhibitory or suppressive influences.

response, $y(t) = \Sigma\delta(t\text{-}t_n)$, with respect to the stimulus envelope, $S_{DR}(t, X_k)$. Mathematically this is expressed as

$$RF(\tau, X_k) = \frac{1}{\sigma^2_{DR}} E[(y(t) - k_0) S_{DR}(t - \tau, X_k)] = \frac{8}{TM^2} \sum_{n=1}^{N} \left[S_{DR}(t_n - \tau, X_k) - \frac{M}{2} \right] \qquad (7)$$

where, as before, M is the peak-to-peak modulation depth of the envelope in decibels, N is the number of action potentials, k_0 is the zeroth-order kernel (i.e., the mean spike rate), T is the experimental recording time, and $\sigma^2{}_{DR} = M^2/8$ is the variance of the modulation envelope (note that this is the amplitude variance of a sinusoid with amplitude $M/2$). Hence the STRF is conveniently obtained by averaging the pre-event (Figure 4) stimulus spectro-temporal envelope and normalizing by the stimulus variance.

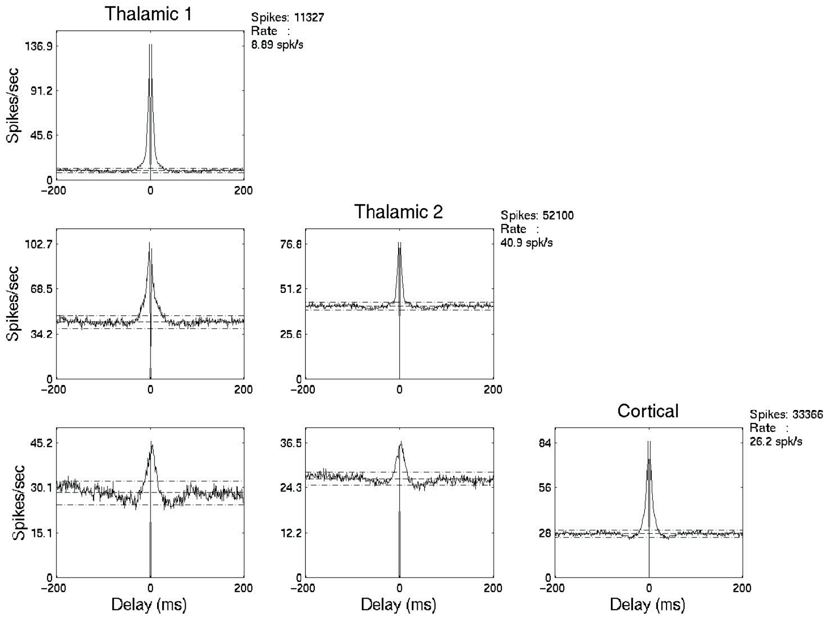


Figure 5 Thalamocortical correlograms of the same cells in Figure 1, now under ripple-driven conditions. Layout and legend are the same as in Figure 1. The synchronous oscillations have been markedly suppressed.

Intuitively, the STRF depicts the mean stimulus envelope which elicits a neural response. As an example, Figure 4 shows the STRF for a thalamic "on center / off surround" type neuron under dynamic ripple stimulation. The time axis corresponds to the time preceding the neural event which occurs at $\tau = 0$. For this example the mean stimulus envelope assumes an off-on pattern along the temporal axis indicating that the preferred temporal envelope at the neuron's CF is initially off and subsequently turns on with a time course of 10 ms. Along the spectral axis, X_k, the STRF has an on-region at the neuron's CF and flanking off regions above and below the CF, indicating that the neuron is inhibited by sounds that fall in the flanking regions. Hence the STRF provides a pictorial description of the average stimulus that produced a response.

3.2 Effects of Ripple Stimuli on Dynamic State

3.2.1 Suppression of Oscillatory Synchrony with Ripple Stimulation

Having developed a parametric, naturalistic stimulus, we return to the issue of global, oscillatory states in the thalamocortical system. As described above, synchronous oscillations in the 7-14 Hz range may occur under either spontaneous or simple-stimulus-driven conditions. Figure 5 shows cross-correlograms from the same neurons as in Figures 1 and 2, now driven by the dynamic ripple stimulus. Notice that the strong oscillations within and across the thalamus and cortex are strongly suppressed by the ripple stimulation.

It appears that, unlike simple stimuli, naturalistic dynamic ripple stimuli *consistently* suppress the degree of global, oscillatory synchrony in the thalamocortical system, thereby providing a controlled dynamic state with which to assess receptive field properties. It is thus possible, with the proper choice of stimulus, to avoid confounding data from distinct dynamic states and, at the same time, to derive an unbiased and nearly complete spectro-temporal characterization of a neuron's response properties.

4. Conclusions

Our observations may be summarized as follows:

(1) Global, synchronous oscillations in the 7-14 Hz range may occur across the thalamocortical system under both spontaneous and simple-stimulus driven conditions.

(2) The dynamic ripple stimulus provides a naturalistic alternative to simple stimuli.

(3) Under dynamic ripple stimulation, the oscillatory synchrony in the thalamocortical system tends to be suppressed.

We believe this is an instance where a principled computational approach to stimulus generation allows us to probe the auditory system more deeply and with greater confidence than through traditional means, not only by virtue of the computational power of the Wiener analysis but also through the apparent control of dynamic state that these naturalistic stimuli provide. We must also emphasize that there is nothing about these methods that makes them exclusively apt for describing the central auditory system. Our conclusions should apply across other neural levels and modalities.

Acknowledgements

Research supported by NIH grants DC 02260 and NS34853, as well as by the Whitaker Foundation and the Ford Foundation.

References

[1] Aersten, A. M. H. J., Olders, J. H. J. and Johannesma, P. I. M. "Spectro-temporal receptive fields in auditory neurons in the grass frog: analysis of the stimulus-event relation for natural stimuli." *Biol. Cybern.*, 39: 195–209, 1981.

[2] Aersten, A. M. H. J., Olders, J. H. J and Johannesma, P. I. M. "Spectro-temporal receptive fields in auditory neurons in the grass frog: Analysis of the stimulus-event relation for tonal stimuli." *Biol. Cybern.*, 38: 235–248, 1980.

[3] Aitkin, L. M., Dunlop, C. W. and Webster, W. R. "Click-evoked response patterns of single units in the medial geniculate body of the cat." *J. Neurophysiol.*, 29: 109–123, 1966.

[4] Boer, E. de "Correlation studies applied to the frequency resolution of the cochlea." *J. Aud. Res.* 9: 209–217, 1967.

[5] Brosch, M. and Schreiner, C. E. "Time course of forward masking tuning curves in cat primary auditory cortex." *J. Neurophysiol.*, 77: 923–943, 1997.

[6] Calford, M. B. and Semple, M. N. "Monaural inhibition in cat auditory cortex." *J. Neurophysiol.*, 73: 1876–1891, 1995.

[7] Cohen, L. *Time Frequency Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[8] DeAngelis, G., Ohzawa, I. and Freeman, R. "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development." *J. Neurophysiol.*, 69: 1091–1117, 1993.

[9] DeAngelis, G., Ohzawa, I. and Freeman, R. "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation." *J. Neurophysiol.*, 69: 1118–1135, 1993.

[10] DiCarlo, J. J., Johnson, K. O. and Hsiao, S. S. "Structure of receptive fields in area 3b of primary somatosensory cortex in the alert monkey." *J. Neurosci.*, 18: 2626–2645, 1998.

[11] Edeline, J-M., Cotillon, N., Nafati, B., Hars, B. and Hennevin, E. "Spindle-like oscillations evoked by acoustic stimuli in the thalamo-cortical auditory system." *Soc. Neurosci. Abstracts*, 24: 1879, 1998.

[12] Eggermont, J. J. and Smith, G. M. "Synchrony between single-unit activity and local field potentials in relation to periodicity coding in primary auditory cortex." *J. Neurophysiol.*, 73: 227–245, 1995.

[13] Eggermont, J. J. "Stimulus induced and spontaneous rhythmic firing of single units in cat primary auditory cortex." *Hear. Res.*, 61: 1–11, 1992.

[14] Etholm, B. "Activity of single medial geniculate units in response to single and double clicks." *Acta Otolaryng.* (Stockh), 81: 91–101, 1976.

[15] Galambos, R. and Davis, H. "Inhibition of activity in single auditory nerve fibers by acoustic stimulation." *J. Neurophysiol.*, 6: 287–303, 1943.

[16] Goldstein, J. L. "Auditory nonlinearity." *J. Acoust. Soc. Amer.*, 41: 676–689, 1967.

[17] Hermes, D. J., Eggermont, J. J., Aertsen, A. M. H. J. and Johannesma, P. I. M. "Spectro-temporal characteristics of single units in the auditory midbrain of the lightly anesthetized grass frog (*Rana temporaria*) investigated with tonal stimuli." *Hear. Res.*, 6: 103–126, 1982.

[18] Johannesma, P. I. M., Aersten, A. M. H. J., Cranen, B. and Van Erning, L. "The Phonochrome: A coherent spectro-temporal representation of sound." *Hear. Res.*, 5: 123–145, 1981.

[19] Kenmochi, M. and Eggermont, J. J. "Autonomous cortical rhythms affect temporal modulation transfer functions." *Neuroreport* 8: 1589–1593, 1997.

[20] Kowalski, N., Depireux, D. A. and Shamma, S. A. "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra." *J. Neurophysiol.*, 76: 3524–3534, 1996.

[21] Marmarelis, P. Z. and Marmarelis, V. Z. *Analysis of Physiological Systems Modeling. The White Noise Approach.* New York: Plenum Press, 1978.

[22] Miller L. M., Escabí M. A., Read H. L. and Schreiner C. E. "Synchrony in the auditory thalamocortical system of the cat, and its modulation with dynamic ripple stimulation." *Soc. Neurosci. Abst.*, 24: 1879, 1998.

[23] Nelken, I., Rotman, Y. and Yosef, O. B. "Specialization of the auditory system for the analysis of natural sounds." In *Central Auditory Processing and Neural Modeling*, J. Brugge (ed.), New York: Plenum Press 1998.

[24] Phillips, D. P. and Irvine, D. R. "Responses of single neurons in physiologically defined primary auditory cortex (AI) of the cat: Frequency tuning and responses to intensity." *J. Neurophysiol.*, 45: 48–58, 1981.

[25] Reale, R. A. and Brugge, J. F. "Auditory cortical neurons are sensitive to static and continuously changing interaural phase cues." *J. Neurophysiol.*, 64: 1247–1260, 1990.

[26] Schreiner, C. E. and Calhoun, B. M. "Spectral envelope coding in the cat primary auditory cortex: Properties of ripple transfer function." *Aud. Neurosci.*, 1: 39–61, 1994.

[27] Schreiner, C. E. and Mendelson, J. R. "Functional topography of cat primary auditory cortex: Distribution of integrated excitation." *J. Neurophysiol.*, 64: 1442–1459, 1990.

[28] Schreiner, C. E. and Urbas, J. V. "Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields." *Hear. Res.*, 32: 49–64. 1988.

[29] Steriade, M. and Llinás, R. "The functional states of the thalamus and the associated neuronal interplay." *Physiol. Rev.*, 68: 649–742, 1988.

[30] Tennigkeit, F., Puil, E. and Schwarz, W. F. "Firing modes and membrane properties in lemniscal auditory thalamus." *Acta Otolaryngol. (Stockh)*, 117: 254–257, 1997.

[31] van Dijk, P., Wit, H. P. and Segenhout, J. M. "Wiener kernel analysis of inner ear function in the American bullfrog." *J. Acoust. Soc. Am.*, 95: 904-919, 1994.

AUDITORY SCENE ANALYSIS

AUDITORY SCENE ANALYSIS

Malcolm Slaney

IBM Almaden Research
650 Harry Road
San Jose, CA 95120, USA

The auditory system performs many amazing tasks, all aimed at understanding the acoustic world around us. As I type these words there is music playing in the background, a workman moving rocks just outside my house, as well as the sound of my computer keyboard. Yet my auditory system has no trouble hearing each sound separately and assigning it to the proper object. Our ability to separate out all of these sounds is known as the cocktail party effect [2]; at a large gathering of people we can easily shift our attention from one conversation to another. This ability is quite remarkable, especially considering that we can organize our auditory perceptions even with only a single ear.

The basic principles that allow us to make sense of the auditory world is known as auditory scene analysis (ASA). Consider the visual situation. When looking at a scene, even as complicated as a flock of birds flying overhead, we have no problem seeing the flock as a single object, distinct from the trees and the clouds. The basic principles that allow us to group the pieces of this scene include common motion cues, color and shape expectations, as well as continuity. All of these principles have an acoustic analogue. These and other principles of auditory scene analysis are reviewed in great depth in Albert Bregman's seminal book [1].

The three papers in this section address three very different parts of the auditory scene analysis problem: musical segregation, neurophysiological modeling and speech perception.

The first chapter, by Uwe Baumann, describes a system for grouping harmonics of a sound and identifying auditory objects. Two of the strongest cues that cause us to group sounds together are common harmonicity and common onsets. Many objects, such as the laryngeal vocal folds and musical instruments, generate sound by periodically interrupting the flow of air. A periodic action leads, in the spectral domain, to a number of sinusoids, all harmonically related. The fact that the sinusoids are harmonically related is a good indication that the sinusoids are associated with the same object. Likewise, if a number of frequency components are all turned on at the same time they probably come from the same object.

Baumann describes an algorithm which uses these principles to group components of a musical sound. He starts with a high-resolution cochlear model to capture the basic spectral information in the signal. By identifying the peaks in this spectrum he can easily find what he calls part-tones. The problem then becomes a matter of extending any part-tones that are imperceptibly interrupted and then grouping the part-tones to form objects. This algorithm is tested by analyzing a musical sound with overlapping bass and soprano notes.

The second chapter, by Susan Denham, looks at the neurophysiology of a completely different part of the problem. Our auditory system has an amazing ability to not only quickly interpret a sound, but is also very tolerant of gaps and noise. A simple method to deal with such interference is to include a low-pass filter in the model. But such a low-pass filter would increase the latency of the system, which appears not to happen.

Denham describes synaptic and neural models to account for these behaviors. The primary feature of these models is that they include a limited resource in what is often called a reservoir model. When a stimulus first arrives at the cell, the neuron is primed and ready to respond immediately to a new stimulus. But the response to later portions of the same stimulus are not as strong because the limited resource has been depleted. She talks about the behavior of this model in response to several types of stimulus important in auditory scene analysis.

Finally, in the third chapter Georg Meyer and his colleagues talk about a simple form of auditory scene analysis using speech stimuli. The simplest possible cocktail party involving speech consists of two different overlapping vowels. Humans can identify both vowels, especially if they are at different pitches. This is true even when both vowels have about the same power and their pitches are fixed.

Meyer proposes a model based on modulation maps. Previous work in this area has either assumed a harmonic sieve, much like Baumann's chapter in this section, or a model based on pitch perception using autocorrelation. In Meyer's model the cochlea's spectral analysis and the amplitude-modulation detectors combine to form a two-dimensional auditory map, where concurrent vowels at different pitches are nicely separated. The vowel-identification task is solved by comparing the response at different modulation frequencies to prestored templates.

These three chapters and the references below are representative of the work that is being done on auditory scene analysis. The chapters address speech, music and the neurophysiology. But much remains to be understood about auditory scene analysis. We need better models that incorporate the effects of binaural perception, language models and common fate. Much of the current work talks about the basic modeling principle [1] of old-plus-new, but none of the these works talks about the role of attention or how our experiences guide the perceptual organization. These tasks are left as exercises for the reader.

References

[1] Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.

[2] Cherry, E. C. "Some experiments in the recognition of speech, with one and two ears." *J. Acoust. Soc. Am,* 25: 975–979, 1953.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (Eds.)
IOS Press, 2001

A Procedure for Identification and Segregation of Multiple Auditory Objects

Uwe Baumann

ENT-Department, Ludwig-Maximilians-Universität München
Marchionini-Str. 15, D-81377 München, Germany

1. Introduction

One of the most advanced signal processing skills of the human auditory system is its ability to direct attentional processes to follow the sound of a selected acoustic source in an environment of multiple, simultaneously sounding voices. For the purpose of building robust speech recognition systems and for the advancement of hearing aids it is desirable to implement signal-processing strategies which can cope with an auditory environment consisting of a mixture of different sound sources in order to enhance or extract the desired information coming from one of the sources. Perception of polyphonic music is an example of this ability. With regard to its highly systematic structure, polyphonic music was chosen as a model to investigate human listeners' strategies for grouping and for obtaining the information about musical voices.

The sound separation system outlined in this chapter is based on a perceptual model developed by Terhardt [21]. Figure 1 outlines his model. Several hierarchically ordered procedures (PROC) process signals derived from a transformation (TRANS) module. They communicate via a memory region (object buffer, OBJ BUF) wherein the extracted objects are stored. For the generation of a compulsory physical reaction (reflex) an additional connection is made for each process to the "motor system" (MOTOR SYS). In the process of ascending the hierarchy, the object buffer size is enlarged in order to hold increasingly complex information. Concerning audition, the object buffer at the beginning of the process provides simple harmonic information, whereas consecutive buffers store phones, syllables, words and sentences. The observer (called "self" by Terhardt) can attach his or her attention to any object buffer. It is estimated that the default attention is switched to buffers with the most meaningful information.

Although straight "bottom–up" models for perceptual processing have been questioned in the auditory domain (and other modalities) [5] [17], it seems worthwhile to investigate the merits and drawbacks of this approach.

2. Model

2.1 Overview of The System

A computational procedure, outlined in Figure 2, was implemented in order to separate polyphonic music into the original voices [2]. A hierarchical combination of auditory spectral analysis, psychoacoustical weighting functions and psychological elements as well as findings of the Gestalt theory are employed in this process.

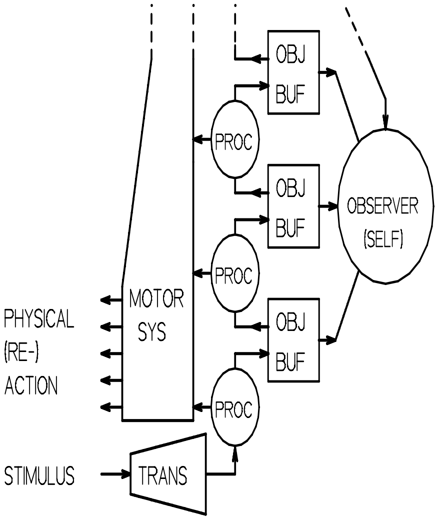


Figure 1 Perceptual model according to Terhardt [21].

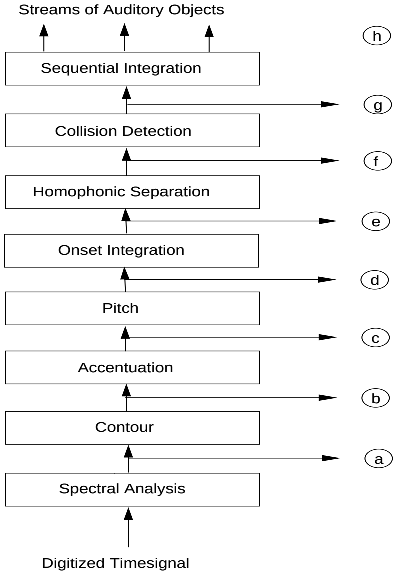


Figure 2 Flowchart of the model.

Several independent stages contribute to the abstraction and selection of meaningful contours among the spectral components. The aim is the formation of components pertinent to auditory objects. The ongoing sequence of auditory objects form a specific auditory object pattern. Circled letters denote the output of the procedure belonging to that level and are referenced to the following figures. Outputs ⓐ to ⓔ of the model are based on psychoacoustic considerations, while outputs ⓕ to ⓗ are motivated by gestalt rules. The hierarchical organization is strictly "bottom-up," although within each processing stage a certain degree of feedback might be applied. Subsequent figures refer to the outputs of stage ⓐ to ⓗ.

A brief two-voiced music example (Figure 3) demonstrates the functions of the procedure. The tone sequences were created on a programmable synthesizer; each tone of the soprano voice consisted of three harmonics and each tone of the bass line consisted of six harmonics.

2.2 Aurally Adequate Representation of Sound

After analog-to-digital conversion a special auditory-like spectral analysis (SPECTRAL ANALYSIS), with high temporal and frequency resolution, is applied to the signal [22]. Figure 4 schematically illustrates the absolute magnitude of the example's frequency-time spectrum.

The next stage of the process (CONTOUR) includes a variety of sub-modules. A peak-picking procedure is applied to each analysis interval T_A ($T_A = nT_S$, T_S sampling interval) to obtain pairs of frequency, f, and level, L, for each *part-tone* or harmonic. The set of part-tones for each interval, T_A, forms a *part-tone pattern* and the consecutively calculated part-tone patterns constitute a *part-tone time pattern* (PTTP, details in [11, 13], for examples see [14]). Phase information of the individual part-tones is completely disregarded. Figure 5 left displays the part-tone time pattern of the example tune from Figure 3 (sound level is indicated as line thickness). Compared to the spectrogram (Figure 4), the PTTP (Figure 5 left) is more easily readable, and the two voices are visually identifiable.

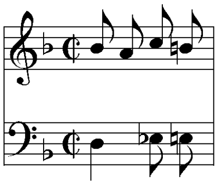


Figure 3 Score of the example music piece.

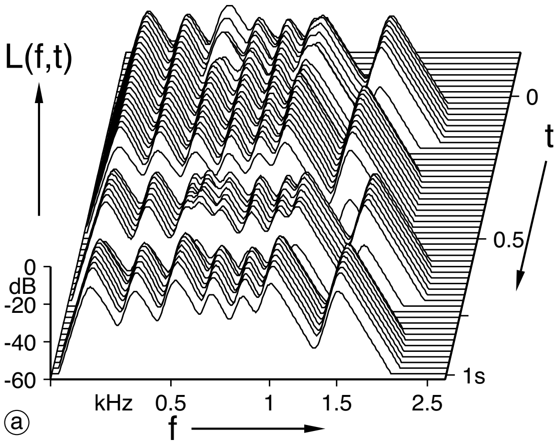


Figure 4 Spectrogram obtained with an ear-related spectral transformation according to [22].

Previous representations of sound using the PTTP did not attach perceptual information (e.g. pitch salience or partial masking). Therefore an estimate of the audibility of a part-tone was not possible. This chapter will show that the attachment of psychoacoustic information to every part-tone is crucial for the segregation task. Hence, the next sub-module of the CONTOUR estimates pitch salience for each part-tone by considering simultaneous masking. This is done by determining the level above the masked threshold (*LX*, excess level) and applying a weighting function which considers the dominance region of spectral pitch (details in [2, 18]). The range of spectral pitch weight, *WS*, is within $0 \le WS < 1$. Minor values of *WS* account for imperceivable pitches whereas values close to 1 indicate strong salience which might lead to perception of a separate pitch in a complex sound consisting of multiple part-tones. After completion of these submodules of the contour process, each part-tone information block *f, L* is expanded by pitch salience information with excess level *LX* and spectral pitch weight *WS*. The next two submodules of the contour process integrate aspects of temporal auditory perception. Up to this point of the model, the processing of information is based on time frames of the underlying spectral transformation. To obtain a

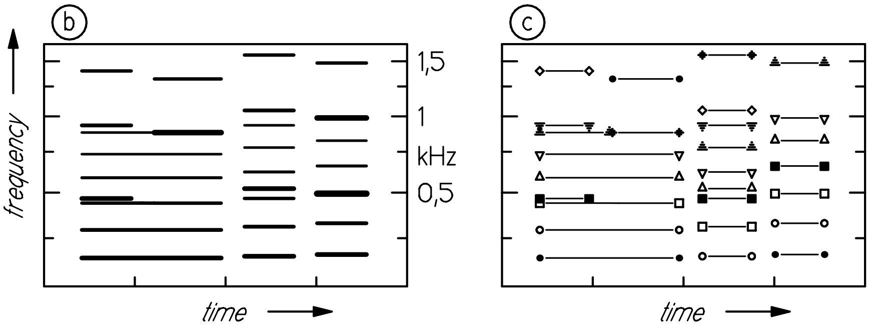


Figure 5 Part-tone time pattern (PTTP, left), Part-tone lines (right).

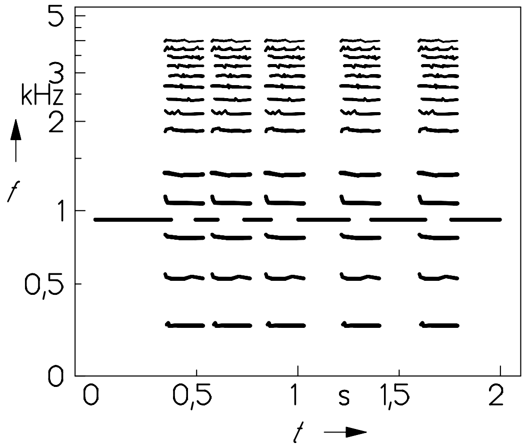


Figure 6 Part-tone time pattern representation. A sinusoidal tone is interrupted five times by an harmonic sound with an oboe timbre. If the sound pressure level of the sine tone is weak, a continuing tone is audible.

meaningful time contour representation, a procedure was implemented which links consecutive part-tones within a narrow range of level and frequency to a *part-tone line*. A unique identification code and the number of part-tones contributing to the line is attached to each part-tone line. The information collected across time frames serves to separate tonal and noisy components of the signal. An auralization of part-tone lines is easily achieved with a resynthesis procedure, using each contour pair f, L as control inputs to a sinewave oscillator.

2.3 Continuity rule

Another important feature of the line-linking submodule of the CONTOUR process is the implementation of rules pertaining to aspects of the continuity effect. To illustrate the impact of this measure, an example of continuity is outlined in the part-tone time pattern representation in Figure 6. A sinusoidal tone is interrupted five times by a harmonic sound with an oboe timbre. If the sound pressure level of the sinusoid is weak and the duration of the interruption is short (additionally, the difference in frequency and level before and after the gap has to be small), an unexpected perception occurs: instead of hearing an interrupted tone, a continuous tone is heard.

The consequences of the continuity effect on the design of sound separation systems are dramatic. If the continuity effect is not taken into account, a sound separation system may be able to segregate the harmonic sound from the interrupted sinusoid, but the auditory representation of the extracted sinusoid would be not appropriate since the listener is unable to perceive the short breaks when the mixture of sine tone and oboe sound is presented.

The first condition for the occurrence of continuity is the duration of the gap. After termination of a part tone line, candidates for membership are searched for an adjustable time range T_L. If candidates have been found, the next stage is to calculate the amount of partial masking.

Houtgast, among many others, investigated the *pulsation threshold*, which is directly related to the continuity effect [12]. The pulsation threshold is defined as the level when interruption of the test signal becomes noticeable. A relationship between simultaneous masking patterns and pulsation threshold has been demonstrated by Fastl [7]. Hence a reasonable model for the development of a continuing impression when listening to interrupted sound, is to calculate the excess level, LX_L, of the presumably masked tone. Therefore, "virtual" (i.e., unobservable using spectral analysis) part-tones are inserted with level and frequency set to values derived by linear approximation from "real" part-tones before and after the gap. If the LX_L of these "virtual" part-tones is below a certain threshold, the second condition for the occurrence of continuity is fulfilled and the gap between two part tone lines is closed with these "virtual" part-tones.

Part-tone lines are symbolized in Figure 5 (right) with different symbols for each line. A splitting of the sixth harmonic of the first bass tone occurs due to the onset of the second soprano note.

2.4 Temporal Aspects of Pitch Perception

The block labeled ACCENTUATION takes time-dependent aspects of pitch perception into account and modifies the spectral pitch weight, *WS*. Fastl has shown that pitch salience of a sinusoid depends on its duration [8]. Short sinusoids ($t < 50$ ms) acquire low salience, while signals longer than 200 ms achieve the highest degree of pitch salience. Therefore, a duration-dependent weighting function was introduced to modify the spectral pitch weight of each part-tone line.

Small onset or offset asynchrony of partial tones improves their detectability [15][10][1]. If the asynchrony is less than 30 ms the onset of a complex tone is still perceived as simultaneous. To account for this effect, the spectral pitch weight, *WS,* of every part-tone line is enhanced for onset and offset of the line.

2.5 Common onset/offset rule

Part-tone lines with similar onset times are combined in the next processing stage (labeled ONSET INTEGRATION). When a physical process initiates a sound, energy is generated in numerous frequency bands. With high probability the energy in each of those bands will start at the same moment, but there has to be some tolerance of asynchrony to accommodate the onset characteristics of musical instruments or speech. Using the results of Grey and Rasch, a time window, T_C, was introduced to simulate, to a certain extent, the perceptual fusion of asynchronous onsets [9][16]. Since the filter response of the underlying spectral transformation is not time symmetric, only a small correction for the different group delays in each analysis band is necessary to compensate for the longer response times of filters with lower center frequencies.

Although the offset of partials generated by musical instruments varies to a larger extent than the onset time, an additional rule for offset similarity was introduced. Therefore, a second time window, T_D, was defined for searching part-tone lines with similar offset.

The result of this stage is referred to as the auditory object pattern (AOP), since the AOP resembles the human capability of integrating simultaneously occurring part-tone lines into a single percept. Figure 7 (left) displays the AOP of the simple example tune. Five auditory objects were detected and marked with different symbols. Due to the simultaneity of the last two notes no segregation of the two voices occurred at this stage.

2.6 Pitch estimation

Separation of simultaneous tones is accomplished by calculating the fundamental pitch of every auditory object, accepting ambiguous results whenever indicated (block PITCH). The algorithm for calculation of virtual pitch according to Terhardt is well suited for this purpose, because missing fundamentals and pitch ambiguities are highlighted [18, 20, 19]. For every AOP a pitch calculation is executed every analysis interval, T_A, to accommodate small pitch changes, such as vibrato or jitter. This leads to a pattern of pitch actions, from

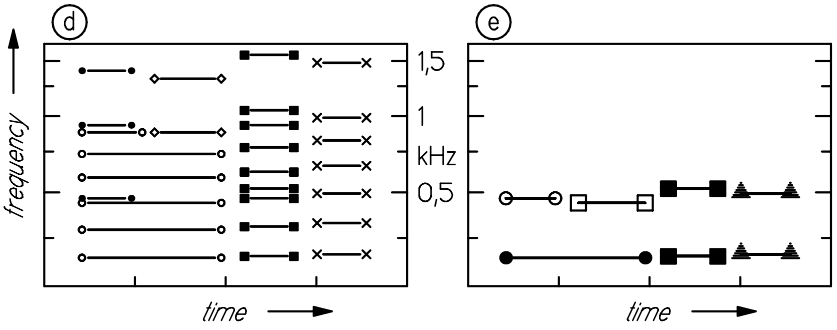


Figure 7 Auditory object pattern (AOP, left), results from pitch calculation (right).

which the most salient pitch lines are extracted and attached to the AOP. Since ambiguous pitch results are allowed, two simultaneous sounds with different fundamental frequency will result in two (or even more) virtual pitch estimates. Regarding the example, each of the two last tones of the example tune were attached with two pitch estimates. The right half of Figure 7 shows the outcome of the pitch calculation algorithm for the five auditory objects obtained.

2.7 Segregation of Simultaneous Onset and Offset

Whenever ambiguous fundamental pitch estimations were detected, the next processing step segregates the harmonics belonging to each pitch height. For each candidate part-tone line, the frequency ratio to the pitch line is calculated for each part-tone during their temporal overlap. The generated set of ratios is analyzed and an averaged deviation is calculated and compared with a pre-defined threshold value.

The ability of this process to segregate homophonously sounding voices is referred to as HOMOPHONIC SEPARATION. The precision of the pitch algorithm is extremely important for resolving ambiguities in instances where a large number of harmonics intermingle. For example, the last auditory object of the tune in Figure 3 consists of the bass, E3, and the soprano part, H4. The part-tone time pattern (Figure 5, left) displays only slight hints for a second voice beside the bass tone. A precisely balanced algorithm is necessary to predict the perceived pitches and to separate these cases. Figure 8 (left) illustrates the separation of the last two tones.

2.8 Regarding Duplex Perception

As observed with the last tone of Figure 8 (left), the third harmonic of the bass voice is intermingled with the fundamental of the soprano tone and the sixth harmonic with the second harmonic. If the soprano voice could be extracted, the resulting bass tone would miss these two harmonics and its timbre would be distorted. To avoid this, the next processing step (COLLISION DETECTION) searches for these intermingled harmonics and tries to share these collisions between the auditory objects generating the harmonics (Figure 8, right).

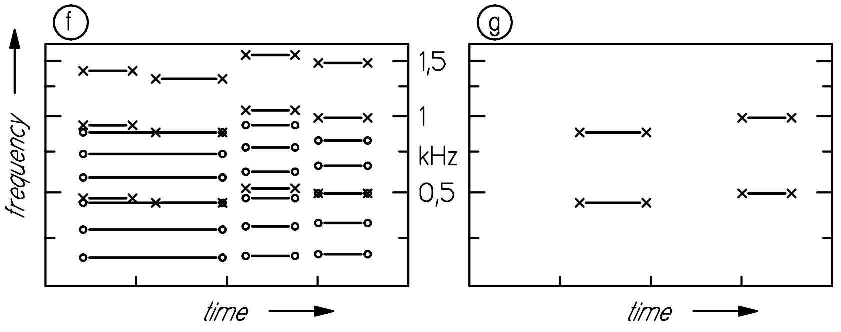


Figure 8 (left) Segregation of simultaneous part-tone lines. (right) Collision detection of intermingling harmonics. Shared harmonics are outlined.

Looked at more closely, the overlapping harmonics carry two aspects of information at once; fundamental and second harmonic for the soprano, third and sixth harmonic for the bass voice. Disregarding this duplex information will give poor segregation results.

2.9 Sequential Grouping

The final stage, termed SEQUENTIAL INTEGRATION, combines and links auditory objects to form a musical line or melody. This is perceived by a human listener attending to the desired voice, for example the bass melody. A comprehensive simulation of this auditory streaming capability is extremely hard to satisfy. For a simplified approach, timbre characteristics were not considered. Only rules for the temporal coherence of tone sequences according to van Noorden were applied [23]. An auditory stream is constructed by determining for each auditory object an unequivocal successor based on the minimal distance of the average fundamental pitch.

Figure 9 shows the extracted and restored voices. The soprano voice is on the right; the bass voice is on the left. Nearly all harmonics have been properly attached to the two voices, only the sixth harmonic of the first bass tone is disrupted by the onset of the second soprano tone.

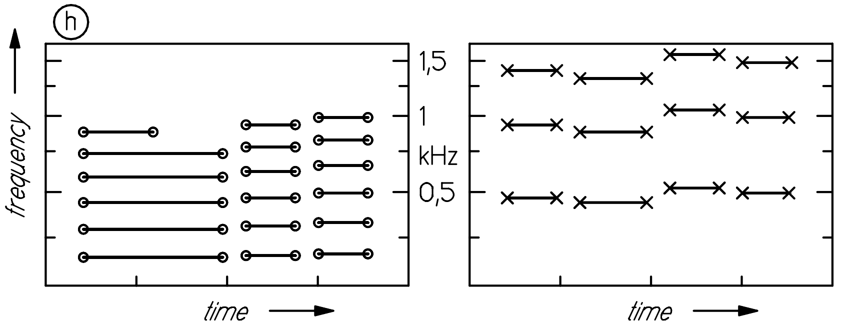


Figure 9 Segregated and restored musical voices. Left: Bass. Right: Soprano.

The example is successfully segregated by applying the procedure outlined in Figure 2. The main problems of automated voice separation are as follows:

(1) high time–frequency resolution of the underlying spectral analysis is necessary,
(2) regarding the continuity effect and formation of auditory objects based on coincident onsets is essential,
(3) pitch estimation with usage of ambiguities is crucial,
(4) restoring musical voices without distribution of shared harmonics results in unsatisfactory timbre fluctuations.

3. Results

An evaluation of the procedure with several examples of polyphonic music and speech utterances was performed. The quality of the segmentation depends on the complexity of the material. Simple two-voiced polyphonic music with a small amount of reverberation is segregated into single voices with only minor changes of timbre, whereas the time structure and melody are preserved. Figure 10 (left) displays the score of a more elaborate two-voiced piece of music.[1] The minuet was played on a digital synthesizer with a sound consisting of 6 harmonics for each voice.

Figure 10 right shows the output of stage ⓑ. The employment of a sophisticated spectral transformation method is of extraordinary importance. If two harmonics are coming close together, as occurs with the first tone of the minuet example in the frequency region around 500 Hz, a transformation with minor frequency selectivity can not resolve the bass voice third harmonic relative to the soprano fundamental. Another important property visible in Figure 10 is the high number of shared harmonics. Nearly every tone shows overlapping harmonics between the soprano and the bass voice. In some cases (e.g., F4 soprano, F3 bass voice at $t = 2$ s) only 3 harmonics remain identifiable. Hence, an application of a pitch-estimation strategy that can predict virtual pitches is significant. If tones are performed with no breaks between notes (legato) or with a large amount of reverberation, a connection of part-tone lines of successive notes may occur if the parameters of the line-linking process are not

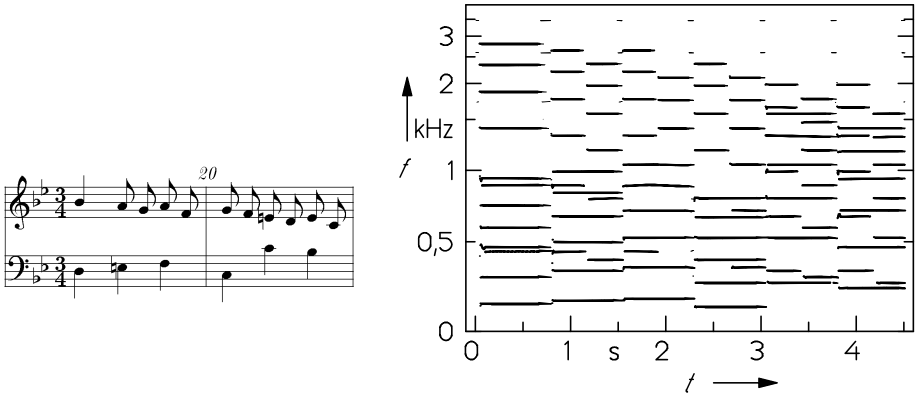


Figure 10 Left: Bars 19–20 from J. S. Bach's Minuet B^b-minor (Kleines Notenbüchlein f. Anna Magdalena Bach). Right: Part-tone time pattern of the minuet played on a synthesizer with 6 harmonics/voice.

[1] This example was also used by Brown and Cooke [4].

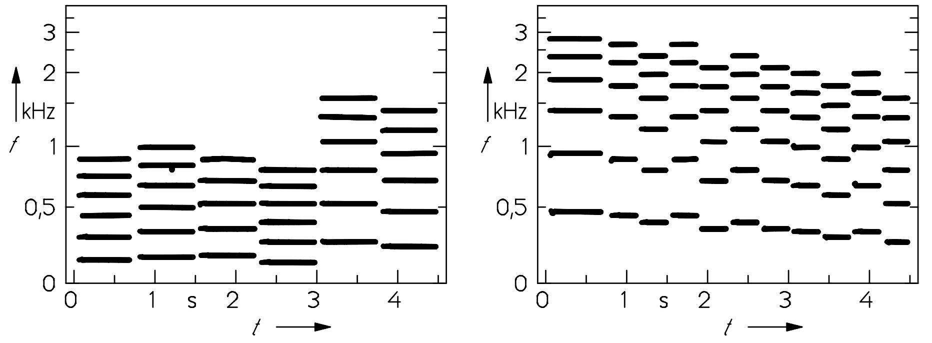


Figure 11 Left: Extracted minuet bass voice. Right: Extracted minuet soprano voice.

properly set. The octave step of the bass voice C3–C4 at about $t = 3$ s serves as an example for this situation. Due to the hierarchy of the sound separation system described in this article, a linking of the second harmonic of C3 and the fundamental of C4 would cause an undesirable misinterpretation.

The segregated and restored minuet bass voice is displayed in Figure 11 (left), the soprano voice on the right-hand side. All of the harmonics have been correctly assigned. Whenever harmonics come too close and intermingle, a modulation occurs. This modulation can lead to a perception of roughness when listening to the resynthesized voice.

Figure 12 displays the part-tone time pattern derived from the analysis of a segment of brass-band music. A long-lasting trumpet tone is accompanied by insertions from tuba, clarinet and saxophone. The extracted trumpet is displayed in Figure 12 right. Four harmonics have been attached to this auditory object. Due to the onset of the saxophone at $t = 1.1$s, the fifth and all higher harmonics were interrupted. For this reason the timbre of the resynthesized trumpet appears not to be as brilliant as the original, but the sound can easily be recognized.

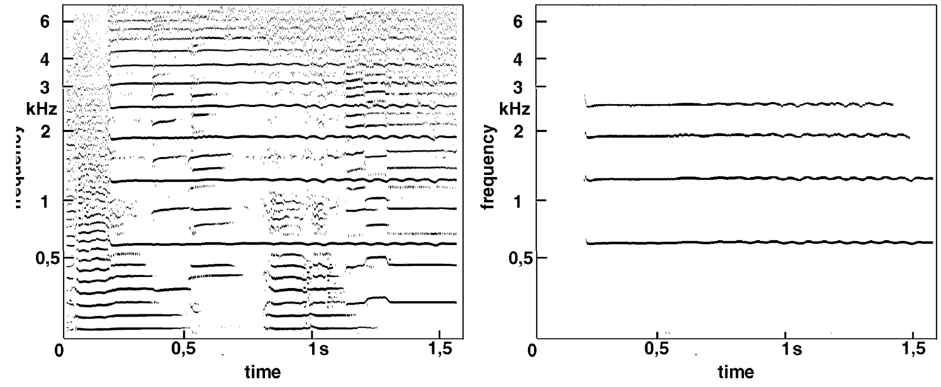


Figure 12 Left: Part-tone time patterns of a musical piece played with brass and woodwind instruments - tuba trumpet, clarinet, saxophone. Right: Extract trumpet voice.

In addition, the sound separation system extracted four auditory objects pertaining to tuba onset and two onsets of the clarinet, starting at $t = 0.4$ s. A closer inspection of the second clarinet onset reveals that two clarinets with different fundamental frequency were sounding together. However, due to the extremely sparse harmonic representation, the homophonic separation process failed to split these simultaneous tones. The saxophone note occurring at $t = 1.2$ s was correctly assigned to three onsets. The extracted and resynthesized auditory objects are easily identifiable, although a reduced brightness is noticeable. The reader is referred to [22] for acoustic demonstration and display of the results.

4. Discussion

A computational model employing a “bottom-up” strategy to simulate some aspects of auditory scene analysis has been outlined. The separation of musical voices is accomplished with a collection of “auditory elements” (part-tone lines) and grouping of elements which have common onsets and similar offsets. Although information about timbre is not evaluated, segregation of polyphonic music has been demonstrated. Concurrent musical sounds with overlapping harmonics are segregated with limited success using a strategy that distributes shared harmonics. Clearly, if the energy of the intermingling partial tones differs to a large degree, this simple approach will yield timbre errors. Another limitation is that the layout of the procedure assumes harmonic sounds. Although the applied pitch algorithm is capable of predicting the perceived pitch of inharmonic sounds such as gongs and bells, the separation process assumes a pure-harmonic relationship to predict harmonic collisions correctly.

Compared with previous attempts to segregate polyphonic music, the procedure presented in this chapter demonstrates superior performance. The algorithm proposed by Brown and Cooke [4] segregated only five soprano tones (out of eleven) and five bass tones with intense timbre distortion, due to incorrect assignment of harmonics for the minuet example displayed in Figure 10. Although their elaborate model utilized a two-dimensional timbre space to allocate harmonics as well as phase (periodicity) information in an autocorrelation map to enable pitch tracking, their approach showed only poor results for the task of segregating concurrent musical voices. Since the Brown and Cooke model is related to the procedure proposed here, a closer look at the differences in the approaches is useful to gain insight to the essential prerequisites for the task of musical voice segregation. One of the most important premises for a successful representation of auditory objects is the usage of a spectral analysis with optimized time and frequency resolution. Brown and Cook employed a gammatone filterbank with relatively large bandwidth ($b = 0.2$ ERB), therefore the representation of auditory objects showed a liability for modulations resulting from adjacent spectral components. Furthermore, they disregarded the continuity effect in their rules to form auditory elements and excluded tracks with short interruptions. Another essential difference is the usage of an “exclusive allocation” strategy and removal of elements from the auditory scene after assignment to an auditory object. Hence, shared harmonics are not resolved. Moreover, their model made use of an autocorrelation method to determine the local pitch contour of each element and therefore no consideration of pitch ambiguities was possible.

Ellis [6] noted that the problem of sound reconstruction from abstract analysis is extremely under constrained and “invention” of extra parameters based on some ideal example is required. The difficult issue of constructing a hierarchy of abstractions and parameterizations that discards only information not important to the perceived nature of the sound has been tackled to some extent by the procedure proposed in this chapter. An extension of the model with such "ideal examples" seems necessary to enhance the quality of the reconstructed musical voices.

5. Summary and Conclusion

A procedure was implemented on a computer to separate polyphonic music in the original constituent voices. A hierarchical combination of auditory spectral analysis, psychoacoustic weighting functions and psychological elements germane to the Gestalt theory served as the basis for this process. Several independent stages contributed to the task of abstraction and the selection of meaningful contours of spectral components. The use of rules for the relationship of spectral components allowed the formation of auditory objects. The ongoing sequence of auditory objects formed an auditory object pattern. The model was tested with several examples of polyphonic music. Depending on the complexity of the structure, two voiced music sources are separated with only minor timbre changes. Although not developed for the separation of speech, the algorithm is also capable of segregating syllables in a spoken sentence.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG SFB 204) and carried out at the Institute of Man–Machine Communication of the Technische Universität München. The author is grateful for the feedback, criticism and suggestions he received from all these people. Specifically, the suggestions of Ernst Terhardt improved the ideas presented here. Acoustical demonstrations and colored part tone time patterns corresponding to the figures in this article can be found on a web page [3]. Special thanks to Claus v. Rücker for his Framemaker support.

References

[1] Baumann, U. "Segregation and integration of acoustical objects in automatic analysis of music." *Proc. 3rd Intern. Conf. Music Percept. Cogn.*, I. Deliège (ed.), ESCOM, pp. 282–285, 1994.

[2] Baumann, U. *A Procedure for Identification and Segregation of Multiple Auditory Objects* (German). Munich: Herbert Utz Verlag, 1995.

[3] Baumann,U. http://ghn86x.hno.med.uni-muenchen.de/baumann/ios_book/ios_demos.html, 2001.

[4] Brown, G. J. and Cooke, M. "Perceptual grouping of musical sounds: A computational model." *J. New Mus. Res*, 23(2): 107–132, 1994.

[5] Churchland, P., Ramachandran, V. and Sejnowski, T. "A critique of pure vision." In *Large-Scale Neuronal Theories of the Brain*, C. Koch and J. Davis (eds.), Cambridge, MA: MIT Press, 1994.

[6] Ellis, D. P. W. "Hierarchic models of hearing for sound separation and reconstruction." *IEEE Workshop Apps. Sig. Proc. Audio Acous.*, 1993.

[7] Fastl, H. "Pulsation patterns of sinusoids vs. critical band noise." *Percept. Psychophys.*, 18: 95–97, 1975.

[8] Fastl, H. "Pitch strength of pure tones." *Proc. 13th Intern. Conf. Acoustics*, Belgrade, pp. 11–14, 1989.

[9] Grey, J. M. and Moorer, J. A. "Perceptual evaluations of synthesized musical instrument tones." *J. Acoust. Soc. Am.*, 62: 454–462, 1977.

[10] Hartmann, W. M. and Johnson, D. "Stream segregation and peripheral channeling." *Music Perception*, 9(2): 155–184, 1991.

[11] Heinbach, W. "Aurally adequate signal representation: The part-tone-time-pattern." *Acustica*, 67: 113–121, 1988.

[12] Houtgast, T. *Lateral Suppression in Hearing*. TNO Report, TNO Institute for Human Factors, Soesterberg, 1974.

[13] Mummert, M. *Speech Coding by Contourizing an Ear-adapted Spectrogram and its Application to Data Reduction* (German). Düsseldorf: VDI-Verlag Reihe 10, 1998.

[14] Mummert, M. http://home.t-online.de/home/Markus.Mummert/index_e.html, 1998.

[15] Rasch, R. A. “The perception of simultaneous notes such as in polyphonic music.” *Acustica*, 40: 21–33, 1978.

[16] Rasch, R. A. “Timing and synchronisation in ensemble performance.” In *Generative Processes in Music; The Psychology of Performance, Improvisation, and Composition*, J. Sloboda (ed.), Oxford: Clarendon, pp. 70–90, 1988.

[17] Slaney, M. “A critique of pure audition.” In *Proc. Computational Auditory Scene Analysis Workshop,* D. Rosenthal and H. Okuno (eds.), Montreal, pp. 13–18, 1995.

[18] Terhardt, E. “Calculating virtual pitch.” *Hearing Res*, 1: 155–182., 1979.

[19] Terhardt, E., Stoll, G. and Seewann, M. “Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions.” *J. Acoust. Soc. Am.*, 71: 671–678, 1982.

[20] Terhardt, E., Stoll, G. and Seewann, M. “Algorithm for extraction of pitch and pitch salience from complex tonal signals.” *J. Acoust. Soc. Am.*, 71: 679–688, 1982.

[21] Terhardt, E. “From speech to language: On auditory information processing.” In *The Auditory Processing of Speech: from Sounds to Words*, M. E. H. Schouten (ed.), Berlin: Mouton de Gruyter, pp. 363–380, 1992,

[22] Unkrig, A. and Baumann, U. “Spectral analysis and detection of frequency contour by use of filters with asymmetrical slopes” (German). In *Fortschritte der Akustik — DAGA ‘93*, Bad Honnef: DPG GmbH, pp. 876–879, 1993.

[23] van Noorden, L. P. A. S. *Temporal Coherence in the Perception of Tone Sequences*. Thesis, Technical University, Eindhoven, 1975.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

CORTICAL SYNAPTIC DEPRESSION AND AUDITORY PERCEPTION

Susan L. Denham

Centre for Neural and Adaptive Systems
School of Computing, University of Plymouth
Plymouth PL4 8AA, UK

1. Introduction

There are many aspects of auditory perception, such as the growth of loudness with duration and the effects of masking, which indicate that the auditory system performs some sort of temporal integration in processing incoming acoustic signals. However, the auditory system is also capable of fine temporal resolution, as evidenced by gap detection, double click discrimination, and also in the short latency and lack of jitter of onset responses in cortex [28]. This has been termed the resolution-integration paradox (i.e., how is it possible for a system to integrate information over long periods while retaining fine temporal resolution?). Most accounts satisfying the integration criterion use long time constants and therefore fail to behave swiftly enough to explain fine temporal resolution, and vice versa [28].

The time constants typically associated with sub-cortical processing differ substantially from those in the cortex. In comparison with the speed and precision associated with processing in the auditory periphery, the temporal response properties of neurons in primary auditory cortex (AI) can appear to be surprisingly sluggish. For example, in the thalamocortical transformation of incoming signals a great deal of the temporal fine structure is lost [5]; best modulation frequencies measured in AI are generally below 15 Hz [24], and the effects of a masker on a probe tone can be detected up to 400 ms after masker offset [3]. The focus in this chapter is therefore on the temporal response properties observed in AI. What gives rise to these phenomena and can they be explained by some common mechanism? As yet, there have been no models proposed that can satisfactorily explain the observed behaviour of neurons in AI. Explanations in terms of intracortical inhibitory circuits have been proposed but inhibition does not provide an adequate account, at least in the case of forward masking which is unaffected by the application of a GABA antagonist [3]. On the other hand, simple threshold neural models cannot replicate such behaviour without some form of inhibition or by means of very long time constants operating on the input signals, which as discussed above, would then prevent the model from satisfying the requirements for good temporal resolution.

Recently it has become apparent that cortical synaptic dynamics may be an important factor affecting the behaviour of biological neurons [17][18][1][25][23]. When synapses are repeatedly activated they do not simply respond in the same way to each incoming impulse, and synapses may develop a short-term depression or facilitation, depending on the nature of the pre- and post-synaptic cells, as well as on the characteristics of the particular synapse involved [25][23]. New experimental work has helped to elucidate the dynamical properties of cortical synapses, which appear to significantly influence the temporal sensitivity of cortical circuitry. Within current neural-network models synapses are generally modeled as simple gains and it is interesting to explore whether models of cortical processing can be usefully enhanced by the inclusion of a richer synaptic model. If synapses are not simply viewed as passive weighting elements in neuronal circuits, but rather as dynamical systems in their own right, then perhaps many of the response properties observed in AI might be explained in a relatively simple way.

To explore this hypothesis, a model of cortical synaptic depression was used to investigate the computational properties of a neuron model that includes dynamic synapses. This model was found to account for a surprisingly wide range of experimental observations, including those outlined above. On the basis of the model it is suggested that the dynamics of thalamocortical synapses may largely explain the temporal integration observed in AI. In addition, the model also provides a novel explanation for some puzzling effects of apparently subthreshold stimuli [20][3].

The remainder of the paper is organized as follows. First, the dynamic synapse and neuron models are described and their behaviour is illustrated. The combined model is then used to replicate a number of experiments including those investigating the transfer of information from thalamus to cortex [5], best modulation frequencies [24,13], the time course [3] and the effect of masker duration [12] on cortical forward masking, the disruptive effect of subthreshold stimuli [20], and the relationship between stimulus envelope properties and onset latency [9]. The simulation procedures used, the assumptions made and the limitations of the approach taken in these simulations are described. In the subsequent discussion we explore the implications of the model for auditory streaming and grouping and for auditory perception in general.

2. The Dynamic Synapse Model

The dynamic synapse model we use here was presented in [26] and shown to replicate the experimental results reported in that paper, and in [18], on the activity-dependent redistribution of synaptic efficacy. In fact, this model of the postulated dynamics of neurotransmitter release had already been proposed much earlier by Grossberg [7][8]. There it was derived from a set of psychological postulates and used, *inter alia*, to explain the excitatory transients in transmitter release after a rest period and related to the effects of synaptic depression, which had been observed experimentally by Eccles [6]. This synaptic depression model has been further developed and used subsequently by Grossberg in more recent years, for example to explain a number of important perceptual features involving the visual cortex. In the area of auditory modeling, a very similar model was also developed by Meddis [19] to describe transduction in cochlear inner hair cells.

The dynamic synapse model characterizes the synapse by defining a "resource," e.g. the amount of neurotransmitter in the synapse, a proportion of which can be in one of three states: *available, effective, inactive*. The dynamical behaviour of the proportions of the resource that are in each of these states is determined by a system of three coupled differential equations (1)–(3) below. In these we use notation similar to that in [8] (see equations (58)–(63)):

$$\frac{dx}{dt} = g \cdot y(t) \cdot l(t) - a \cdot x(t) \tag{1}$$

$$\frac{dy}{dt} = \beta \cdot w(t) - g \cdot y(t) \cdot I(t) \tag{2}$$

$$\frac{dw}{dt} = a \cdot x(t) - \beta \cdot w(t) \tag{3}$$

where *x(t)* is the amount of *effective* resource (e.g., activated neurotransmitter within the synaptic cleft), as a proportion of the total resource, *y(t)* is the amount of *available* resource (e.g., free neurotransmitter in the synapse), and *w(t)* is the amount of *inactive* resource (e.g., neurotransmitter being reprocessed).

The input signal, *I (t),* represents the occurrence of a presynaptic action potential (AP) and is set equal to unity at the time of arrival of the AP and for a small period of time, *δt,* thereafter, and otherwise is set to 0. The constant, *β,* determines the rate at which the inactive resource, *w(t)* (e.g., neurotransmitter which has been reprocessed), is released to the pool of available resource on a continuing basis, and α represents the rate at which the effective resource becomes rapidly inactive again (e.g., as a result of neurotransmitter reuptake), subsequent to being activated. The instantaneous efficacy of the synapse is determined by the variable, *g(t),* which can be interpreted as the fraction of available resource released as a result of the occurrence of the presynaptic AP. It takes a value in the range of zero to one.

The key idea behind the model is that there is a fixed amount, *K,* of total resource available at the synapse, a proportion, *g. y(t),* of which is activated in response to presynaptic activity, rapidly becomes inactive, and is then subsequently made available again through reprocessing. Thus, if the synapse is very active (i.e., it is bombarded by a large number of action potentials occurring over a short period of time), the amount of available resource, *y(t),* is rapidly reduced. There must then follow a period during which the synapse can recover in order to respond fully once more. This process, illustrated in Figure 1, appears to replicate the experimentally observed characteristics of synaptic depression, as reported in (for example) [18][26].

The EPSP at the synapse, *e(t)*, is computed from *x(t)* in (1) using the following equation for the passive membrane mechanism [26]:

$$\tau_{EPSP} \cdot \frac{de}{dt} = \gamma \cdot x(t) - e(t) \tag{4}$$

3. The Neuron Model

The neuron model is described by the following system of equations, which has been adapted from a model described in [16]:

$$\tau_E \frac{dE}{dt} = -E(t) + V(t) + G_K(t) \cdot (E_K - E(t)) \tag{5}$$

$$s(t) = 1\text{, if } E(t) \geq \theta(t)\text{, else } s(t) = 0 \tag{6}$$

$$\tau_{G_K} \frac{dG_K}{dt} = -G_K(t) + \eta \cdot s(t) \tag{7}$$

$$\tau_\theta \frac{d\theta}{dt} = -(\theta(t) - \theta_0) + s(t) \tag{8}$$

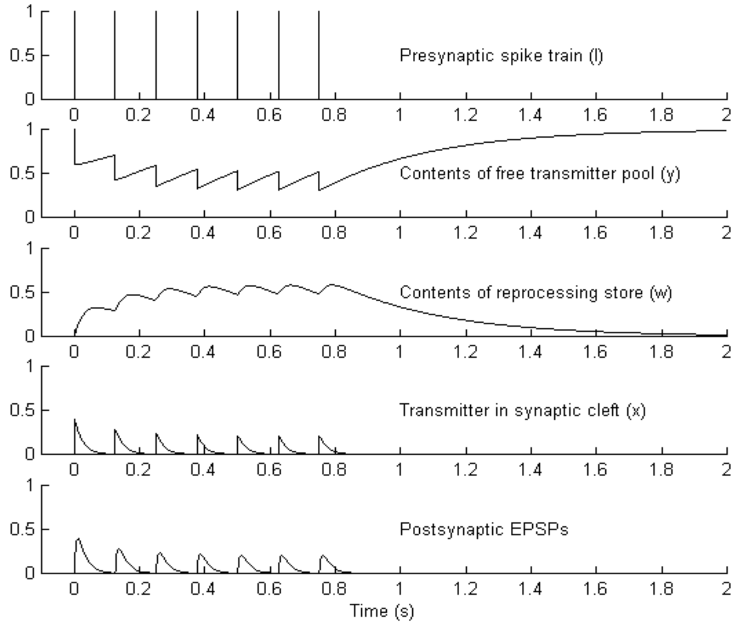


Figure 1 The response of the synaptic model to an incoming spike train.

where, $E(t)$ is the variation of the neuron's membrane potential relative to its resting potential, $V(t)$ is the driving input found by summing all the synaptic EPSPs, $G_K(t)$ is the potassium conductance, divided by the sum of all the voltage-dependent ionic membrane conductances, E_K is the potassium equilibrium potential of the membrane relative to the membrane resting potential, $\theta(t)$ is the firing threshold potential, θ_0 is the resting threshold, $s(t)$ is the variable which denotes firing of the cell, τ_E, τ_{EPSP}, τ_θ, and τ_{GK} are time constants, and γ, χ and η are constant parameters.

In this system of equations, $s(t)$ is set to 1 to signal the occurrence of an action potential (i.e., $E(t)$ reaches a value above the firing threshold, $\theta(t)$; otherwise $s(t)$ is zero). Equation (8) is introduced purely to provide a refractory period. It allows representation of an absolute period and a relative period. For the first few milliseconds the value of $\theta(t)$ is very large, preventing any firing. As $\theta(t)$ decays between spikes, the threshold for firing decreases with time elapsed since the last spike. A further spike can occur, therefore, in this period if the value of $E(t)$ is sufficiently large. When $s(t)$ is zero, the potassium conductance term, $G_K(t)$, decays to zero via equation (6). When $s(t) = 1$, the value of G_K is increased instantaneously by an amount, η, and then decays again. The behaviour of the neuron model is illustrated in Figure 2. In this case we have not explicitly modeled the action potentials generated when the cell fires, but in the simulations below generally use the spiking variable, $s(t)$, as the output from the model.

4. Simulation Results

Not all cortical synapses are depressing; for example, synapses between cortical pyramidal neurons and bi-tufted GABAergic interneurons synapses are strongly facilitating [23].

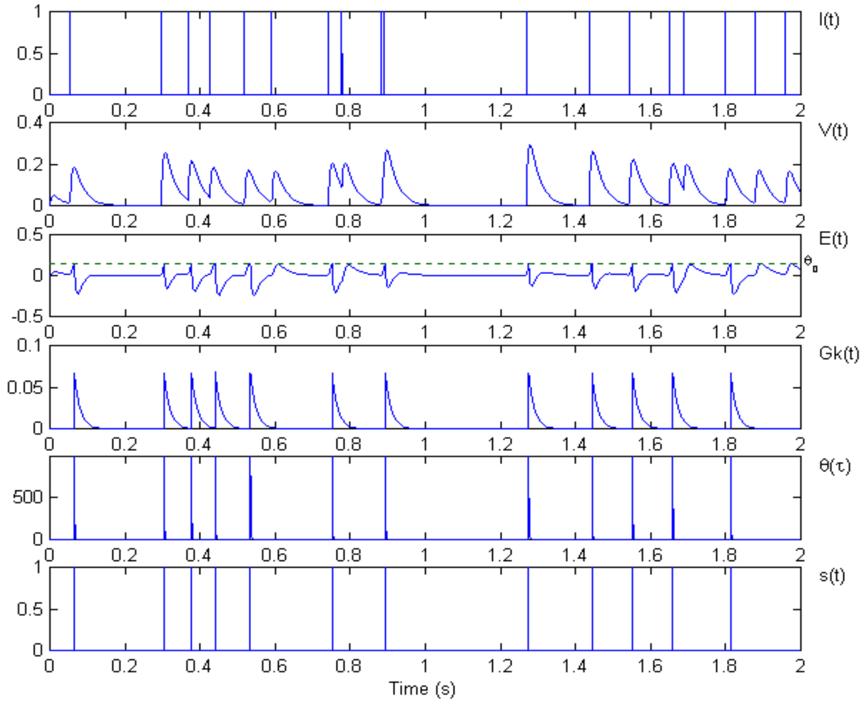


Figure 2 Response of the neuron model with a dynamic synapse to an incoming spike train showing the synaptic EPSPs, the resulting change in membrane potential, as well as the sharp increases and passive decay of GK(t) and q(t).

However, thalamocortical synapses appear to be depressing; they are mediated by non-NMDA excitatory amino acids, depress rapidly and remain desensitized for some time [25]. In the simulations that follow, it can be seen that the response characteristics of the model neuron, when the dynamic synapse model is included, turn out to be very similar to that found in primary auditory cortex. As a result it is suggested that the depression of thalamocortical synapses may provide at least a partial explanation for the responses observed.

4.1 Loss of Temporal Fine Structure in the Thalamocortical Transformation of Incoming

Differences between the response properties of thalamic and cortical neurons were investigated by Creutzfeldt *et al* [5]. Activity in thalamic relay cells and subsequent activity in paired pyramidal cells in AI was recorded, and it was found that even when thalamic activity was clearly synchronized to the stimulus up to 200 Hz, the paired cortical cell was unable to follow the details of the signal beyond about 20 Hz. The plots in Figure 3 show the response of the model to spike trains generated to resemble typical thalamic activity in response to stimuli of the frequencies indicated. Total activity for 20 presentations is plotted both for the presynaptic spike trains and the model response. The model behaviour closely resembles that found experimentally [5]. The model responds to details of the stimuli occurring at 10 Hz and to a lesser extent to details at 20 Hz, but for higher stimulus frequencies, the model only responds strongly at the onset of the signal. The reason for this is that at high frequencies successive presynaptic spikes arrive before the synapse has time to recover. This causes

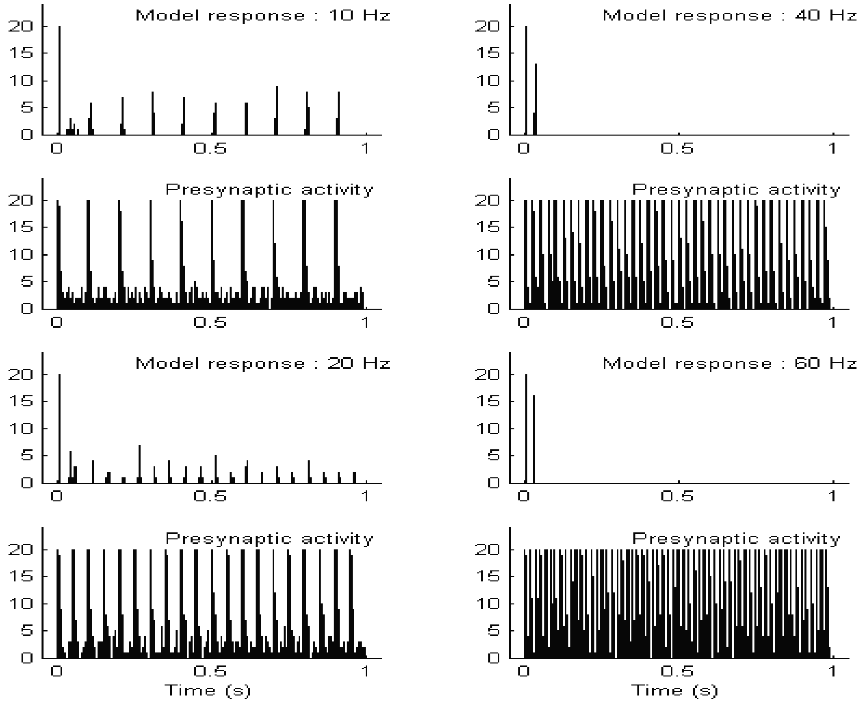


Figure 3 Simulation of the transmission of signals between thalamic relay and cortical pyramidal cells. Spikes were generated probabilistically, resulting in the distributions shown at the bottom of each quadrant, and used as inputs to the model. This input activity resembles the activity in thalamic relay cells recorded experimentally in response to signals with periodicity indicated [5]. The model (top of each quadrant) qualitatively replicates the behaviour of paired pyramidal cells in AI, which showed almost no response except at signal onset when stimuli exceeded 20 Hz.

a strong depression of the synapse, resulting in the generation of very small postsynaptic EPSPs that are insufficient to raise the cell membrane potential above the firing threshold.

4.2 Frequency Response of the Extended Neuron Model

The frequency response of the neuron model with a depressing synapse is illustrated in Figure 4B. Although the synaptic dynamics have been tuned to match those found experimentally in the somatosensory cortex, it is interesting to note that the model clearly responds preferentially to frequencies under 10 Hz, as is also found in AI. It seems to be the case that the dynamics of cortical depressing synapses may be quite similar across different cortical areas.

For comparison the response of a neuron model without a depressing synapse is shown in Figure 4C. Clearly, such a model cannot replicate the behaviour observed experimentally without the addition of delayed inhibitory inputs which increase in strength with stimulus frequency. Alternatively, modeling the synapse as a low-pass filter but with a very low cut-off frequency could result in a similar frequency response, but would fail simultaneously to account for the short response latency found in AI [9]. The benefit of the proposed model is that it can account both for the low-pass frequency response and short onset latency (cf. Section 4.8) within a single neuron model.

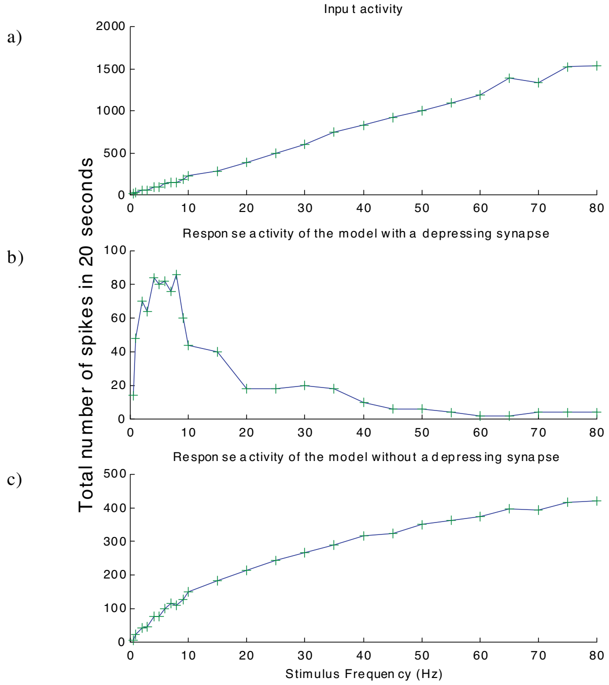


Figure 4 Frequency response of the model. The response of the neuron model with and without a depressing synapse to an incoming spike train of the frequency indicated, simulated for 20 seconds. The plot shows the total number of times the cell fired during the 20-second period. Stochastic presynaptic spike trains were used, with the probability of a spike set so as to generate, on average, the number of spikes per second indicated.

4.3 Limitations of the Simulations

For many of the experiments simulated the nature of the thalamocortical signals is unknown, which makes it difficult to know whether the stimuli used as inputs to the model are realistic. However, the details of the acoustic stimuli used in the experiments are generally well documented and therefore it is desirable to be able to simulate the experiments using similar acoustic stimuli. For this reason a well-documented and tested peripheral model, DSAM [21], was used to generate signals characteristically found in auditory nerve fiber recordings in response to acoustic stimuli. The problem with this approach is that the rest of the subcortical auditory system has not been similarly modelled. Therefore, in the following simulations the output from the peripheral model is reprocessed to ensure that the firing rate remains below about 200 Hz by enforcing a reasonable refractory period. Clearly this ignores the computations which occur in the rest of the auditory system. However, it is surprising how many results the model can replicate; a situation that would almost certainly be improved upon by more accurately modelling the thalamo-cortical signals. While recognising that this simplification is likely to result in a poor approximation of actual thalamic relay cell activity, it is difficult at this stage to do much better, and has the added benefit of making the simulations tractable.

For the remainder of the simulations, the acoustic signals specified are processed by the DSAM peripheral model that includes, an outer- and middle-ear transfer functions, a gamm-

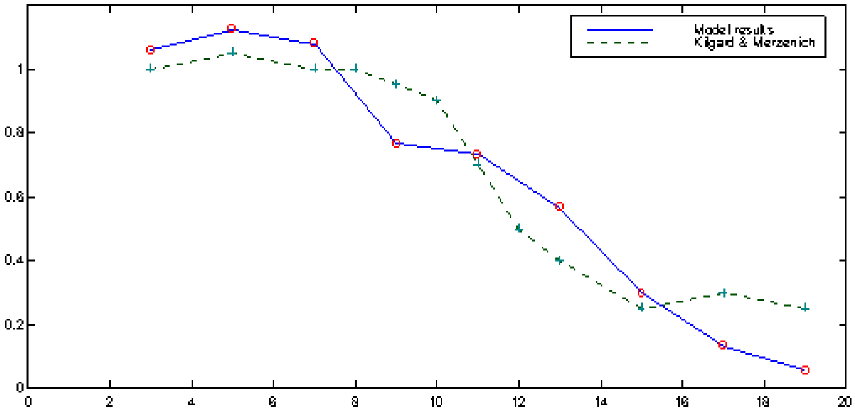


Figure 5 Response to repeated tones at the given repetition rates; model results 'o__o' and experimental results '+--+' [13]. Normalised repetition rate transfer functions are found using a stimulus consisting of 6 tones pulses at the repetition rate indicated and then calculating the mean response to the last 5 tones in the sequence divided by the response to the first tone; each tone has a duration of 25 ms.

atone filterbank, and Meddis' inner hair cell model. A simple stochastic spike generator model is used, and a convergence of twenty inner hair cells to one auditory-nerve fiber is assumed. The spike trains are then processed to ensure that refractory periods are generally greater than 20 ms. However, when more than one spike occurs simultaneously, as is possible with a combinations of 20 spike trains per channel, the refractory period is allowed to decrease in proportion to the degree of coincidence. This has the benefit of not destroying the enhanced onset response generated by the inner hair cells.

4.4 Best Modulation Frequencies

Rate-modulation transfer functions were extensively investigated by Schreiner and Urbas [24], who found that the best modulation frequencies in AI were generally below 15 Hz. More recently very similar normalized rate modulation data was presented [13]. To demonstrate the validity of the modeling approach taken Figure 5 illustrates a comparison between these experimental results and the response of the model to similar acoustic stimuli, preprocessed in the way described above. As can be seen, the model response closely replicates the experimental results.

4.5 The Time Course of Forward Masking

Although there are undoubtedly a number of factors that contribute to the phenomenon of forward masking, it is clear that the depression of thalamocortical synapses must contribute to the total effect. Explanations for forward masking have also been sought in terms of lateral or forward inhibition. However, it has been shown that masking continues to exist even in the presence of a GABA_A antagonist and therefore even if inhibitory inputs have some part to play they cannot provide a full account [3]. Both cortical forward masking and that evidenced behaviourally have been shown to last far longer than explicable in terms of peripheral adaptation [3][4][22]. The model clearly provides a mechanism for forward masking, since synapses that have been previously activated require time to replenish their transmitter stores and respond less strongly when depleted. The time course of synaptic recovery appears to be consistent with the time course of cortical forward masking. The

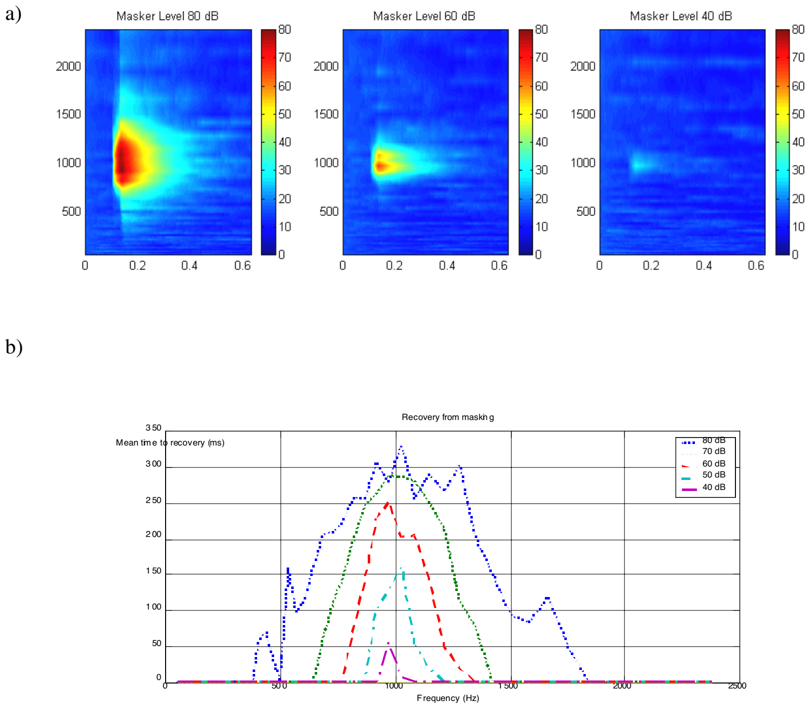


Figure 6 Distribution and time course of transmitter depletion at synapses across the tonotopic axis in response to a 1000 Hz masker of 30-ms duration at the intensities indicated. a) The color scale indicates the percentage depletion relative to transmitter levels at the start of the masker for maskers of the three intensities indicated. b) Time for transmitter to recover to within 5% of initial mean levels after masker offset for each masker intensity indicated.

tonotopic distribution of masking is also consistent with a model of forward masking in terms of the depression of thalamocortical synapses since it has been shown that masking is closely related to the receptive fields of cortical neurons [3,4]. Figure 6 shows the depletion at synapses across the tonotopic axis in response to masking stimuli at the intensities indicated. A comparison between the distribution and time course of synaptic transmitter depletion and Brosch and Schreiner's plots of the time course and distribution of masking [3], shows that there is a remarkable similarity between the two.

An important aspect of this model is that it demonstrates that cortical forward masking could be dependent on presynaptic rather than postsynaptic activity. This offers a simple explanation for the puzzling experimental observation that masking is sometimes detected even in response to maskers that do not actually activate the target cell [3]. If masking is a result of transmitter depletion of thalamocortical synapses, then it would be quite possible for such synapses to become depleted by thalamic activity even though there is insufficient incoming activity to actually cause the cortical cell to fire, which is how the response to the masker was determined [3]. Since these synapses would nevertheless be depleted, the probe tone could therefore be masked by the "sub-threshold" masker.

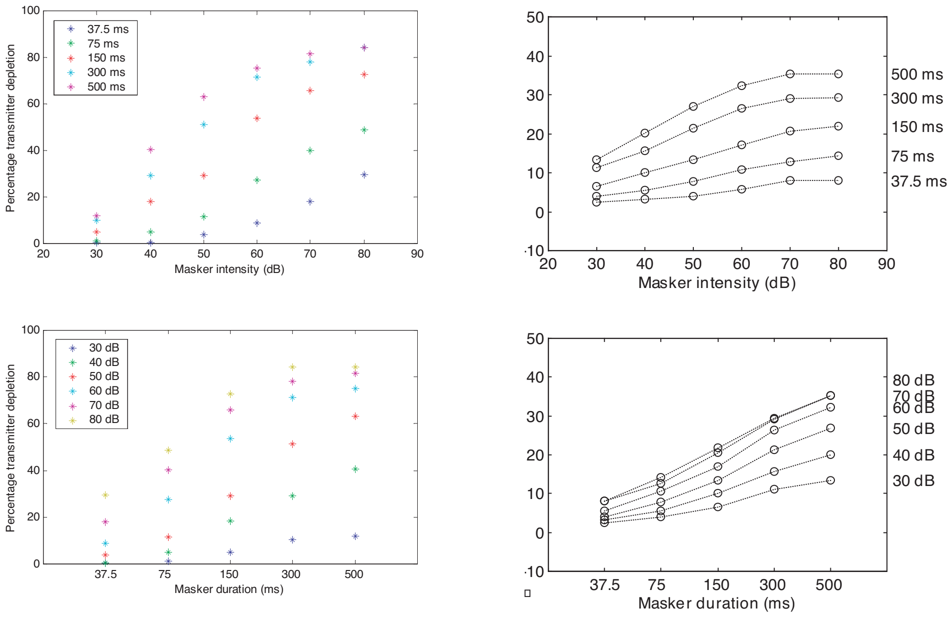


Figure 7 The effect of masker duration. Transmitter depletion relative to mean levels at the start of the masker is plotted for masker duration and intensities indicated. As can be seen the model is clearly sensitive to both masker duration and intensity, as found by Kidd and Feth [12].

4.6 *The Effect of Masker Duration on Forward Masking*

In psychophysical experiments it has been shown that the degree of masking is affected by the duration of the masker and masking increases with masker duration [12]. This was also found to be the case by Brosch and Schreiner [3] in their recordings in AI. However, the sensitivity to duration was observed even when the AI cell responded only at the onset of the masker, and although the effect of masker duration was noted, it was not suggested how this could occur. The model investigated here suggests a simple explanation — as long as there is some tonic incoming activity during the masker, then transmitter depletion at the thalamocortical synapses will be related to masker duration. Therefore, if as we hypothesize, the degree of masking is related to the degree of transmitter depletion at thalamocortical synapses, then the sensitivity to masker duration follows. The paper by Brosch and Schreiner [3] did not include any detailed results on masker duration, so in Figure 7, a comparison between Kidd and Feth's results [12] and the model's response is shown. One drawback should be noted; although these results are qualitatively the same, it is not clear how the degree of transmitter depletion in the model can be directly related to the probe threshold shifts plotted by Kidd and Feth.

4.7 *Disruption of Synchronisation Responses by Subthreshold Stimuli*

In a recent paper [20], Nelken suggested that his experiments showed a correlate of comodulation masking release. Activity was record in AI in response to noise modulated at

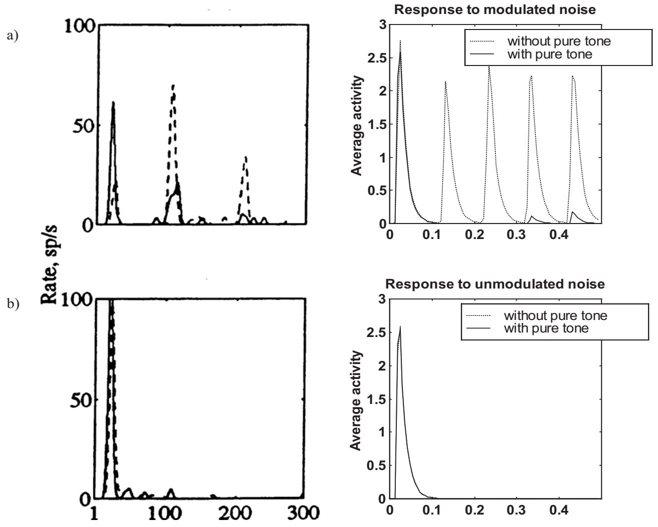


Figure 8 (a) Response of neurons in AI [20], left column, and the model, right column, to a wideband noise stimulus trapezoidally modulated at 10 Hz, without '---' and with '___' a continuous pure tone. (b) Experimental and model responses when the noise is unmodulated.

10 Hz, and was found to synchronize to each noise pulse as expected. However, when a very soft, even subthreshold, continuous pure tone with frequency corresponding to the cell's best frequency, was added to the noise, then this synchronization was disrupted. In contrast, when the pure tone was added to an unmodulated noise then the response to the noise alone was indistinguishable from that to the noise plus tone. Nelken suggested that the cortex might therefore be able to detect masked sounds by means of their disruption of the more powerful masker.

Once again a simple explanation of Nelken's results is suggested by the model, which can easily replicate the experimentally observed behaviour as long as there is some tonic thalamic activity in response to the pure tone. Because the activity in response to the pure tone continues through the silent gaps between the noise pulses, this effectively prevents the recovery of the synapses between noise pulses and so the synchronized response is disrupted. This explanation is also consistent with Nelken's unpublished observations that the synchronized response to the noise alone was far more reliably obtained when the noise was trapezoidally modulated, than when sine wave modulation was used. Figure 8 shows Nelken's experimental results and the model's responses to similar stimuli.

4.8 Onset Latency

Neurons in AI generally respond to the onset of stimuli and to transients in acoustic signals. The factors that influence the timing of the onset response are unknown, but Heil has recently published a number of papers in which the relationships between onset latency in AI and various characteristics of the stimulus envelope were investigated [9]. It was shown, for example, that for a linear rise function, the onset latency in AI was related to the rate of change of peak pressure and was independent of rise time and plateau peak pressure. In Fig-

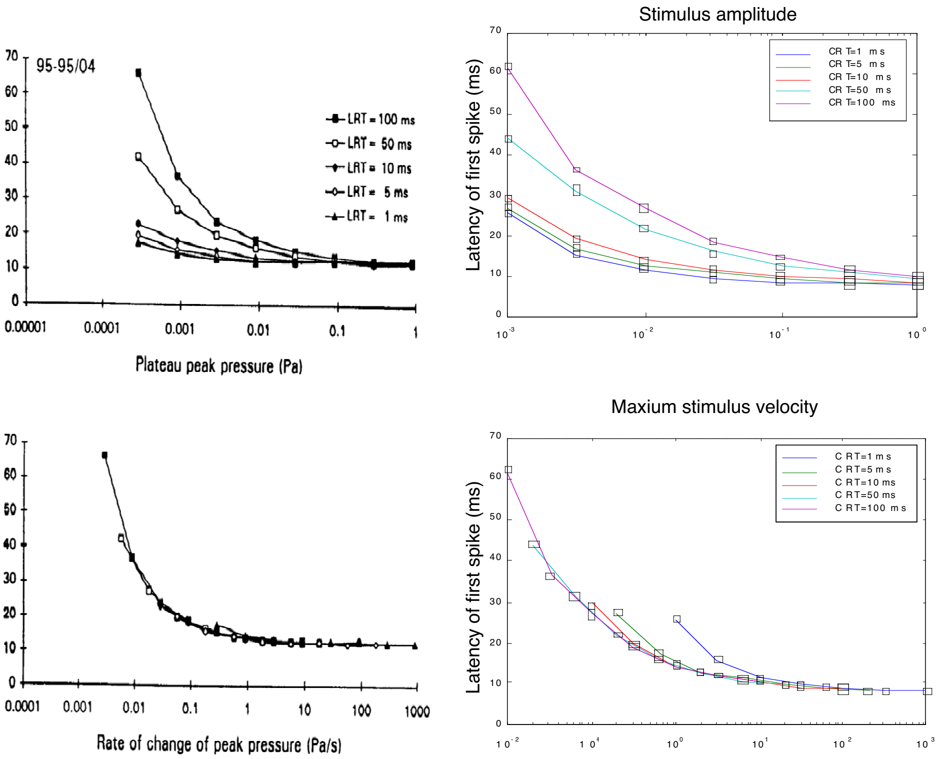


Figure 9 Onset latency for a linear rise function. The left hand column shows Heil's results [9] for this stimulus type, plotted against plateau peak pressure (top) and rate of change of peak pressure (bottom), the right hand column shows the model's response latencies for similar stimuli.

ure 9 it can be seen that the model's behaviour is very similar to that observed by Heil [9] When the latencies are plotted against rate of change of peak pressure, the latencies for different rise times superimpose quite closely. However, for stimuli that are close to the response threshold this relationship does not hold up so well — an effect also noted by Heil but not evident in the results included here. Heil also found that when a cosine-squared rise function was used the onset response latency was related to the acceleration of plateau peak pressure. The model's response to such stimuli does not replicate this result very well. However, this may be due to the simplifications made in the subcortical modeling, particularly the failure to accurately capture the enhanced onset response in the inputs used, rather than a shortcoming in the synaptic model, further work is necessary to understand the problem.

5. Discussion

In this paper it has been shown how a model neuron that incorporates dynamic synapses responds to a number of different stimuli. The results seem to indicate that synaptic depression at thalamocortical synapses may explain a number of aspects of the response properties of neurons in AI.

The nonlinearity of the dynamic synapse model allows it to behave in many situations like a low-pass filter whilst also retaining a fast onset response. In response to repeated stimulation much above 10 Hz, synaptic depletion prevents the cell from responding except at the onset of the stimulus. However, the synaptic dynamics are not slow and after a period of rest the synapse can respond with a large EPSP to the onset of a new stimulus, which can result in a response of short latency. Since the reliability of a depressing synapse also appears to be related to the amount of available transmitter [23], an aspect not included in this model, this means that after a period of rest such synapses will tend to respond very reliably as well. This is therefore consistent with the generation of onset responses of short latency and with little jitter. Although the cell tends to respond only at the onset of stimuli, important processing can continue to occur in the dendrites throughout the duration of the stimulus. This allows the cell to exhibit a sensitivity to stimulus duration, even when only responding at stimulus onset. In addition, some of the apparent nonlinearities of responses measured in AI, such as the influence of subthreshold stimuli or interaction between different components of a complex stimulus [20], could be accounted for in this way.

Since synaptic depression operates at thalamocortical synapses which are the path through which sensory signals must pass in order to get to cortex, it seems likely that the dynamics of depressing synapses have a major role to play in sensory processing. Synaptic depression appears to result in a relatively infrequent sampling of the sensory inputs by the cortex where such information is presumably integrated with ongoing cognitive processes. This bears a remarkable similarity to Viemeister's "multiple looks" model which was formulated in order to explain temporal processing in auditory perception and to resolve the resolution-integration paradox [27][28]. In his model, it was envisaged that "looks" or samples from a short time-constant process are stored in memory and can be accessed and processed selectively, depending on the task.

Another effect of synaptic depression is to greatly enhance the response to the onsets of signals. This could also act to promote grouping across frequency channels. Synaptic depression effectively provides a kind of lateral inhibition acting in the temporal domain, which may help to increase the temporal contrast of stimuli [3]. It has also been suggested that although thalamocortical sensory signals on their own cannot elicit lasting activity, facilitation at pyramidal NMDA synapses might act to enhance the response to incoming signals of interest [25]. This could provide a mechanism for the flexible processing of sensory signals, depending on factors such as previous experience or the current state of attention.

The frequency response of the model, illustrated in Figure 5B, bears a strong relationship to speech modulation transfer functions, with frequencies around 4 to 6 Hz being the dominant frequency of the envelope of speech signals. Syllables in speech are generally, although not always, distinguished by an amplitude peak preceded and closed by an amplitude trough [11]. Therefore, when the model is stimulated by a speech signal, it has a tendency to fire at the onsets of syllables within the signal. Synaptic depression may therefore give rise to a syllable-like segmentation of speech signals within AI. Such segmentation could occur in parallel across the tonotopic axis, independently within each frequency channel. This suggestion is consistent with the experimentally observed response to species-specific calls of neurons in AI, which tend to fire primarily at the onset of segments or syllables within calls, irrespective of the characteristic frequency of the neuron [5][29][30]. One effect this would have is to increase temporal synchrony across the tonotopic axis, thereby promoting the grouping of related frequency components of a call. Synchronous activity is likely to be important for the effective transmission of signals to further processing centres which integrate information across frequency channels.

In experiments in which species specific calls were manipulated [30], it was also shown that speeding up or slowing down the signal (or reversing it) all resulted in reduced responses. We suggest that the reasons for this difference between manipulations and that the behaviour of the model can help to explain these results. In the first case, when the signal is slowed down, activity is still generated in response to syllable onsets, but since these occur at a slower rate, the total amount of activity per second decreases. In the second case, when the signal is speeded up, synaptic depression would prevent synchronization to syllable onsets as effectively as for the control case. Finally, reversing the signal results in a reduced response, not because of a change in timing of the stimulus but because of the change in the nature of the transients in the signal. As shown in Section 4.8, on onset latency, sharp transients with abrupt rises are far more effective in generating responses than those with slow rise times. Reversing the speech signal means that the transients generally become less abrupt and therefore generate reduced activity. It seems reasonable to suppose that communication sounds have evolved to optimize their detection by the cortex, and that the communication sounds used are those most salient within AI – hence, the similarity between the modulation transfer functions of speech signals and those measured in auditory cortex. Interestingly, although derived very differently, the behaviour of the model is very similar to the RASTA filter developed by Hermansky and Morgan and which was found to markedly improve speech recognition in noise [10].

In general, as a result of synaptic depression, far stronger responses are likely to be evoked in AI at the onset of new sounds than at the onset of sounds which have recently be heard. This may be a useful trigger for the recognition of a new sound source in the auditory scene and could underlie Bregman's "old + new" heuristic [2] (i.e., the parts of a signals that resemble those previously encountered may be attributed to the previous sound source and the new parts which evoke a stronger response may then be processed separately). However, although it seems that primitive auditory streaming might arise in the thalamocortical system [14][15], it is not at all obvious what role synaptic depression might have in this process. While the time constants associated with synaptic depression are consistent with the time constants used in our model of auditory streaming [14], in some ways the effects of synaptic depression seem diametrically opposed to those expected to promote streaming. Although synaptic depression could support the recognition of a new sound source or stream, a stimulus that is repeatedly heard would cause less and less activity in AI. How then could a foreground stream perceptually "pop out" as occurs in streaming experiments? Clearly, much more work is required to adequately address this question.

6. Conclusions

By taking synaptic dynamics into account in modeling these experiments, it has been possible to account for a number of previously unexplained results in a fairly straightforward way. On the basis of these investigations it is suggested that the dynamics of thalamocortical synapses may help to explain the temporal integration observed in AI and in auditory perception.

The temporal response properties in the auditory system change markedly from the auditory periphery to the cortex and one reason for this might be changes in the synaptic dynamics. The synaptic model may therefore prove a useful extension to current models of auditory processing in simulating the temporal characteristics of responses recorded experimentally both in cortex and subcortically.

Acknowledgements

The author would like to thank Israel Nelken and Peter Heil for their interest in this work and for their helpful discussions and suggestions, and Michael Denham for his helpful advice during the preparation of this paper.

References

[1] Abbott, L. F., Varela, J. A., Sen, K. and Nelson, S. B. "Synaptic depression and cortical gain control." *Science,* 275: 220–224, 1997.

[2] Bregman, A. S. *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[3] Brosch, M. and Schreiner, C. E. "Time course of forward masking tuning curves in cat primary auditory cortex," *J. Neurophysiol*., 1997.

[4] Calford, M. B., Semple, M. N. "Monaural inhibition in cat auditory cortex." *J. Neurophysiol*., 73: 1876–1891, 1995.

[5] Creutzfeldt, O. Hellweg, F. C., Schreiner C. "Thalamocortical transformation of responses to complex auditory stimuli." *Exp. Brain Res*., 39, 87–104, 1980.

[6] Eccles, J. C. *The Physiology of Synapses*. New York: Academic Press, 1964.

[7] Grossberg, S. "Some physiological and biochemical consequences of psychological postulates." *Proc. Natl. Acad. Sci, USA,* 60: 758–765, 1968.

[8] Grossberg, S. "On the production and release of chemical transmitters and related topics in cellular control." *J. Theor. Biol.,* 22: 325–364, 1969.

[9] Heil, P. "Auditory cortical onset responses revisited. I. First-spike timing." *J. Neurophysiol*., 2616–2541, 1997.

[10] Hermansky, H. and Morgan, N. "RASTA processing of speech." *IEEE Trans. Speech Audio Proc*, 2(4): 578–589, 1994.

[11] Jusczyk, P. W. *The Discovery of Spoken Language*. Cambridge, MA: MIT Press, 1997.

[12] Kidd, G. and Feth, L. L. "Effects of masker duration in pure-tone forward masking." *J. Acoust. Soc. Am.,* 72: 1384–1386, 1982.

[13] Kilgard, M. P. and Merzenich, M. M. "Plasticity of temporal information processing in the primary auditory cortex." *Nature Neurosci*., 1: 727–731, 1998.

[14] McCabe, S. L. and Denham, M. J. "A model of auditory streaming." *J. Acoust. Soc. Am.,* 101: 1611–1621, 1997.

[15] McCabe, S. L. and Denham, M. J. "A thalamocortical model of auditory streaming." *Proc. IEEE Intern. Joint Conf. Neural Networks (IJCNN'98),* pp. 1541–1546,1998.

[16] McGregor, R. J. *Neural and Brain Modelling*. San Diego: Academic Press, 1989.

[17] Markram, H., Lubke, J., Frotscher, M., Sakmann, B. "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs." *Science,* 275: 213–215, 1997.

[18] Markram, H., Tsodyks, M. "Redistribution of synaptic efficacy between neocortical pyramidal neurons." *Nature*, 382: 807–810, 1996.

[19] Meddis, R. "Simulation of mechanical to neural transduction in the auditory receptor." *J. Acoust. Soc. Am.,* 79: 702–711, 1986.

[20] Nelken, I., Yosef, O. B. "Processing of complex sounds in cat primary auditory cortex." *Proc. Nato ASI on Computational Hearing*, S. Greenberg and M. Slaney (eds.), pp. 19–24, 1998.

[21] O'Mard, L. P., Hewitt, M. J. and Meddis, R. "DSAM: Development system for auditory modelling," *http://www.essex.ac.uk/psychology/hearinglab/lutear/home.html*.

[22] Relkin, E. M. and Smith, R. L. "Forward masking of the compound action potential: Thresholds for the detection of the N_1 peak." *Hear. Res.,* 53: 131–140, 1991.

[23] Reyes, A., Lujan, R., Rozov, A., Burnashev, N., Somogyi, P. and Sakmann, B. "Target-cell-specific facilitation and depression in neocortical circuits." *Nature Neurosci*., 1: 279–285, 1998.

[24] Schreiner, C. E. and Urbas, J. V. "Representation of amplitude modulation in the auditory cortex of the cat. I. Comparison between cortical fields." *Hear. Res*., 32: 49–64, 1988.

[25] Thomson, A. M. and Deuchars, J. "Temporal and spatial properties of local circuits in neocortex." *Tr. Neurosci*, 17(3): 119–126, 1994.

[26] Tsodyks, M. V. and Markram, H. "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability." *Proc. Natl. Acad. Sci, USA,* 94: 719–723, 1997.

[27] Viemeister, N. F. and Wakefield, G. H. "Temporal integration and multiple looks." *J. Acoust. Soc. Am.,* 90: 858–865, 1991.

[28] Viemeister, N. F. and Plack, C. J. "Time analysis." In *Human Psychophysics*, W. A. Yost, A. N. Popper and R. R. Fay (eds.), New York: Springer-Verlag, 1993.

[29] Wang, X. "What is the neural code of species-specific communication sounds in the auditory cortex?" *Proc. 11th Int. Symp. on Hearing*, Grantham, UK, pp. 456–462, 1997.

[30] Wang, X., Merzenich, M. M., Beitel, R. and Schreiner, C. E. "Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: Temporal and spectral characteristics." *J. Neurophysiol.,* 74: 2685–2706, 1995.

APPLYING A MODEL OF CONCURRENT VOWEL SEGREGATION TO REAL SPEECH

Georg F. Meyer, Dekun Yang and William A. Ainsworth

Centre for Human and Machine Perception Research
MacKay Institute of Communication and Neuroscience
Keele University, Keele, Staffordshire ST5 5BG, UK

1. Introduction

People are often involved in situations where more than one sound is present at the same time and are able to tolerate competing signals when listening to a target voice. This phenomenon is known as the *cocktail party effect*. One way to explain the good human performance in noisy conditions is to consider the auditory environment as a complex scene containing multiple objects and to hypothesize that the auditory system is capable of grouping these objects into separate perceptual streams based on certain primitive features identifying each object. One such feature is the fundamental frequency of voiced speech sounds [2] [15] [21] [26]. When concurrent voiced sounds are present, they can be segregated into separate perceptual streams containing parts of the auditory scene originating from a single sound source in order that the signal-to-noise ratio within each stream is improved for the subsequent speech recognition process.

The past decade has seen an explosive growth in studies of segregation of concurrent vowels. This trend has been inspired primarily by the desire to study auditory perceptual mechanisms. Double-vowel experiments are a very suitable test for models of auditory scene analysis because a large body of experimental and modelling data exists and the effects have been reproduced in many laboratories and for a number of different languages.

Auditory scene analysis is usually described as a two-stage process — a low-level, primitive, stream formation stage and a high-level, schema-based, recognition stage [4]. Double vowels fit this paradigm very well because the stimulus parameters likely to influence the primitive grouping stage, such as fundamental frequency, duration and relative amplitude are easy to control. The pattern matching stage can be kept relatively simple because of the stationary nature of the stimuli. It is possible to evaluate different stream formation strategies and grouping cues within the same framework.

Psychophysical experiments have shown that listeners use the fundamental frequency (f_0) cue to group harmonic features for vowel recognition [2] [15] [21] [26]. In these experiments subjects are asked to identify pairs of simultaneously presented vowels whose f_0s are systematically varied. It has been shown that intelligibility increases significantly within ca. a two-semitone f_0 difference (12% of the lower f_0) when the stimuli are long in duration (200 ms). For short vowels (50 ms) the intelligibility gain due to the f_0 difference is less pronounced. Also, the relative level of the two vowels in a pair can be varied over a wide range without affecting the recognition performance gain if the vowels are presented at different f_0s.

In an attempt to explain the psychophysical findings, computational models have been developed that simulate the major auditory and perceptual process by which listeners may exploit a difference in f_0 when identifying concurrent vowels. Models that aim to explain human performance have been presented previously (cf. the review in [5]). All models predict human performance as a function of f_0 difference [2] [5] [18]. The models fall into two broad categories. The first category is based on unidimensional representations of the signal spectrum. The harmonics of stationary signals are resolved and the regular spacing between harmonics can be used to recover streams, provided the f_0 can be estimated. Examples of such techniques are the harmonic-selection technique proposed in [19] and the time-domain-comb-filter model described in [5]. The harmonic-selection model is not a plausible model for human performance because harmonics in the speech pitch range are not resolved above around 1 kHz in auditory filters; thus a selection algorithm must fail. The second category of models is based on two-dimensional signal representations. An auditory filterbank usually forms one axis of this representation, while the second axis is based on signal characteristics within each channel of the filterbank. Examples of this type are the models in which autocorrelation is used to form the second axis of the signal representation [2] [18]. Such models are based on a temporal analysis of signals obtained by peripheral filtering and hair-cell transduction. The models segregate mixed vowels by grouping excitation patterns across channels based on the periodicity information in autocorrelation functions within each channel.

This chapter addresses the question of how such a streaming process might be implemented computationally for the fundamental frequency cue, as well as for evaluating its performance for natural speech signals. Section 2 presents a set of experimental results pertaining to human performance for recognizing concurrent vowels. Section 3 describes the model of concurrent vowel segregation. Section 4 evaluates the model performance on segregating concurrent vowels for real speech.

2. Human Perception

Although various experimental results on human performance in concurrent vowel identification have been reported in the literature, the experiments typically investigate the effect of a single stimulus parameter and are conducted under specific conditions that are not always maintained in experiments investigating other parameters. In this section we present a set of experimental results generated within a unified experimental framework that show human performance in recognizing concurrent vowels while manipulating the fundamental frequency, duration and relative amplitude of the vowels. The experimental results will serve as a benchmark to validate the computation model presented in the following section.

We used synthetic vowel stimuli in the experiments. The French long vowels [a,e,i,o,u,y] were synthesized using a parallel implementation of Klatt's synthesizer [9] with 16-bit resolution and a sampling rate of 20 kHz. The stimuli to be recognized were generated by adding pairs of non-identical vowels leading to fifteen possible vowel combinations. One vowel of the pair always had an f_0 of 100 Hz, the f_0 of the other vowel was set to either 100 Hz, 106 Hz, 112 Hz or 126 Hz. The relative amplitude of the constituent vowels was scaled for three level differences between the signals: 0 dB, 6 dB and 12 dB. The signals were windowed for three signal durations of 51.2 ms, 102.4 ms and 204.8 ms. Half-Hanning windows were applied to the initial and final 25.6 ms of each stimulus.

Four subjects were presented with the vowel pairs and asked to identify the pair heard. The signals were played diotically via Beyer Dynamic DT660 headphones at ca. 54 dB SPL. All experiments were conducted in a sound-proof room. Each subject performed an initial training session where single vowels were played over the full f_0 range. After the familiarization session, an experiment using vowels with 0 dB relative amplitude was run. This ses-

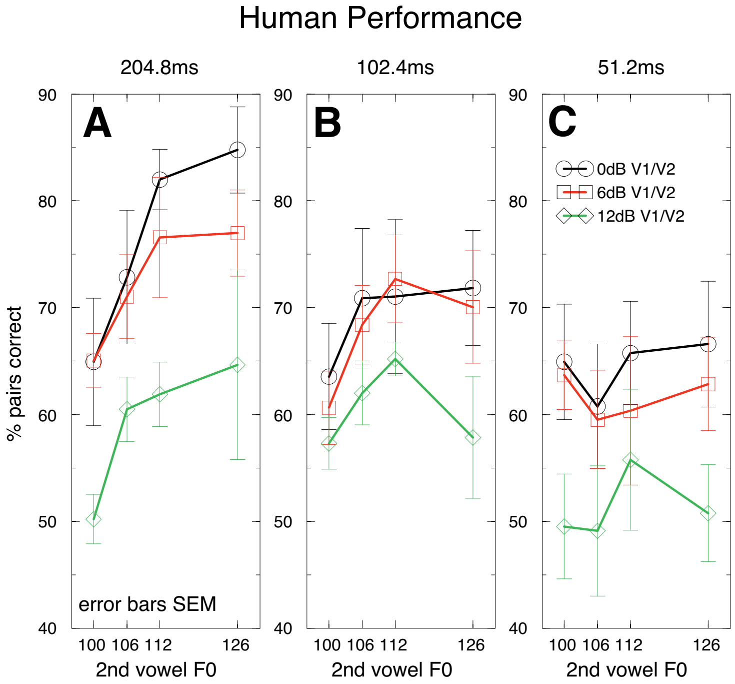


Figure 1 Subject performance for the double-vowel recognition task. Pairwise recognition is plotted in each graph for three relative level settings, 0 dB, 6 dB and 12 dB. One vowel is always presented at 100 Hz f_0. The f_0 of the second vowel is shown on the graph. Panel A shows recognition performance for the 204.8-ms stimuli, panel B is for the 102.4-ms stimuli while panel C shows data for the 51.2-ms signals.

sion is not included in the data, as all subjects showed improvement over the first session. Each session contained a quasi-random sequence of 360 combinations of the 15 possible vowel pairs, the three possible levels of relative amplitudes and four possible f_0 differences. Stimulus durations were constant within each experiment.

Subject performance is shown in Figure 1. Panel A shows the data for the 204.8-ms stimulus duration, panel B is based on the 102.4-ms stimuli and panel C is for the 51.2 ms signals. Each panel shows the percentage of correctly recognized pairs against f_0 of the second vowel for each relative amplitude level. If 204.8 ms segments are played, subject performance improves from 65% of pairs correct to 85% correct if both vowels have the same RMS amplitude. As the level difference between the vowels increases, overall performance is reduced, but comparable performance increases are observed (12.5% at 6 dB, V1/V2 and 15% at 12 dB, V1/V2). An analysis of variance shows that the f_0 difference ($F = 12.81$, $p < 0.001$) and the relative amplitude ($F = 29.41$, $p < 0.001$) are significant factors. Figure 1B shows responses to the 102.4 ms stimuli. The improvements with f_0 are no longer very pronounced, but still significant ($F = 2.78$, $p = 0.045$). The effect of relative amplitude is clearly visible and significant ($F = 3.92$, $p = 0.023$). For very short duration signals (Figure 1C) no effect of f_0 difference is visible in the data ($F = 0.44$, $p = 0.72$), while the relative signal level is significant ($F = 6.05$, $p = 0.03$). The recognition scores for stimuli where both vowels have the same fundamental frequency (100 Hz, 2nd vowel) lie between 50% and 65%, depending on the relative amplitude. In the absence of segregation cues, the data suggest that listeners use an independent, high-level recognition mechanism. Zwicker [26] suggests a spectral subtraction process where listeners recognize a dominant vowel first and then "mentally subtract" its spectrum from the mixture to yield a second vowel.

The experimental results confirm that the f_0 difference is a powerful grouping cue for vowel segregation, provided that the stimuli are at least 100 ms in duration. For shorter signals no significant performance improvements are seen in our data. This is consistent with the findings of Assmann and Summerfield [2]. Small performance increases, even for 50-ms duration data, are described in [3]. In the 200-ms and 100-ms conditions, a sharp rise over the first 2 semitones in f_0 difference is observed. This is also consistent with previous data and suggests that some form of fundamental frequency analysis is carried out. If level differences between the vowels are introduced, subject performance is reduced, but the reduction in performance appears to be independent of the effects introduced by the f_0 difference manipulations. Where no grouping cues are present, subjects still achieve average recognition rates that are well above chance level. When both vowels are presented at the same amplitude, subjects recognize 65.8% of all the pairs (chance performance = 6.67%). A spectral subtraction mechanism, as proposed by Zwicker [26], is a plausible explanation, but since only the relative amplitude, not the pitch difference or signal duration alters the recognition rate where both f_0s are equal, the proposed model does not consider this aspect of the data.

3. The AM-Map Model

In this section we describe the amplitude-modulation-map-based model for segregating concurrent vowels. The model generates an amplitude modulation (AM) map from an input speech signal through four stages:

(1) peripheral filtering using an auditory filterbank,
(2) hair-cell transduction via half-wave rectification to extract the AM excitation patterns of the speech signal,
(3) bandpass filtering to remove high-frequency components, and
(4) spectral analysis of the AM excitation patterns.

The model uses a 2-D map to represent AM components of the speech signals filtered by an auditory filterbank. Provided that the f_0s are known, segregation is achieved by grouping signal components with common modulation frequencies in the channels.

Autocorrelation-based models rely on the observation that, while a representation based on the average discharge rate in an auditory filterbank is unable to resolve speech harmonics in the high-frequency range directly, precise temporal information is present in the discharge pattern seen in each channel. A secondary processing step can therefore be used to generate a representation relying on the signal fine-time structure. Autocorrelation analysis of the pattern in each channel highlights periodicities linked to the signal fundamental frequency [14]. This representation has been used, with some success, for the separation of concurrent vowels [2] [18] [23].

Autocorrelation analysis has two major computational drawbacks — it is an inherently non-linear operation, and the energy in the resultant representation is not localized on the perceptually relevant f_0 axis. In practice this means that it is not possible to recover spectra directly from the representation but that the analysis is used to compute the dominant pitch in each of the filterbank channels. The segregation process then groups all channels with common periodicity into perceptual streams. All energy within each channel is attributed to

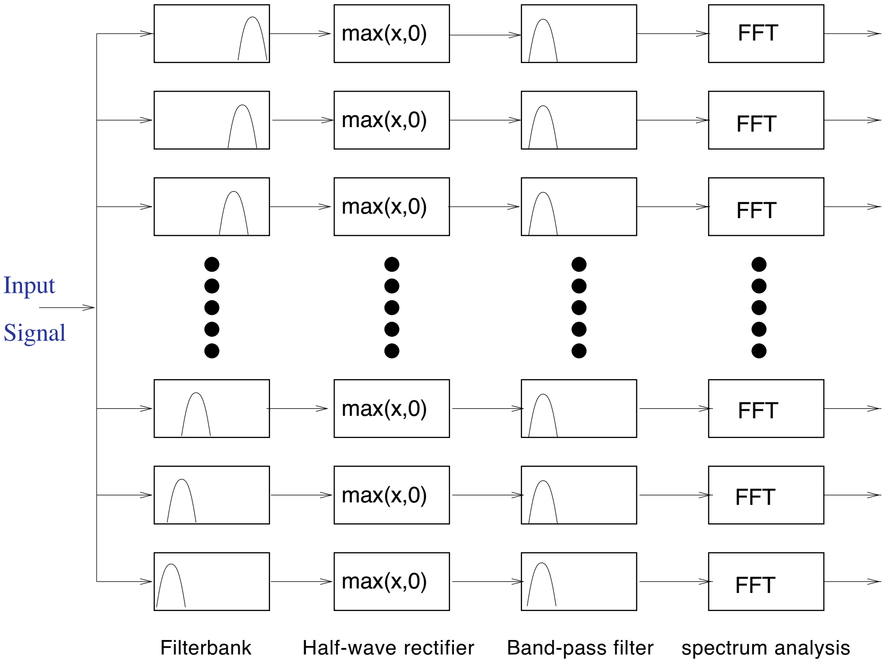


Figure 2 Schematic diagram of the AM-map model. An AM map is computed through four stages: (1) peripheral filtering via an auditory filterbank to perform cochlear frequency analysis, (2) half-wave rectification to perform hair-cell transduction, (3) band-pass filtering to remove the high-frequency components, and (4) spectral analysis, via the Fourier transform, to extract modulation frequencies.

one source. Because of the broad spectral distribution of energy in speech sounds this "all or nothing" segregation means that spectra recovered using the autocorrelation method are necessarily distorted. The modulation map algorithm we propose does not have these drawbacks though, in contrast to autocorrelation analysis, requires long analysis windows to segregate concurrent voiced sounds.

Human listeners also need relatively long stimulus durations to successfully segregate concurrent vowels. The autocorrelation model proposed by Meddis and Hewitt [18], which builds on the model proposed by Assmann and Summerfield [2], uses an exponentially decaying analysis window with a duration of 30 ms. This short interval is appropriate considering that the representation derived by autocorrelation analysis for periodic signals does not change significantly after a full period of the lowest frequency component in the signal has been processed. It also means that the autocorrelation models, as they currently stand, are unable to account for the improvement in performance as signal duration is extended from 50 to 200 ms. This, however, does not invalidate autocorrelation analysis as a model for human performance because the models could easily be extended to account for human data by adding a long-duration integration stage at the identification level.

We propose an alternative to autocorrelation models, based on a spectral representation of the discharge pattern observed in each channel. The alternative representation proposed for speech segregation is similar to the modulation spectra proposed by Kollmeier and Koch [10] but uses different parameters and does not attempt to use binaural information. This type of representation, a map of channel characteristic frequency against modulation frequency, has been demonstrated physiologically at the level of the inferior colliculus [13] [22].

If a signal is passed though an auditory filterbank, then the discharge pattern in each simulated nerve fiber encodes the average energy in the channel as both average discharge rate and as fine timing information (cf. [8] [12] for a review of time coding in the auditory system). If two objects in an auditory scene have different modulation frequencies such as, for instance, two simultaneous vowels with different fundamental frequencies, then this envelope information can be used to segregate the sources. The processing steps involved are discussed in detail below.

The first stage in the information processing in the model is an auditory filterbank. The signal is split into thirty-two 0.5-Bark-spaced channels with characteristic frequencies ranging between 0.1 kHz and 4.7 kHz (Figure 2). Each filter is a linear, fourth-order recursive gammatone filter [6] [11]. The output of each channel is scaled to approximate human hearing thresholds.

An important component of any auditory model is a model of hair-cell transduction. An example for such a model is the Meddis hair cell model [16][17], which models the three main effects observed in hair cell transduction: a non-linear compression with adaptation, half-wave rectification and low-pass filtering of the input signal. The rectification and filtering actions of the model are critical to an amplitude demodulation system. However, the log-compression negatively affects the extraction of AM components at moderate to high signal levels. The model proposed in this chapter includes an explicit half-wave rectification and filtering stage, but does not include any further nonlinearity.

The carrier frequency in each channel is removed by low-pass filtering the half-wave rectified signal. The filter used is a first-order, low-pass filter. The resulting signal has a non-zero mean, as a consequence of the half-wave rectification. Before computing the Fourier transform this mean, as well as any low-frequency beats, is removed by a second, high-pass filter. This filter is also implemented as a recursive, first-order filter. Filter time constants of $T_l = 2$ ms (low-pass filter) and $T_h = 4$ ms (high-pass filter) are used.

After the rectification and filtering stages have been applied to each channel, a windowed time slice is applied and a Fourier analysis carried out on the output of each channel. The amplitude spectrum is plotted along the abscissa for each of the 32 channels in the system.

Strictly speaking, the process is not exactly equivalent to envelope extraction where the filterbank is able to resolve single harmonics. For these channels only the harmonic closest to the characteristic frequency of the filter is observed in the modulation spectrum. As auditory filters widen with characteristic frequency this limitation only applies for very low characteristic frequencies. The modulation spectrum of these channels shows these harmonics because the low-pass filter only cuts off above 300 Hz. This a mixture of resolved harmonics at low frequencies and envelope information at higher frequencies means that the Hilbert envelope is not suitable for the construction of modulation maps at low channel frequencies.

The map shows energy in the modulation spectrum for each channel in the auditory model. Carrier and envelope frequency for each object in the scene can be read off the two axes. Spectra can be recovered by sampling the map along the target f_0.

If a single voiced speech sound is used to drive the model, a characteristic striped pattern appears. Energy is localized in ridges corresponding to the fundamental frequency and its harmonics. If the energy in all ridges is summed the vowel spectrum is recovered. As a consequence of the demodulation stage, energy is localized primarily within the partial spectra, located at modulation frequencies corresponding to the first five harmonics of the signal. A

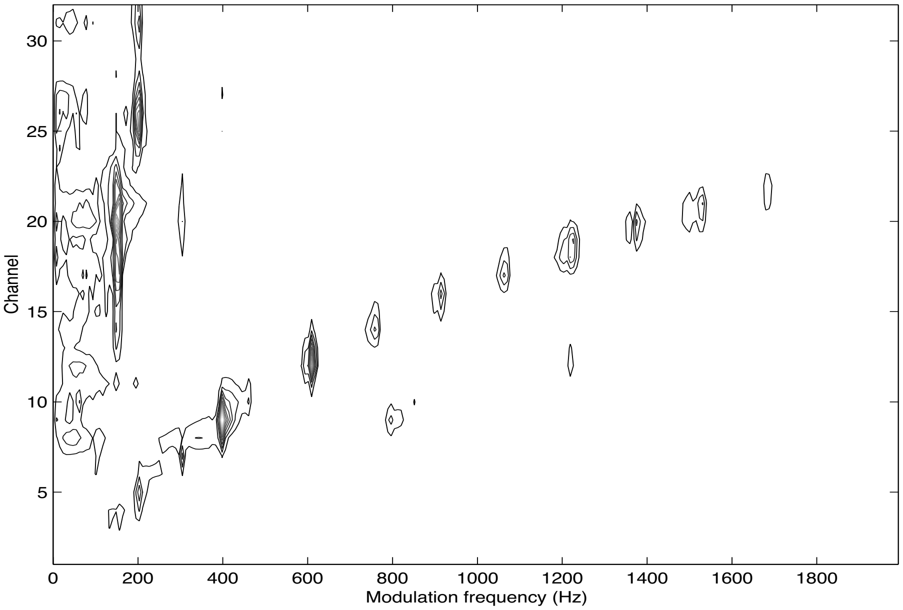


Figure 3 A contour plot of an AM map for the concurrent vowels [er] with an f_0 of 152 Hz and [iy] with an f_0 of 200 Hz, where spectral analysis is performed by a conventional DFT with a 128-ms Hanning window. The AM information of vowels is well encoded as the harmonic ridges in the AM maps.

key feature of the maps is that the representation is sparse. Additional vowels, provided their fundamental frequencies are different, can be accommodated with little spectral overlap. Figure 3 shows a contour plot of an AM map for the concurrent vowels [er] with an f_0 of 152 Hz, and [iy] with an f_0 of 200 Hz (spectral analysis is performed by a conventional DFT with a window of 128 ms). We observe that AM information of vowels is well encoded as the harmonic ridges in the AM maps.

The AM representation is used to model the major auditory and perceptual processes by which listeners exploit a difference in f_0 when identifying the constituents of double vowels. When a concurrent vowel pair with different f_0s of the constituent vowels is fed to the segregation model, the AM map shows the harmonic ridges corresponding to the harmonics of the two f_0s. Provided that the resolution of modulation frequency is sufficient to localize the harmonic ridges in the AM map, segregation can be achieved in two stages. The first stage is to group the harmonic ridges corresponding to the fundamental of the target sound. The second stage is to sum the grouped harmonic ridges to recover the vocalic spectrum. Because of the sparse and well-localized distribution of energy in the AM map, spectra of concurrent voiced speech sounds can be extracted with minimal distortion by exploiting f_0 information. This is the main advantage of the AM-map representation over autocorrelation-based processes.

4. Model Performance on Synthetic Data

Amplitude modulation maps are used to model f_0-based streaming. To this end, the perceptual experiments were repeated, using the model rather than human listeners. Spectra were extracted at the two, known (rather than estimated) pitches. Each of the extracted spectra was compared against a set of templates obtained by driving the maps with isolated vowels and extracting the resulting spectra. Cross-correlation coefficients, taking the full spectra (not the peak positions as in [2]), were computed. Each extracted spectrum was compared against a set of templates and the template leading to the highest correlation coefficient was taken to identify the recognized vowel. More complex and perceptually relevant descriptors of the extracted spectra could be used, but this "minimalist approach" makes the process more transparent.

The system recognition performance was tested on stimuli used in the experiments with human listeners. Model performance was evaluated for three stimulus durations (51.2, 102.4 and 204.8 ms) for each of three relative levels (0 dB, 6 dB and 12 dB, V1/V2). Both spectra were recovered by extracting the initial five partial spectra from the map using the known, rather than the estimated f_0.

Pattern matching was performed against templates obtained by averaging the place representation for isolated vowels with fundamental frequencies ranging between 100 and 200 Hz, in 5-Hz steps. The cross-correlation coefficient was computed for all templates and the template with the correlation highest to the recovered spectrum was chosen as the recognized vowel.

The model predicts qualitative effects observed in human recognition performance data reasonably well. The model performance on the synthetic data is illustrated in Figure 4. One of the most striking differences is that without a fundamental frequency difference the model predicts none of the pairs correctly. This is not surprising because the model is a pure segregation model. If no segregation cues are present it has to fail because, in this case, both streams contain the same data. In essence, the model recognizes the same vowel twice.

For long (204.8 ms) stimuli, recognition performance rapidly rises to 100% for 0 dB V1/V2. The relative increase in performance as a function of f_0 difference shows trends similar to the human data. The model performance for large f_0 difference values is about 10% higher than the performance of human listeners.

If the signal duration is reduced to 102.4 ms, the pattern is similar, but overall performance levels reduce. The effect of the manipulation of the relative amplitude of the vowels is more pronounced when the two vowels are well segregated: 112 Hz and 126 Hz at 204.8 ms and 126 Hz at 102.4 ms.

For very short signal durations the model performance shows a gradual increase in performance while subject data show no significant f_0 effects. The relative level manipulation also does not show a clear effect.

Human data suggests that some high-level process, which allows the recognition of simultaneous vowels without segregation cues is used. The model performance never exceeds 60% of the pairs correct in the 50-ms duration condition, which is roughly the human baseline performance. If an independent high-level process is assumed, then performance increases due to the segregation process that would only be visible once they exceeded this baseline performance. In the 51.2-ms signal case, visible increases in performance would not be expected. The increase in performance is consistent with perceptual data reported by McKeown [15].

5. Model Performance on Real Speech Data

While synthetic vowels are appropriate stimuli to use in perception and modelling experiments, such signals do not reflect the spectral and temporal variability of real speech. One is confronted with difficulties when dealing with real speech signals. First, since real speech signals are non-stationary the vowels under study are of short duration which limits the resolution of the vowel spectra. Second, the variation of vowel spectra is an inherent problem of

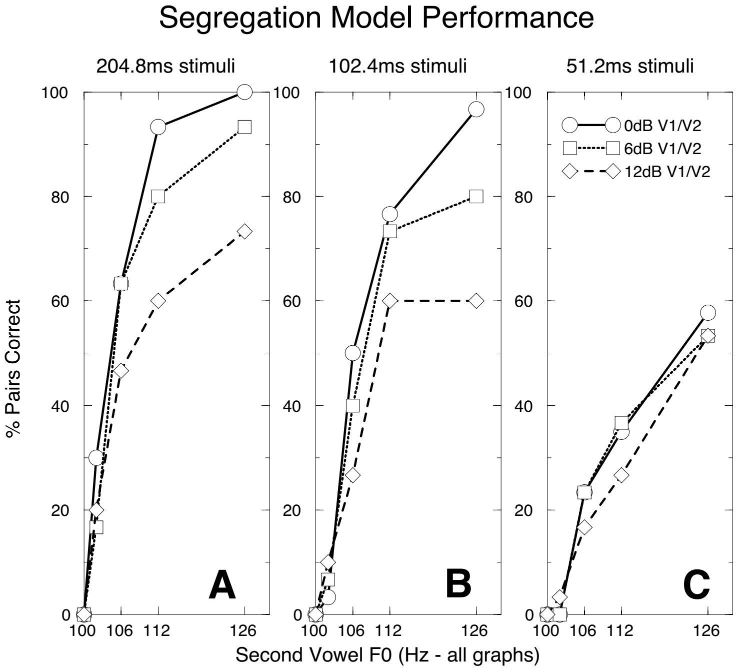


Figure 4 Pairwise recognition performance of the segregation system based on modulation maps. Panel A shows recognition rates (in percent) against the second vowel f_0 for 0 dB, 6 dB and 12 dB relative rms levels. Panel B and C show the same data for 102.4 ms (B) and 51.2 ms (C). Recognition performance increases in all cases as f_0 differences are introduced, but the slope decreases as the duration is reduced. The relative level has a larger effect when the vowels are well separated (i.e. at high f_0 differences for the 204.8 ms and 102.4 ms conditions). If both vowels have the same fundamental frequency, the pairwise recognition fails because the algorithm is unable to segregate the vowels into separate streams.

real speech. Therefore, it is necessary to investigate the performance of models under more realistic speech conditions.

To date, most computational models have been evaluated only on synthetic vowels and little work has been devoted to the investigation of segregating concurrent vowels extracted from real speech. As the emphasis of the research carried out to date has been on the perceptual basis rather than on the computational aspect of auditory scene analysis, it is understandable that the development of the enabling techniques to make this approach a practical proposition has received less attention. However, it is desirable to develop computational models which can work for real speech processing. In this section we present experimental results for evaluating the f_0-guided segregation model described in the previous section.

The evaluation was performed on a real speech database known as the TIMIT corpus. The reason for choosing this corpus lies in its popularity as a phonetically rich, speaker-independent, real speech database that is used extensively in the speech-processing community. We consider the problem of segregating and recognizing concurrent vowels whose constituent members are the five long vowels [aa], [iy], [ae], [oy] and [uw]. All vowels among the five classes were extracted from a total of 3,536 sentences spoken by 442 different speakers. Each vowel was 128 ms in duration. Concurrent vowels are generated by mixing randomly selected pairs of vowels from the extracted signals.

We carried out two experiments to evaluate the performance of the model. The first experiment was designed to measure the capability of recovering vowel spectra based on the AM representation. The second experiment tested segregation performance by measuring the recognition rate of the segregated vowels. Since there is no "ground truth" for f_0 information available for real speech analysis it is difficult to assess the accuracy of f_0 estimation. Bearing in mind that our main aim is to investigate the validity of the segregation model, we adopted the following strategy to minimize the influence of f_0 estimation in our experiments. We estimated f_0 from the isolated vowels before mixing them to generate concurrent vowels. We then assigned the f_0s of mixed vowels by the f_0s obtained from the isolated constituent vowels. We used a linear-prediction-based method [24] to determine the f_0s of the isolated vowels extracted from the TIMIT database. The method adopts the sinusoidal model for speech waveforms and parameterizes the amplitude and frequencies of the sinusoidal components using a high-order, linear-prediction model. The combination of the sinusoidal model and linear-prediction technique leads to a high-resolution, linear-prediction spectrum in which the f_0 can be reliably determined from the harmonic peaks. It was found in the experiments that the f_0 estimation is reasonably good for the segregation.

In the first experiment 3,000 pairs of concurrent vowels were extracted from the TIMIT corpus and used for the evaluation. We fed the vowels to the model and recovered the vowel spectra using the AM-map representation. If the segregation model is perfect, then the spectrum recovered from the mixed vowels should be the same as that extracted from the corresponding isolated vowels because the same f_0 is involved in the processes. Figure 5 shows the comparison of spectra obtained from mixed vowels and isolated vowels. We can see the difference between the spectrum extracted from the mixed vowels and that from the isolated vowels. To quantify the difference we calculated the overall spectral difference over all channels and normalized it by the isolated spectrum. The average difference over the 3,000 pairs of concurrent vowels was 11.5%.

The spectral distortion caused by segregation can be attributed to two factors: (1) the interaction between constituents of mixed vowels and (2) the limited resolution of the spectrum. First, the interaction between constituents of mixed vowels is an inherent problem associated with the model based on the AM representation. When a concurrent vowel is present, unwanted AM components may emerge due to the beating of harmonics from different constituent vowels. In the experiment we found that the unwanted AM components are small and have little influence in segregation, especially when two constituent vowels belong to different vowel classes. Second, the limited resolution of the spectrum is a common problem in real speech analysis because the duration of real vowels is short. In order to overcome this difficulty it is necessary to employ some spectral analysis techniques to replace the conventional Fourier transform in the AM representation.

In the second experiment the recognition performance of the segregated vowels was evaluated. As in the first experiment, we tested 3,000 pairs of concurrent vowels. Constituent vowels were recovered based on the segregation model. Vowel recognition is performed via vowel classification. Linear discriminant analysis (LDA) [7] was used to achieve the vowel classification. The appeal of LDA for vowel classification comes from its capability to accommodate the spectral variation of vowels within classes. Spectra of real vowels vary depending on the context (e.g., the word with which the vowels are associated). It is desirable to accommodate the variation in order to achieve reliable vowel classification. LDA deals with the variation by reducing the dimensionality via linear projection. The linear projection is chosen in such a way that the ratio of the between-class scatter and the within-class

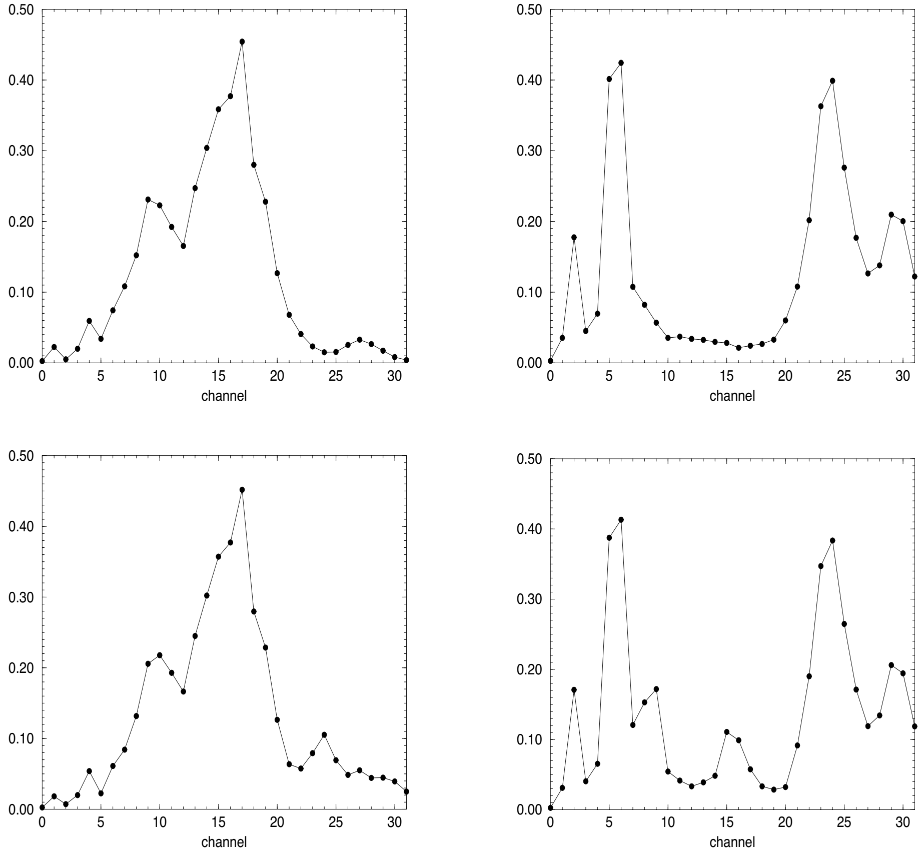


Figure 5 Comparison of the spectrum of isolated vowels and the spectrum of the segregated vowels obtained using the AM map model. Panel top-left is a spectrum of an isolated vowel [er]. The top-right panel is a spectrum of an isolated vowel [iy]. Concurrent vowels are generated by adding these two vowels with equal power. The panel, bottom-left, is the spectrum of the segregated vowel [er]. The panel, bottom-right, is the spectrum of the segregated vowel [i].

scatter is maximized. The dimension of the reduced space is N-1, where N is the number of classes. In the experiment we determined the projection matrix using all available vowels extracted from the TIMIT corpus as the training data. Segregated vowels were then transferred into the reduced feature space using the linear projection matrix. Vowel classification is performed in the reduced feature space based on a Euclidean distance measure. That is, each vowel is assigned to the class by which the Euclidean distance between the vowel and the template is minimized.

In the experiment the segregation performance was measured by the recognition rate of the target vowel, which is defined as the constituent vowel having a relatively higher power. Figure 6 shows the recognition performance of the segregation model based on the AM map. The concurrent vowels were 128 ms in duration. A conventional FFT was used to compute the AM maps. The target-to-interferer ratio varied from 0 to 12 dB. In each condition three trials were carried out in which 3,000 concurrent vowels were used. The target vowel recognition rate is the average rate over all six classes. For comparison, the target vowel recognition was performed without segregation and the results are also plotted in Figure 6. It can be

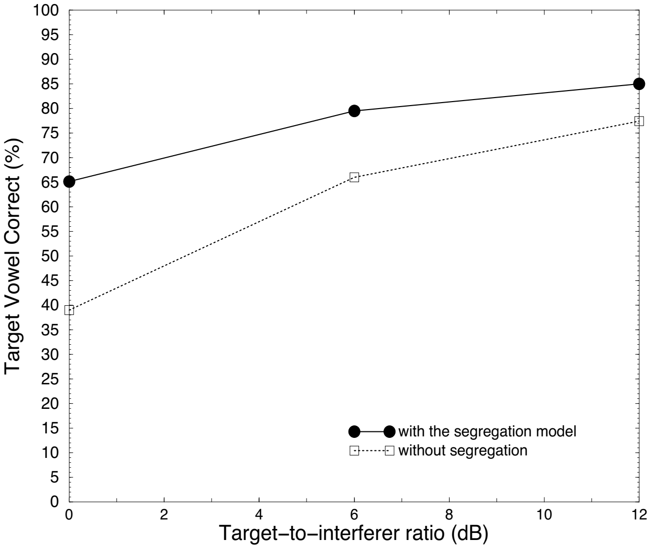


Figure 6 Recognition performance of the segregation model based on the AM map. The recognition rate of the target vowel is plotted against the target-to-interferer ratios (solid line). For comparison the recognition performance without segregation is also plotted (dotted line). The concurrent vowels were 128 ms in duration. A conventional FFT was used to compute the AM maps.

observed that the segregation based on the AM map produced better target vowel recognition.

We can see from the two experiments that the performance of the model depends largely on the resolution of the AM map. Since the duration of speech signals is limited, the only way to increase the resolution of the AM map is to apply some advanced spectral analysis techniques for attaining high-resolution AM representation. The zero-padding mechanism is not effective in this respect because it only performs the interpolation operation in the frequency domain. It has recently been shown that the reassigned spectrum technique can be used to improve the resolution of AM representation for segregating synthetic concurrent vowels [20], which provides a convenient means with which to overcome the resolution problem.

The experimental results show that the proposed model is robust, which makes it very suitable for engineering applications. In contrast to Parsons' harmonic-selection model [19], the proposed model does not have to resolve high-frequency harmonics. In practice, it is difficult to resolve high-frequency harmonics because small f_0 estimation errors produce additive effects for high-frequency harmonics and the smearing of spectral information occurs where the signal's f_0 changes within the analysis window. Therefore, the proposed model is expected to outperform Parsons's model for real speech. The detailed comparison of these two models is one of our on-going projects.

6. Conclusions

In this chapter we have proposed a signal representation and segregation model that predicts human ability to recognize concurrent vowels as the f_0 difference, relative amplitude and duration are varied. The model predicts human performance data qualitatively for the three parameters. For 100- and 200-ms stimuli the introduction of a two-semitone fundamental frequency difference leads to a significant increase in recognition performance. The relative level of the two component vowels in each pair affects recognition performance. Even at a 12-dB relative level both human listeners and the model show a significant increase in performance with the introduction of small f_0 differences. It is worth noting that the f_0-guided segregation model alone cannot explain human performance for stimuli where both vowels have the same fundamental frequency.

The model is intended as an abstract approximation to the signal representations observed in the inferior colliculus [13] [22]. The representation is computed by applying a Fourier transform to a window of activity seen in an auditory filter. While it is clear that the proposed algorithm is not, in any way, physiologically plausible, it is important to note that the model is a close analogue to the comb-filter model proposed by Cheveigné [5]. The use of the Fourier transform allows high-frequency resolution which comb filters can only achieve if multiple filter stages are cascaded. The time-frequency trade-off inherent in the Fourier transform accurately simulates the reduction in human performance observed as stimulus duration is reduced.

The model relies on the signal fine-time structure coded in a bank of auditory filter channels. This information is lost in low-threshold, low-dynamic range auditory-nerve responses because such fibers lock onto the dominant formant frequency [25]. Dynamic range restrictions mean that the envelope of the signal is not coded at moderate-to-high sound pressure levels. A model that critically depends on envelope information cannot therefore be driven directly by populations of low-dynamic-range nerve fibers. Amplitude-modulation maps have been demonstrated at the level of the inferior colliculus [13] [22] and the auditory cortex. High-threshold, auditory-nerve fibers, or onset cells in the cochlear nucleus, have both been shown to code envelope information even at moderate to high sound pressure levels and could form the basis of the proposed scheme [1]. The proposed segregation model requires a long duration analysis window to achieve sufficient frequency resolution in the modulation domain. For very short signal durations a spectral subtraction model may mask the (small) performance gains introduced by the segregation model.

A very significant aspect of the model, in our view, is that the object of the modelling is the underlying signal representation, where the segregation of signal sources with different f_0s is an inherent feature of the representation. The constraints imposed by the algorithm, most notably the time-frequency trade-off inherent in any frequency analysis, cause the model to reflect human performance data without the need for any secondary system or a more detailed explanation.

The segregation performance of the model was evaluated on real speech signals using the TIMIT database. Experimental results show that the model can provide an engineering solution to segregating real vowels from concurrent vowel backgrounds. Based on the model, the spectra of constituent vowels can be recovered from the mixed vowels with a small amount of spectral distortion. A reasonably good target vowel recognition rate can be obtained under various target-to-interferer ratios. Compared with the recognition without segregation, the model significantly improves the recognition performance of the target vowels. The segregation performance of the model largely depends on the resolution of the AM map. By incorporating some advanced spectral analysis techniques into the model, further improvement of segregation performance can be expected.

Acknowledgments

This research was supported by the Science and Engineering Research Council, UK, under the EPSRC Grant GR/L05655.

References

[1] Ainsworth, W. A. and Meyer, G. F. “Recognition of plosive syllables in noise: Comparison of an auditory model with human performance.” *J. Acoust. Soc. Am.*, 96: 687–694, 1996.

[2] Assmann, P. F. and Summerfield, A. Q. “Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies.” *J. Acoust. Soc. Am.*, 88: 680–697, 1990.

[3] Assmann, P. F. and Summerfield, A. Q. “The contribution of waveform interactions of the perception of concurrent vowels.” *J. Acoust. Soc. Am.*, 95: 471–484, 1994.

[4] Bregman, A. S. *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[5] Cheveigné, A. de “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time domain cancellation model of auditory processing,” *J. Acoust. Soc. Am.*, 93: 3271–3290, 1993.

[6] Darling, A. M. “Properties and implementation of the gammatone filter: A tutorial.” *Work in Progress 5*, Dept. of Phonetics and Linguistics, University College, London, 1991.

[7] Duda, R. and Hart, P. *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience, 1973.

[8] Evans, E. F. “Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons.” *Audiology*, 17: 369–420, 1978.

[9] Klatt, D. H. “Software for a cascade/parallel formant synthesizer.” *J. Acoust. Soc. Am.*, 67: 838–844, 1980.

[10] Kollmeier, B. and Koch, R. “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.” *J. Acoust. Soc. Am.*, 95: 1593–1602, 1994.

[11] Kortekaas, R. W. L. and Meyer, G. F. *Vowel Onset Detection Using Models’ of the Auditory Periphery and the Nucleus Cochlearis: Physiological Background*. Institute for Perception Research, TU Eindhoven, Rapport 963, 1994.

[12] Langner, G. “Periodicity coding in the auditory system.” *Hearing Res*, 60: 115–142, 1992.

[13] Langner, G. and Schreiner, C. E. “Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms.” *J. Neurophysiol.*, 60: 1799–1822, 1988.

[14] Licklider, J. C. R. “Three auditory theories.” In *Psychology: A Study of a Science (Study I. Conceptual and Semantic, Vol 1: Sensory, Perceptual and Physiological Formulation),* New York: McGraw-Hill, *pp*. 41–144,1959.

[15] McKeown, J. D. “Perception of concurrent vowels: The effect of varying their relative level.” *Speech Comm.*, 11: 1–13, 1992.

[16] Meddis, R. “Simulation of the mechanical to neural transduction in the auditory receptor.” *J. Acoust. Soc. Am.* 79: 702–711, 1984.

[17] Meddis, R. “Simulation of auditory-neural transduction: further studies.” *J. Acoust. Soc. Am.*, 83: 1056–1063, 1986.

[18] Meddis, R. and Hewitt, M. J. “Modelling the identification of concurrent vowels with different fundamental frequencies.” *J. Acoust. Soc. Am.*, 91: 233–24, 1992.

[19] Parsons, T. W. “Separation of speech from interfering speech by means of harmonic selection.” *J. Acoust. Soc. Am.*, 60: 911–918, 1976.

[20] Plante, F, Meyer, G., Ainsworth, W. A, “Improvement of speech spectrogram accuracy by the method of reassignment.” *IEEE Trans. Speech Audio Proc.*, 6: 282–286, 1998.

[21] Scheffers, M. T. M. *Shifting Vowels: Auditory Pitch Analysis and Sound Segregation*. Ph.D. Thesis, Groningen University, 1983.

[22] Schreiner, C. E. and Langner, G. “Periodicity coding in the inferior colliculus of the cat. II. Topographical organization.” *J. Neurophysiol.*, 60: 1823–1840, 1988.

[23] Weintraub, M. “A computational model for separating two simultaneous talkers.” *Proc. IEEE ICASSP 86*: 81–84, 1986.

[24] Yang, D., Meyer, G. F. and Ainsworth W. A. “Pitch analysis of concurrent speech.” *Proc. Inst. Acoust.* 20: 163–170, 1998.

[25] Young, E. D. and Sachs, M. B. “Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers.” *J. Acoust. Soc. Am., 66:* 1381–1403.

[26] Zwicker, U. T. “Auditory recognition of diotic and dichotic vowel pairs.” *Speech Comm.*, 3: 365–277, 1984.

AUDITORY PROCESSING OF SPEECH

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

AUDITORY PROCESSING OF SPEECH

Steven Greenberg[1] and Malcolm Slaney[2]

[1]International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA

[2]IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA

For human listeners, speech is the most apparent means by which the auditory system impinges on everyday life. Virtually all of our daily interactions with friends, family, and colleagues rely on the auditory modality. At first glance the processing performed in the auditory pathway seems transparent and simple. After all, rarely do we experience difficulty understanding what a speaker says, even in the presence of background noise and reverberation. This apparent perceptual invariance of speech has led some to conclude that the brain decodes the signal by back-computing the articulatory gestures from the acoustics [2]. However, this "motor theory" does not actually solve the speech decoding problem as speakers have many different ways with which to articulate the same words and elementary components ("phones") [1]. The variability at the articulatory level is almost as daunting as that observed in the acoustics. Some other theoretical framework is required to account for the robustness of speech under the wide range of acoustic conditions in which humans communicate.

Two of the papers in this section address the invariance issue from the auditory perspective.

The articulatory feature, *voicing* (which reflects the vibration of the laryngeal vocal folds), is an important property for distinguishing among certain consonants. Automatic speech recognition (ASR) systems generally ignore the signal's periodic properties (associated with the fundamental frequency and its perceptual correlate, pitch) and try to deduce the presence of voicing via analysis of the spectral contour. Although this approach may be adequate for recognition under high signal-to-noise (S/N) conditions, it may not work as well in noisy environments. The chapter by Strope and Alwan address this general problem by application of a variety of different computational techniques, each designed to ascertain the presence or absence of voicing using some form of cue extracted from the fine-temporal structure of the waveform. Their approach can be useful under noisy conditions because the ability to follow quasi-periodic properties of the signal may not degrade very much in such circumstances, making the method more robust than the conventional spectral-envelope approach used in most ASR systems. Rather than relying on a single metric, they use three separate methods in order to estimate the probability of voicing in the context of the contrast between the voiced, alveolar fricative [z] and its unvoiced counterpart, [s]. Although the methods have all been used before to measure acoustic periodicity, they have not previously been applied in tandem for a single recognition task. The basic idea is to detect amplitude modulation over a certain frequency range associated with voicing (ca. 90–160 Hz for male adult speakers, and between 180–330 Hz for an adult female voice). Two of the methods compute the fluctuation in the waveform envelope over the relevant fundamental frequency range (but in different ways). The third method computes the autocorrelation function of the waveform to assess the likely periodicity of the signal. Strope and Alwan find that each method provides a reasonable estimate of voicing, but suggest that using the algorithms in combination yields even more reliable recognition. Their result provides potential insight as to why the auditory system is, itself, likely to use multiple methods for computing specific properties of the speech signal.

A different approach to robust recognition is taken by Tian and colleagues. They are concerned with developing robust "front-end" features for various sorts of noisy environments (in particular, cellular telephone transmission) based on spectral parameters of the signal. They compare a widely used form of spectral estimation, mel-cepstral features (which provide a highly smoothed representation of the spectral envelope distributed over quasi-logarithmic frequency coordinates), with a spectral estimator derived from a model of auditory-nerve excitation patterns. The assumption is that an auditory model is more robust in noisy environments since there are certain properties of neuronal firing that are relatively invariant across a wide range of signal-to-noise ratios. While recognition performance degrades as a function of decreasing S/N for both types of spectral features, those based on the auditory model degrade far less than the conventional mel-cepstral approach. Such results suggest that neuronal firing properties of the auditory nerve may impart a measure of robustness and invariance to signal representations under a wide variety of acoustic background conditions.

The third paper in this section examines not recognition, but synthesis of speech. The requirements for creating realistic, intelligible voices is quite different from recognition. The goal of the latter is to ascertain the identity of the words spoken (and by implication the sequence of the words' constituent phones). A coarse representation of the spectrum usually yields results superior to those based on finer-grained spectral estimations. In synthesis the opposite relation holds — the finer-grained spectral representations sound far more natural than those based on coarse articulatory models. Kawahara, in his chapter, describes an ingenious method for extracting key spectro-temporal properties of the speech signal and using these as the basis for modifying the waveform to vary the phonetic and prosodic properties of the speaker's voice. Thus, he can model the voice quality of an individual speaker with high fidelity and then build word and phrase models using the same speaker's voice. In Kawahara's approach it is essential that the periodicity associated with the signal's fundamental frequency (i.e., voice pitch) be estimated with precision (in contrast to recognition which dispenses entirely with periodicity information). It is also important that the phase characteristics of the waveform modulation associated with each frequency channel be aligned in proper fashion. Such results suggest that the fine spectral granularity of the auditory system may be exceedingly important for sound quality.

Together, the three chapters of this section provide a representative sample of computational approaches derived from auditory models that are currently being used to enhance speech technology. And to the extent they are successful in this endeavor, the studies also shed light on the function of specific properties of the auditory system.

References

[1] Greenberg, S. "Speaking in shorthand — A syllable-centric perspective for understanding pronunciation variation." *Speech Communication*, 29: 159–176, 1999.

[2] Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. "Perception of the speech code." *Psych. Rev.*, 74: 431–461.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

MODELING THE PERCEPTION OF PITCH-RATE AMPLITUDE MODULATION IN NOISE

Brian P. Strope[1,2] and Abeer A. Alwan[1]

[1]*Department of Electrical Engineering*
UCLA, 405 Hilgard Ave., Los Angeles, CA 90095, USA
[2]*Nuance Communications*
1380 Willow Road, Menlo Park, CA 94025, USA

1. Introduction

The robustness of speech communication depends on a structural hierarchy that is deeply embedded with redundancy. This structure exists in our language — sentences, phrases, words, syllables, phonemes — as well as in the rapidly varying acoustic details that cue these perceptions. Together, the stages of this hierarchy form a web of partially orthogonal dimensions. In typically noisy situations the listener is unlikely to perceive each of these representational units in explicit detail. Instead, partially corrupted cues are readily filled-in with expectations derived from other stages in the hierarchy: the listener may miss the phonetic segment but still reconstruct the word (or miss the word, but understand the gist of the phrase, etc.).

This chapter demonstrates that a similar redundancy exists for the detection of voicing in noise. Specifically, after power spectrum cues are removed, listeners can use amplitude modulation cues to detect voicing at low signal-to-noise ratios.

Because of this redundancy, amplitude modulation cues in voiced speech provide a salient, robust sensation of pitch that may be instrumental in recognizing speech in noise. In the current study, three psychoacoustic models are used to predict the temporal modulation transfer function (TMTF) and the detection of voicing for high-pass filtered naturally spoken fricatives in noise. Computational models based on waveform-envelope statistics and modulation filtering properties predict the TMTF data with a high degree of precision, and models derived from a summary autocorrelogram representation fit both the TMTF and high-pass filtered data sets.

1.1 Voicing in Speech Analysis and Automatic Recognition

During voiced speech, the vibration of the vocal folds excites time-varying resonances of the vocal tract. Given a sequence of feature vectors representing log-magnitude, spectral estimates of the vocal-tract transfer function across time, most automatic speech recognition (ASR) systems use a hierarchy of non-stationary stochastic models operating at progressively longer intervals of speech analysis (10–30 ms) and statistical modeling (at the representational level of the phonetic segment, word, phrase and sentence) to ascertain what was most likely to have been said [17]. However, ASR systems rarely use pitch or voicing information in this process.

Instead, the signal processing for feature vector extraction usually reflects some form of deconvolution, attempting to shield vocal-tract transfer-function estimates from the impact of the driving function. Linear prediction, for example, is used with a predictor polynomial that is significantly shorter than the anticipated glottal periodicity. Similarly, when homomorphic analysis is used for ASR, the high-quefrency cepstral terms (which can represent the periodic ripple across the spectral estimate resulting from a harmonic driving function), are ignored. Using Mel-frequency cepstral coefficients (MFCC), the initial spectral estimate is first averaged (in time) over multiple pitch periods and then integrated across frequency, providing an approximation of auditory frequency selectivity. The output is then logarithmically compressed and the discrete cosine transform is used to partially decorrelate the log-magnitude spectral estimate across frequency. Higher-order terms in the resulting cepstral vector are ignored. Integrating across time and frequency reduces the variance of the spectral estimate, and together with the truncated cepstral vector, nearly eliminates periodic source information.

Deconvolution is an important step for isolating the phonetic information about "*what was said*," from aspects of the prosodic information pertaining to "*how it was said*." But as the first processing stage it may be eliminating large parts of the perceptually salient information used by humans to identify and recognize speech in noisy environments.

Speech communication has evolved to be robust in noise. Redundancies are, therefore, ubiquitous. Perceiving speech under noisy conditions requires an intelligent use of the potentially unreliable, albeit redundant, multi-dimensional cues spread over wide-ranging time scales. While deconvolution must occur somewhere in the recognition process, blindly eliminating a potential wealth of redundant cues may not be appropriate for the first stage of processing. Thus, rigid blind deconvolution in this first stage is unlikely to be optimal.

1.2 Pitch Perception

Processing voicing information in speech requires analyzing the harmonic structure associated with a quasi-periodic vocal driving function and might therefore be considered as an aspect of pitch perception.

In 1951, Licklider proposed a "duplex" theory [11] to account for many properties of pitch perception, including the perception of the missing fundamental (or residue pitch), as well as the pitch of modulated noise. Licklider envisioned neural machinery that measured the running temporal autocorrelation in each auditory frequency channel. The sensation of pitch, he proposed, is associated with the common periodicities observed across channels.

In 1984 Lyon was able to simulate an implementation of the duplex theory, labeling the graphic output a *correlogram* [12]. Since then, Meddis and colleagues [13] [14] have formalized the simulations and included a final stage that adds the running autocorrelations across each channel generating a *summary correlogram*. Cariani and Delgutte have also shown that similar processing of measured auditory-nerve impulses is sufficient to predict many classic pitch perception phenomena [2]. Other researchers have replaced the autocorrelation function with different mechanisms that measure the temporal intervals in each channel (e.g. [15] [4] [5]).

In general (and as shown in Licklider's original sketches achieved without the aid of computational simulation), simulations using these models provide a graphical output that correlates well with pitch. The time lag of the peak in the summary correlogram is usually found to be the reciprocal of the frequency of the perceived pitch and the height of the peak is often correlated with pitch salience. With few exceptions however, the models are not used to predict psychoacoustic just-noticeable-differences (jnds) with general stimuli. Together with the lack of a clearly identified physiological substrate for the implementation of the required timing measurements, this line of research remains somewhat of an "open-loop."

1.3 Perception of Amplitude Modulation

Processing voicing information in speech might also be thought of as a form of amplitude modulation perception.

In 1979, Viemeister applied a linear systems approach to the detection of acoustic envelope fluctuations [21]. His model was first fit to data describing the detection of sinusoidal amplitude modulation of wideband noise and then used to predict the detection of other harmonic envelopes. Motivated by the close relationship between standard deviation and autocorrelation, Viemeister's model used the standard deviation of a demodulated envelope as the statistic to predict human performance. Although this measure does not characterize the perceived pitch of the amplitude modulation, a more sophisticated simulation involving autocorrelation was not required to accurately fit the detection data. More recently, this model has been extended to predict other amplitude modulation detection data [19] [20].

In 1989, Houtgast measured modulation masking that suggested explicit neural modulation filtering [8]. Narrow-bandwidth noise modulators were found to mask the perception of sinusoidal modulators in a manner similar to the spectral masking of tones by narrow-band noises. Modulation tuning has also been measured physiologically [e.g., 9]. However, other modulation masking experiments, using sinusoids, have been less conclusive [19] [1]. Nonetheless, a model of modulation filtering has been implemented and shown to be correlated with many aspects of amplitude-modulation perception [3].

In essence, modulation filtering replaces the single low-pass filter in the envelope statistic model with a second bank of filters. The modulation filtering simulations also include a better approximation of auditory filtering than the single band-pass filter used in the envelope statistic model.

Therefore, there are at least three modeling approaches which may be helpful for analyzing the periodic envelope fluctuations in voiced speech: autocorrelation or interval-based temporal processing, the measurement of an envelope statistic and explicit modulation filtering. To choose among them, implementations of each were first fit to predict TMTF data and then each was used to predict the discrimination of voicing for strident fricatives in noise.

2. A Strident Fricative Case Study

Fricatives are generated by forcing air through a sufficiently narrow constriction in the vocal tract, resulting in a turbulent, noise-like source. With voiced fricatives the vocal folds also vibrate, adding low-frequency energy to the spectrum. The relative level of the first harmonic, compared with that of the adjacent vowel, has been shown to serve as an effective indicator of voicing distinctions for fricatives [18] [16].

2.1 Characterizing [s] and [z]

For our study, the strident fricatives [s] and [z], along with the vowels [a], [i] and [u] were recorded as CV syllables from four talkers. Figure 1 compares average log-magnitude spectral estimates for [s] and [z]. The voiced [z] has low-frequency energy not present in the [s].

Current ASR systems use the presence of low-frequency spectral energy to discriminate these sounds. However, there are situations where this particular spectral cue can be obscured (e.g. a high-pass channel or with a competing low-pass noise).

Figure 2 shows examples of the temporal waveform for [s] and [z], after each has been high-pass filtered above 3 kHz. Without low-frequency spectral components, the low-fre-

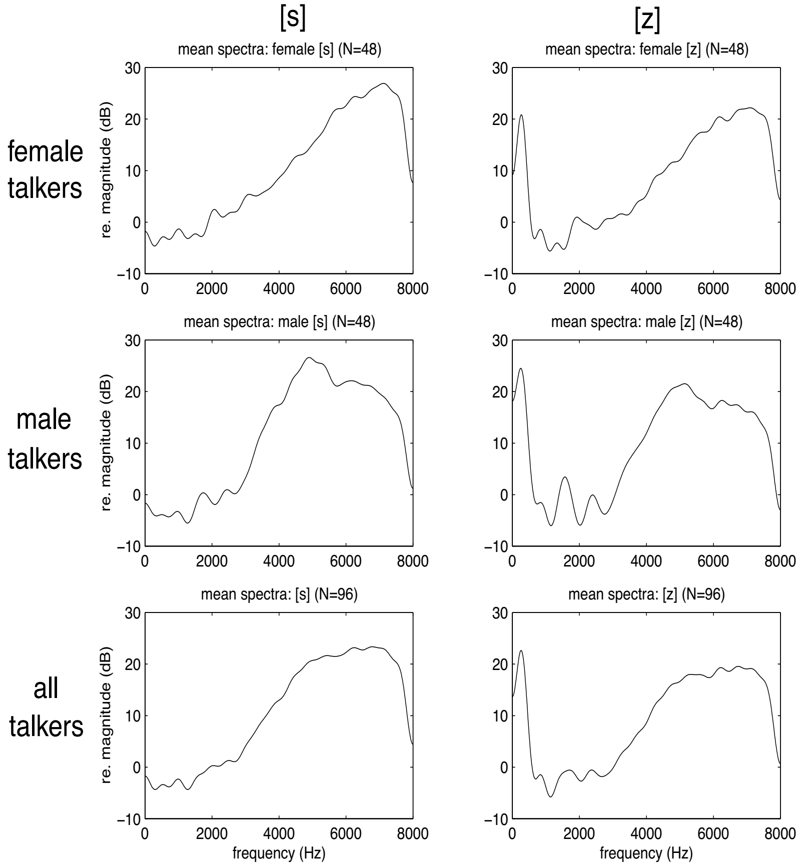


Figure 1 A comparison of average spectral estimates for [s] and [z] spoken by both male and female speakers.

quency pitch-rate information is represented in the envelope of the high-frequency, noise-like carrier. These figures provide evidence that the vibrating vocal folds can modulate the pressure source that drives the turbulence for a voiced fricative. The modulated noise source leads to a potentially redundant voicing cue in a spectral region with significant speech energy. ASR systems that integrate spectral estimates over multiple glottal periods do not distinguish such sounds, while human listeners can distinguish them even at low signal-to-noise ratios (see Section 4).

2.2 Perceptual Measurements

To measure the perceptual sensitivity to this potential voicing cue, the discrimination of these sounds was measured in wide-band noise. The syllable-initial fricatives were temporally isolated from the adjacent vowel, and high-pass filtered above 3 kHz. During the perceptual tests, tokens were centered within a one-second span of spectrally flat noise.

Adaptive tests [16] were used to track the perceptual discrimination of the isolated fricative as a function of SNR at two *d'* levels. For each trial, the subject was required to identify a randomly chosen token as either [s] or [z]. Feedback was provided. The initial SNR was sufficiently high that the fricatives were clearly distinguishable for all subjects. The SNR

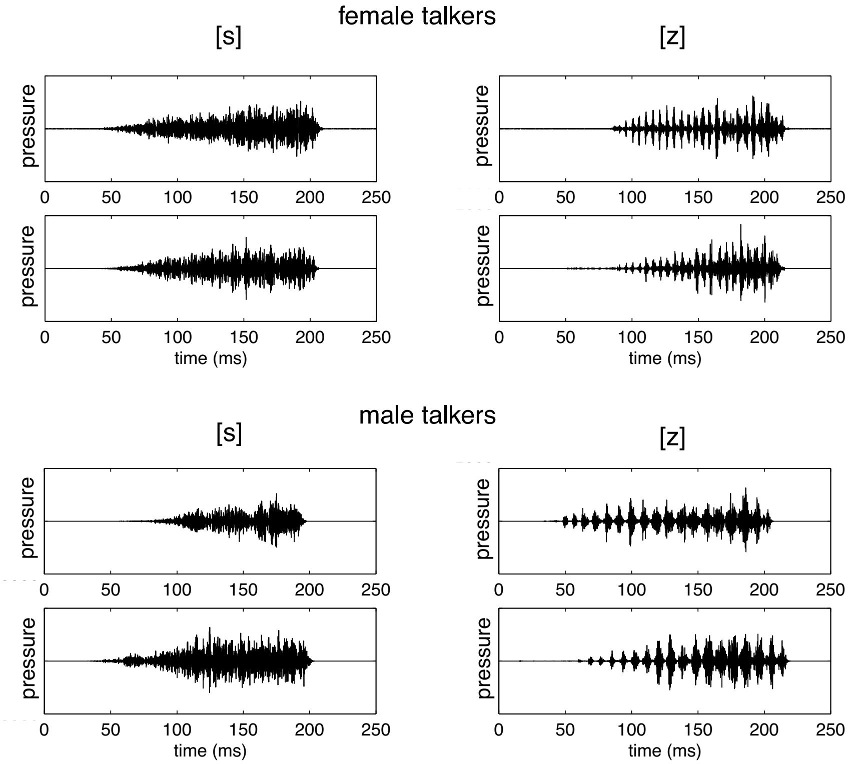


Figure 2 Examples of temporal waveforms after high-pass filtering.

was increased after an incorrect response and decreased after either two or three consecutively correct responses. (A reversal is defined as a change in the direction of the SNR step). The SNR step size started at 4 dB, and was reduced to 2 dB after the first reversal, and to 1 dB after the third. The average of the SNR at the next 6 reversals provided an initial threshold estimate. If the variance in this estimate was less than 2 dB, the measurements stopped, otherwise the experiment continued for up to 6 more reversals. The average of three such measurements provided a final threshold estimate for each subject. When 2 (or 3) correct responses are required, the threshold estimate converges to a 70.7% (or 79.4%) correct response rate. For this experiment, these correspond to *d'* values of 1.09, and 1.64, respectively. Four audiometrically normal subjects participated in the experiment. Average thresholds across these four subjects are shown together with model predictions in Figure 10 below.

3. AM-Detection Mechanisms

The task in this experiment requires detecting periodic envelope fluctuations, which become increasingly weak with the addition of noise. Perhaps the most direct approach is to model this perceptual process using an envelope statistic.

3.1 Envelope Statistic

Figure 3 shows a block diagram of the signal processing in an envelope-statistic model. This classical approach reduces auditory processing to the following steps: auditory filtering

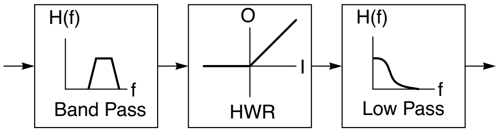


Figure 3 Schematic illustration of the envelope detection process used in the current study.

(measured along the basilar membrane), half-wave rectification (approximated in inner-hair-cell transduction) and low-pass filtering (computed throughout the higher levels of auditory processing). From an engineering perspective, the band-pass filter selects a channel, while the half-wave rectifier serves as a non-linearity that modulates the carrier down to DC, with the low-pass filter tracking the envelope.

The model's sensitivity to amplitude-modulated wideband noise increases with a broadening of the bandwidth in the initial filter, while the reduction of sensitivity with increasing envelope frequency is mostly determined by the final low-pass filter.

3.2 Modulation Filtering

A schematic overview of an implementation of modulation filtering is shown in Figure 4. Building from the envelope-detection processing above, the model includes multiple 4th-order gammatone filters [15] which provide a reasonable approximation of auditory filtering, and replaces the single low-pass filter with a second filterbank that analyzes the envelope spectrum.

The frequency response for the modulation filters used (Q_{3dB} of 2, and -12 dB DC gain) was adapted from [3]. For each filter the implementation used a second-order pole and a first-order (real) zero at DC. The distance of the zero to the unit circle was set to meet the DC specification. The resulting frequency responses are shown in Figure 5.

Both the modulation filtering and the envelope-detection model compute the magnitude of the fluctuations of the envelope of the acoustic waveform. As stated previously, the primary difference is that modulation filtering assumes a second, filtering stage tuned to different envelope modulation rates. Figure 6 compares the processing output of these two models to a noise carrier with no modulation, as well as one with 56% modulation [20 log(m) = 5 dB depth]. Although the standard deviation of the input is the same for the modulated and unmodulated cases, the outputs of both models exhibit relatively more fluctuation in the modulated case.

3.3 Correlational Analysis

An overview of the correlational analysis is shown in Figure 7. This is an implementation of Licklider's model [11] together with a final stage that adds correlation estimates across channels [13] [14]. The first stage is the same gammatone approximation of cochlear filtering, used above. The transduction stage includes half-wave rectification, low-pass filtering, and a 2nd-order Butterworth high-pass filter with a cut-off of 4 Hz. Running autocorrelations are computed in each filter channel and the results are summed across channels.

Our implementation of running autocorrelation for each channel involves two stages. First, the instantaneous product of the current input, and a version of the input delayed by the interval, τ, is computed for all time and all values of τ:

$$x_1(t,\tau) = x(t)\, x(t-\tau).$$

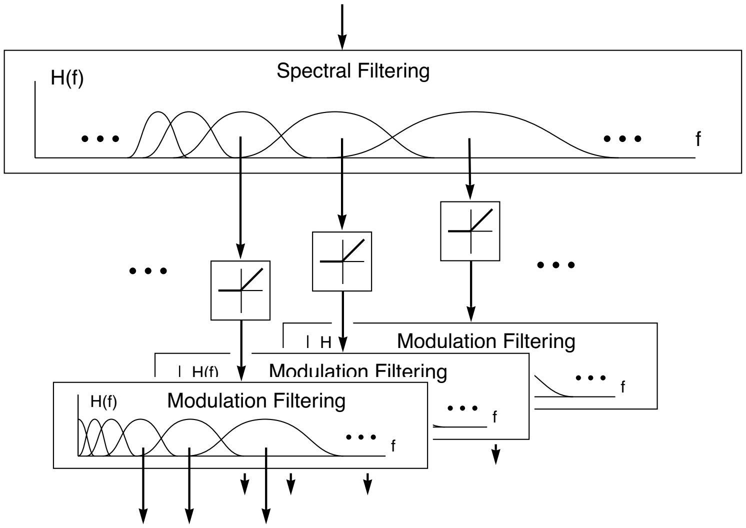


Figure 4 A schematic illustration of the modulation filtering performed in the current study.

Second, to form a running autocorrelation estimate, these sequences are low-pass filtered (for each value of τ) to below one half of the final correlation sampling rate:

$$x_2(t,\tau) = x_1(t,\tau) * h_{lpf}(t).$$

In the evaluations below, the correlation sampling rate was 25 Hz, and $h_{lpf}(t)$ was implemented as a 6th-order Butterworth filter with a -3 dB point at 10 Hz. That is, after the low-pass filter, the running autocorrelations were sampled every 40 ms and then summed across frequency channels to generate a sequence of summary correlogram estimates.

As described above, the position of the peak in the summary correlogram has often been shown to be correlated with the reciprocal of the perceived pitch (in units of frequency), although some models utilize the entire waveform of the summary correlogram [13] [14]. Our analysis represents a compromise between these two approaches. For each sample of the summary correlogram, our statistic is the maximum difference, across all delay values τ, between the summary correlogram values at delays of τ and τ/2:

$$statistic = max\ [sc(\tau)\ \text{-}\ sc(\tau/2)],\ (0 < \tau < 20\ ms).$$

With a sinusoidal envelope, this difference peaks at a value of τ, equal to the period of the sinusoid. Figure 8 includes examples of this decision statistic using the same noise carrier, but with either no modulation or with 56% modulation (i. e., 5 dB depth) at 100 Hz. In the

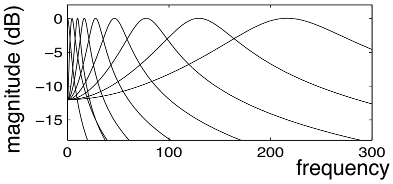


Figure 5 Responses of the modulation filterbank. Each filter is implemented using a complex pole and a real zero

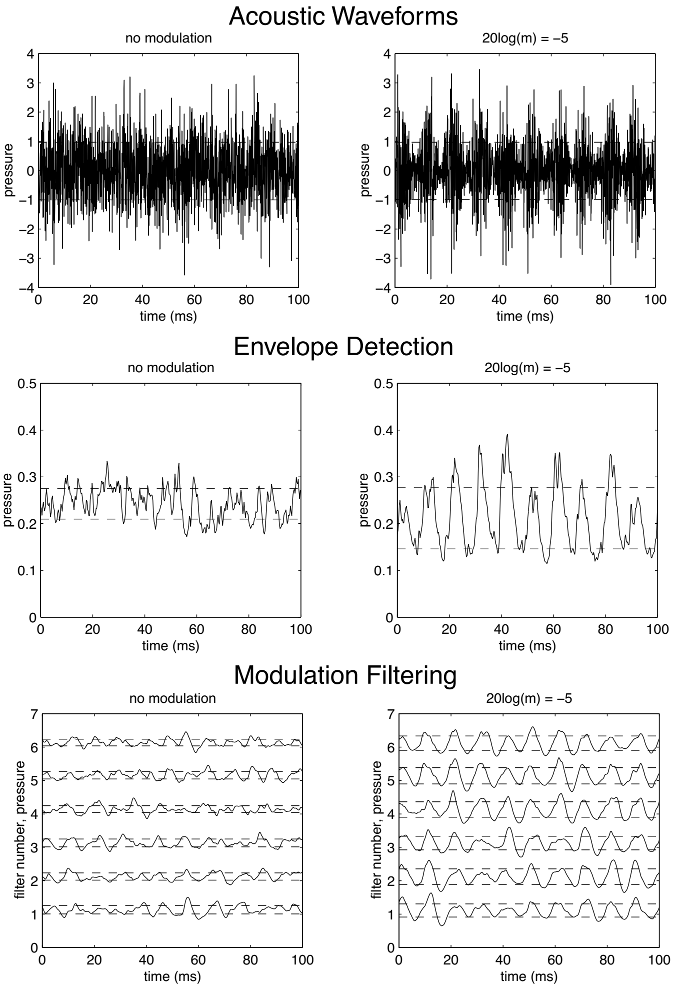


Figure 6 Comparisons of the amplitude modulation detection models. Dashed lines indicate standard deviations. The modulation filtering plots show the outputs of six auditory channels, each filtered by a modulation filter centered at 100 Hz.

modulated case, the first peak (after zero delay) in the summary correlogram occurs at the period of the modulation, 10 ms. When there is no modulation, the summary correlogram approximates an impulse. Adding the individual correlation estimated across channels reduces some of the variance; consistent modulation patterns across channels add together, while inconsistent ones generally cancel each other. However, considerable variation remains across summary correlogram samples (shown in the lower half of Figure 8) due to the stochastic nature of the carrier.

4. Comparing Predictions

The temporal modulation transfer function (TMTF) is a measure of auditory sensitivity to amplitude modulation as a function of modulation frequency. More specifically, the mini-

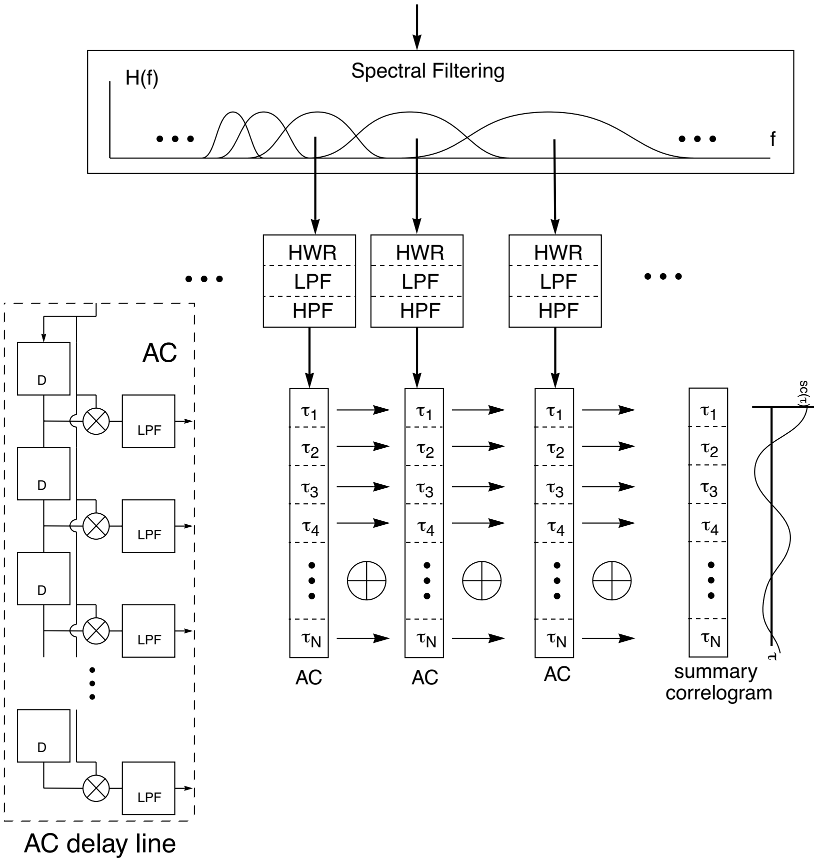


Figure 7 Overview of the correlational processing. Inset shows autocorrelation delay-line detail.

mum detectable sinusoidal amplitude modulation depth is typically measured as a function of modulation frequency using wide-band noise carriers.

Each of the three models was initially adjusted to predict TMTF measurements derived from previous studies [20] [3]. The resulting models were then used to predict the discrimination thresholds for the high-pass filtered [s] and [z] tokens in noise. Because the natural fricatives are non-stationary all three models were evaluated using multiple measurements in time (or multiple "looks") [22].

For the envelope-statistic model, the best match was found using an initial filter bandwidth of 3 kHz, centered at 5.5 kHz. With these parameters, the filter approximated a matched-filter for the high-pass filtered [s] and [z] segments. The low-pass filter was a 1st-order Butterworth with a cut-off of 90 Hz. The normalized fourth-moment statistic [19] [20] was used.

To obtain multiple measurements in time, the output of the envelope detection mechanism was segmented using partially overlapping, 50-ms windows that had 10-ms raised-cosine onset and offsets, as well as a 30-ms steady-state center. The windows were incremented by 40 ms. The window length was chosen to ensure multiple periods in each window for the pitch-frequency range of interest. By modulating the DC offset in the envelope, the

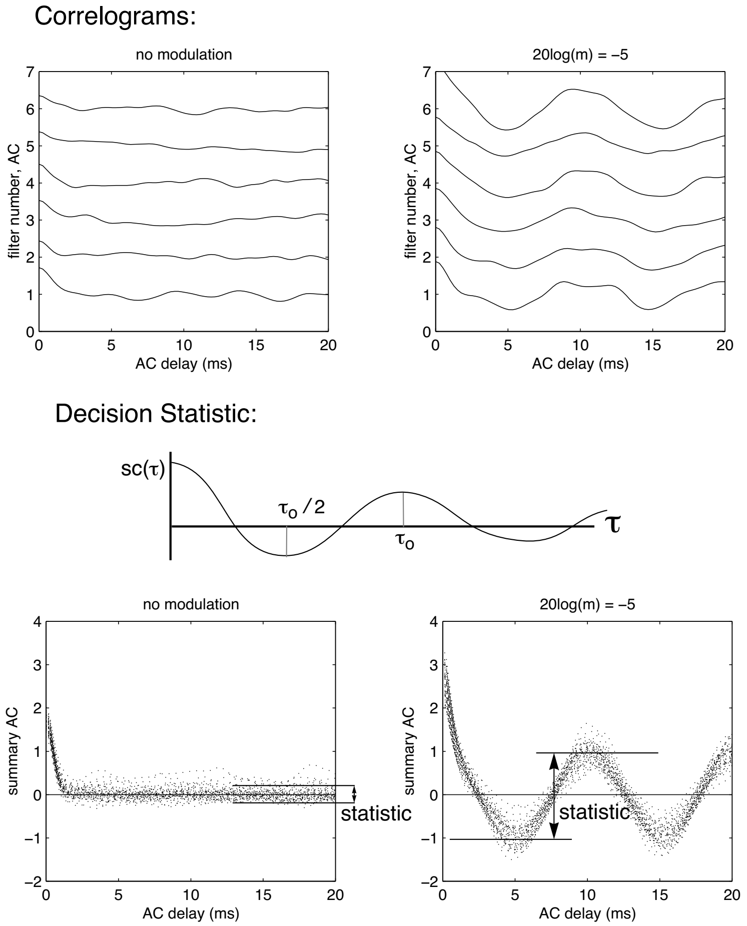


Figure 8 Samples of the correlogram output and super-imposed examples of the summary correlogram decision statistic. Input signals are the same as in Figure 6.

shape of the window can dominate measurements using the standard deviation or the fourth-moment. Therefore, the DC offset for each 50-ms window was removed before weighting by the raised-cosine and then added back before computing the statistic.

Threshold predictions were obtained by using the difference in the decision statistic in signal and non-signal intervals over 100 simulations in order to estimate *d'* for each "look." Assuming independence of individual measurements, a total detection *d'* was estimated as the length of a *d'* vector containing all looks [7]. With a stimulus duration of 500 ms used for the TMTF data, the vector included 12 elements (or looks). A line was fit to the log of total *d'* estimates as a function of the log of the modulation depth. From this line, the modulation threshold was estimated from the point where the line crossed the *d'* threshold of 1.26 tracked in the perceptual TMTF measurements [20] [3].

With the modulation filtering and correlation models, the initial filtering stage was six, 4th-order gammatone filters with center frequencies range between 4.28 Hz and 6.97 kHz. Filters overlapped at their half-power points, and the bandwidths were set using the equation described in [6]. To predict the TMTF data using modulation filtering only the modulation

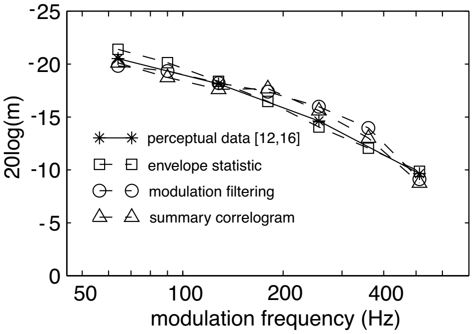


Figure 9 Three predictions of TMTF data: *m* is the modulation depth; perceptual data is an average from [20] and [3].

filter tuned to the probe envelope frequency was considered. When predicting the fricative data two modulation filters centered at 120 Hz and 200 Hz were used. The windowing applied to the envelope-detection simulations was also used for the modulation filtering. The standard deviation was the measured statistic.

As observed previously [3], the modulation filtering was too sensitive to predict human performance without adding a large amount of internal noise. To obtain the best match to the TMTF data, internal noise was added both before and after modulation filtering.

Using the correlation model, the peak distance statistic described above was measured every 40 ms for the summary correlogram. To approximate the shape of the TMTF data, the first-order, low-pass filter was used with a cut-off frequency at 280 Hz.

TMTF threshold predictions for all three models are shown in Figure 9. Each model provides a reasonable prediction across this frequency range. Predicting the voicing detection thresholds for the natural, non-stationary, fricatives in noise required finding the fricatives (or more specifically finding the voicing in the fricative) within the 1-second interval of noise. For all model predictions below, only the three consecutive temporal segments that maximized the difference from the background noise were analyzed, providing three temporal looks per token. Total *d'* values were then estimated as a function of SNR.

Figure 10 shows the *d'* estimates for each model's prediction of the discrimination of the high-pass filtered [s] and [z] tokens in noise. The model based on correlations provided the best prediction.

5. Modeling Implications

The envelope statistic was not sufficient, by itself, to discriminate reliably between [s] and [z] (even at relatively high SNR values) because this measurement does not distinguish the periodic voicing cues in [z] from the aperiodic fluctuations in [s]. Both the modulation filtering and the autocorrelation processing include specific modulation tuning and as a result more accurately fit the observed data.

Reasons for the difference in performance between these two models are less clear, and could be specific to these simulations. By reducing the amount of internal noise, the modulation filtering model provides a better estimate of the [s] and [z] data, but over-estimates the TMTF sensitivity. One primary difference is that the autocorrelation mechanism integrates correlation estimates across frequency, while the modulation-filtering simulations use the more general assumption that each output corresponds to an independent measurement. Inte-

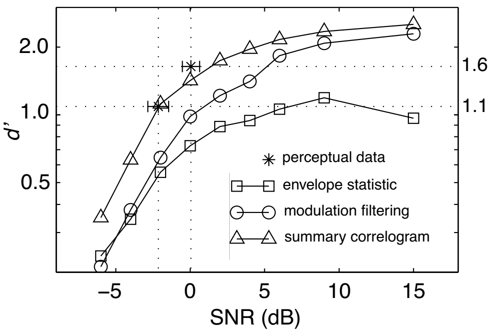


Figure 10 Discriminating high-pass filtered [s] and [z]: data are an average across four subjects.

grating correlation estimates across frequency channels de-emphasizes envelope components uncorrelated across frequency in favor of correlated components. Another difference is that the correlation simulations uses a low-pass filter to limit sensitivity, while the modulation simulation incorporates internal noise.

It is interesting to note that if the auditory system does include a cross-channel interval-based representation, redundancies in this representation are likely to make it inefficient to maintain across many regions of the pathway. Efficient decorrelation of the (potentially smooth and periodic) summary correlogram might approximate a cosine transform. Such periodic transformations exist in other perceptual systems [23]. In this case the decorrelated representation would have many of the properties of the (demodulated) output of a modulation filterbank. The difference is that the envelope analyzed is first processed to identify common correlations across a broad frequency range.

6. Conclusion

This chapter has identified a secondary temporal cue that can reliably distinguish between [s] and [z] on the basis of voicing. This amplitude-modulation cue had not been identified in previous studies of voiced fricatives [18] [16]. Once the cue has been identified it is not clear what processing should be used to reliably extract it. Three possibilities were investigated in this study.

While cross-channel, interval-based processing has been quite successful in predicting many properties of pitch perception, we have shown that these mechanisms can also predict TMTF thresholds and the detection of voicing for high-pass filtered fricatives in noise. Simulations using envelope-statistic and modulation-filtering models fit TMTF data, but do not predict the isolated speech data.

Acknowledgements

We thank the subjects for their efforts, and Jim Hant for helpful discussions and for his assistance in running some of the experiments. This research was supported by NIDCD grant #DC02033 and the NSF.

References

[1] Bacon, S. P. and Grantham, D. W. “Modulation masking: Effects of modulation frequency, depth, and phase.” *J. Acoust. Soc. Am.,* 85: 2575–2580, 1989.

[2] Cariani, P. A. and Delgutte, B. “Neural correlates of the pitch of complex tones: I. Pitch and pitch salience.” *J. Neurophysiol.*, 76: 1698–1716, 1996.

[3] Dau, T., Kollmeier, B. and Kohlrausch, A. “Modeling auditory processing of amplitude modulation. I–II.” *J. Acoust. Soc. Am.,* 102: 2892–2919, 1997.

[4] de Cheveigne, A. “Cancellation model of pitch perception.” *J. Acoust. Soc. Am.*, 103: 1261–1271, 1998.

[5] Ghitza, O. “Auditory nerve representations as a basis for speech processing.” In *Advances in Speech Processing*, S. Furui, M. Sondhi (eds.), New York: Marcel Dekker, pp. 453–485, 1991.

[6] Glasberg, B. R. and Moore, B. C. J. “Derivation of auditory filter shapes from notched-noise data.” *Hearing Res.*, 47: 103–138, 1990.

[7] Green, D. M. and Swets, J. A. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[8] Houtgast, T. “Frequency selectivity in amplitude-modulation detection.” *J. Acoust. Soc. Am.*, 85: 1676–1680, 1989.

[9] Langner, G. “Periodicity coding in the auditory system.” *Hearing Res.* 60: 115–142, 1992.

[10] Levitt, H. “Transformed up-down methods in psychoacoustics.” *J. Acoust. Soc. Am.* 49: 467–477, 1971.

[11] Licklider, J. C. R. “A duplex theory of pitch perception.” *Experientia*, 7: 128–134, 1951.

[12] Lyon, R. F. “Computational models of neural auditory processing.” *Proc. IEEE ICASSP*, 36.1: 1–4, 1984.

[13] Meddis, R. and Hewitt, M. J. “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification.” *J. Acoust. Soc. Am.,* 89: 2866–2882, 1991.

[14] Meddis, R. and O’Mard, L. “A unitary model of pitch perception.” *J. Acoust. Soc. Am.*, 102: 1811–1820, 1997.

[15] Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand, M.“Complex Sounds and Auditory Images.” In *Auditory Physiology and Perception*, Y. Cazals and K. Horner (eds.), Oxford: Pergamon Press, pp. 429–446, 1992.

[16] Pirello, K., Blumstein, S. and Kurowski, K. “The characteristics of voicing in syllable-initial fricatives in American English.” *J. Acoust. Soc. Am.,* 101: 3754–3765, 1997.

[17] Rabiner, L., Juang, B. H. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[18] Stevens, K. N., Blumstein, S., Glicksman, L.,Burton, M., Kurowski, K.“Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters.” *J. Acoust. Soc. Am.,* 91: 2979–3000, 1992.

[19] Strickland, E. and Viemeister, N. “Cues for discrimination of envelopes.” *J. Acoust. Soc. Am.,* 99: 3638–3646, 1996.

[20] Strickland, E. and Viemeister, N.“The effects of frequency region and bandwidth on the temporal modulation transfer function.” *J. Acoust. Soc. Am.,* 102: 1799–1810, 1997.

[21] Viemeister, N. “Temporal modulation transfer function based on modulation thresholds.” *J. Acoust. Soc. Am.,* 66: 1364–1380, 1979.

[22] Viemeister, N., Wakefield, G. “Temporal integration and multiple looks.” *J. Acoust. Soc. Am.* 90: 858–865, 1991.

[23] Wang, K. and Shamma, S. “Self-normalization and noise-robustness in early auditory representations.” *IEEE Trans. Speech, Aud. Proc.*, 2.3: 412–435, 1994.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

FRONT-END DESIGN BY USING AUDITORY MODELING IN SPEECH RECOGNITION

Jilei Tian, Kari Laurila, Ramalingam Hariharan and Imre Kiss

Speech and Audio Systems Laboratory
Nokia Research Center
P. O. Box 100, 33721 Tampere, Finland

1. Introduction

It is well known that the human auditory system is an “expert” speech recognizer. If a computer-based speech recognition system could be designed that sufficiently reflects the processing of the auditory system, the resulting representations should be superior to representations based on non-biological criteria commonly used in computer speech recognition algorithms. The potential advantages of using auditory modeling for automatic speech recognition (ASR) depend on how accurate the models are in simulating the relevant properties of the human auditory system. Developing such models relies on the knowledge we currently possess about the auditory system. This knowledge is acquired by combining data that have been collected using psychophysical, physiological and auditory phenomena. Due to extensive studies of the auditory system we now know quite a lot about the kinds of transformations that occur, at least at the peripheral level of the pathway, and it has become feasible to build computational models that take these auditory properties into account. Different types of auditory representations for speech may make it easier to identify those features of the signal that are most relevant for automatic speech recognition. In addition to the commonly used Mel Frequency Cepstral Coefficients (MFCC) front-end [9], a number of alternative auditory approaches have recently been proposed.

The perceptual linear prediction (PLP) technique proposed in [3] uses some concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum, consistent with many properties of human hearing. Even though PLP is computationally efficient and yields a low-dimensional representation of speech, it considers limited aspects of auditory processing. A joint synchrony/mean-rate auditory speech processing scheme proposed in [11] provided promising results in a case study but its utilization with a commonly used Hidden-Markov-model (HMM) based classifier has not been very successful [4]. Cohen proposed another scheme in [1] but did not apply the method within the HMM framework. Ghitza [2] developed an ensemble interval histogram (EIH) model. In comparison with the commonly used MFCC front-end, on an isolated word database in adverse conditions [4], the reduction of error rate with the EIH model was rather small with a high computational load. In addition, auditory modeling research has been widely carried out for purposes other than automatic speech recognition [5].

In short, many researchers have shown that an auditory modeling approach can lead to enhanced representations of the speech signal. In this chapter we combine certain previously proposed auditory functions and apply them to an ASR front-end in order to obtain improved noise robustness.

2. Auditory Modeling

2.1 The Human Auditory System

Sound travels through the external auditory meatus (the ear canal) to produce a pattern of vibration at the tympanic membrane (ear drum). A series of three small bones (i.e., the ossicles) — the malleus (attached to the tympanic membrane), the incus and the stapes — transmit the pressure variation to the oval window of the cochlea (Figure 1). The system acts as an amplifier because the area of the tympanic membrane is greater than the area of the oval window, increasing the total sound pressure. The cochlea contains three compartments:

(1) the scala tympani, which follows the outer contours of the spiral,
(2) the scala vestibuli, which follows the inner contours and connects with the scala tympani at the helicotrema, and
(3) the scala media, which is not connected with the other two and ends blindly at the apex.

The scala tympani and scala vestibuli are filled with perilymph. The scala media contains the organ of Corti and is filled with endolymph. The stapes acts on the oval window of the scala vestibuli. The pressure is then transmitted to the scala tympani and the round window. This pressure pattern is translated into oscillatory movements of the basilar membrane on which sits the organ of Corti. The sensory receptors are hair cells located in the organ of Corti. There are three rows of outer, and one row of inner hair cells. The apical surface of the outer hair cell stereocilia are embedded in the underside of the tectorial membrane, which is relatively stiff. The stereocilia of the inner hair cells are not directly attached to the tectorial membrane, but their tips lie close to this structure. The oscillatory movements of the basilar membrane result in minute deflections of the stereocilia. Sound creates a traveling-wave pattern along the basilar membrane. Different frequencies result in peak amplitudes of the travelling wave at different locations along the basilar membrane, resulting in discrete stimulation of different populations of hair cells. The vibrations of hair cells are transformed into action potentials in auditory-nerve fibers. The fibers of the auditory nerve are most sensitive to a limited range of frequencies and such selectivity is commonly illustrated by means of a frequency

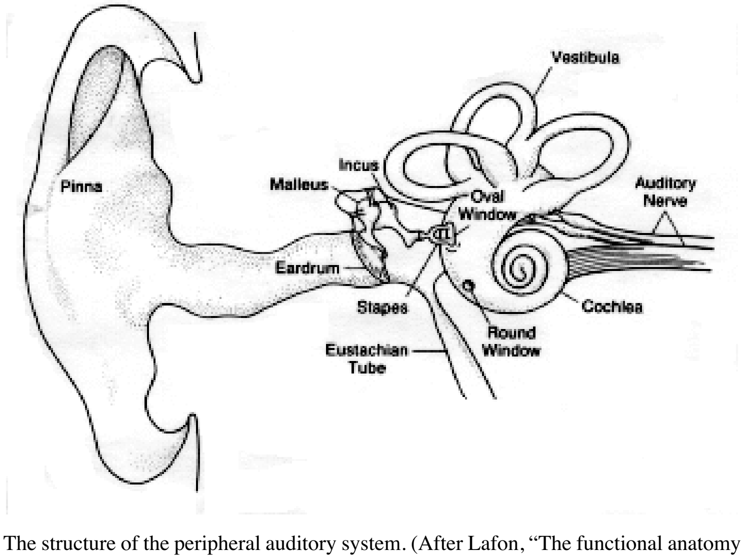


Figure 1 The structure of the peripheral auditory system. (After Lafon, "The functional anatomy of the speecl organs," *Manual of Phonetics,* B. Malmberg)

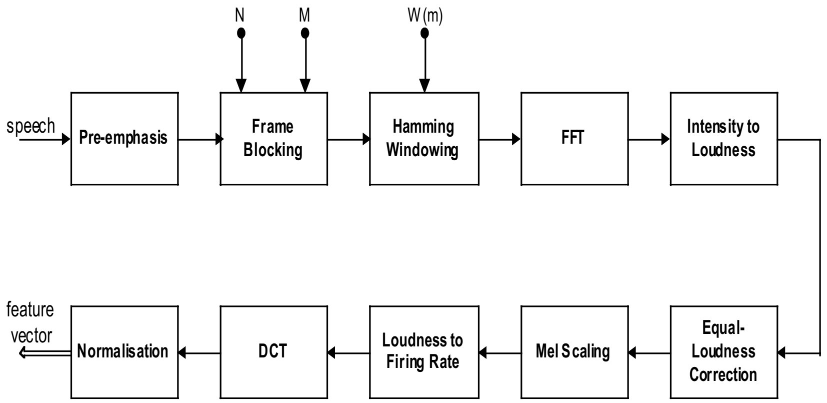


Figure 2 Block diagram of the auditory front-end.

threshold (tuning) curve. The central auditory pathway consists of the cochlear nucleus, superior olivary complex, the lateral lemniscus, the inferior colliculus, the medial geniculate body and the auditory cortex, all specialized to preserve time and frequency information (see [8] for more details).

2.2 *Auditory Front-End*

The auditory system can be divided into the auditory periphery and the central auditory pathway. Since the central auditory pathway is not thoroughly understood, most of the auditory modeling approaches focus on the auditory periphery. The speech processing of the auditory periphery consists of the following basic stages:

(1) low- and high-frequency attenuation in the outer and middle ear,
(2) basilar membrane filtering, and
(3) mechanical to neural transduction.

The outer ear modifies sound by transferring the acoustic vibrations to the eardrum. It consists of a partially cartilaginous flange (the pinna) which includes a resonant cavity at the entrance to the ear canal. The resonances of the outer ear increase the sound pressure at the eardrum, particularly in the range of frequencies of 2–7 kHz (i.e., it functions as a band-pass filter). The transformer action of the middle ear helps to match the impedance of the air in the ear canal to the much higher impedance of the cochlea fluid. In doing so, the middle ear also performs a band-pass function. The outer and middle ear combine to give an approximately flat-topped, band-pass function. The basilar membrane behaves like a filter bank which decomposes sound waves into separate frequency bands. The mechanical motion of the basilar membrane is converted into neural spikes in the post-synaptic auditory nerve. This conversion is performed by the inner hair cells, and is explained in detail in the remaining part of this section.

Despite the extensive research in auditory modeling, the human auditory system is not yet fully understood. Hence, in this chapter we only take into account certain critical auditory functions relevant to speech recognition in order to build an auditory front-end. The basic idea is to incorporate nonlinear frequency scaling, amplitude compression (loudness), short-term adaptation and the firing rate of auditory neurons into the model.

A block diagram of the auditory front-end is given in Figure 2. The digitized speech signal (sampled at 8 kHz) is fed into the first-order, high-pass, pre-emphasis filter as seen in Equation (1). This step compensates for the negative spectral slope of approximately 20 dB per decade in the human articulatory system. This pre-emphasis filter models the functions of the outer and middle ear up to 4 kHz.

$$H(z) = 1 - a_0 z^{-1} \tag{1}$$

The power spectrum of each frame is computed by applying an FFT on the windowed speech after pre-emphasis. Next, intensity-to-loudness conversion (also known as cubic root compression) is applied (i.e., loudness = intensity$^{1/3}$). This operation is an approximation to the power law of hearing and simulates the nonlinear relationship between the intensity of sound and its perceived loudness.

An approximation to the variable sensitivity across frequency at ca. 40 dB is given by equation (2) [3].

$$H(\omega) = 1.151 \cdot \sqrt{\frac{(\omega^2 + 144 \times 10^4)\omega^2}{(\omega^2 + 16 \times 10^4)(\omega^2 + 961 \times 10^4)}} \tag{2}$$

A filter bank can be regarded as a crude model for the transduction of the basilar membrane in the human auditory system. A mapping of acoustic frequency, *f,* to a perceptual frequency *mel* can be defined as

$$mel(f) = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \tag{3}$$

A set of 24 band-pass filters (whose bandwidth increases with frequency) models the basilar membrane. For simplicity, each band-pass filter is centered at the middle of the corresponding mel band and is triangular, starting and ending at the central frequencies of the adjacent *mel* bands.

A model for the transduction of mechanical motion of the basilar membrane to activity of auditory-nerve fibers is described here. The inner hair cells and auditory-nerve fibers are modeled as a transduction from loudness to firing rate. The Schroeder–Hall model [10], based on the generation and depletion of electrochemical "quanta" in a hypothetical inner hair cell, is consistent with neurophysiological phenomena. The model is defined as follows. The quanta of an electrochemical agent are generated in the inner hair cell at a fixed average rate, *r*. The probability of firing of an auditory-nerve fiber is directly proportional to the number of quanta currently existing and to the permeability function related to the instantaneous input stimulus level, *s(t)* (the square root of loudness). The quanta are used up by producing spontaneous firings, g_s, and a natural decay, g_d, without causing any firing. Thus Equation (4) describes the number of quanta as a function of time and the instantaneous firing rate *f(t)* of an auditory neuron:

$$\begin{cases} \dfrac{dn(t)}{dt} = r - (g_d + g_s + c \cdot s(t)) \cdot n(t) \\ f(t) = (g_s + c \cdot s(t)) \cdot n(t) \end{cases} \tag{4}$$

where *n(t)* is the number of quanta at time instant, *t*, and *c* is a constant. The discrete form of the above equation is given by the following iterative equation group:

$$\begin{cases} n(k) = \dfrac{r + n(k-1)}{1 + g_s + g_d + c \cdot s(k)} \\ f(t) = (g_s + c \cdot s(k)) \cdot n(k) \end{cases} \tag{5}$$

By applying a discrete cosine transform (DCT, see Equation (6)) to the firing rates from all the sub-channels, we obtain 13 de-correlated features which form the feature vector for one frame.

$$c_i = \sum_{j=1}^{24} f_j \cdot \cos\left(\frac{\pi \cdot i}{24}(j - 0.5)\right) , 0 \le i \le 12 \tag{6}$$

2.3 Parameters of the Auditory Model

The parameters (r, c, g_d, g_s) of the auditory front-end (equation (5)) are determined according to the relevant physiological data [1] [10]. Because *r* is a fixed firing rate, it can be set arbitrarily to one so that all other quantities are viewed as a fraction of this scale factor. The firing rate in response to a tone burst can be simulated as a decaying exponential form including steady (*A*) and transient (B) states.

$$n(t) = A + B \cdot e^{-t/T} \tag{7}$$

By differentiating Equation (7) and comparing with the result of Equation (4), the time constant, *T*, of the auditory model can be written:

$$T = 1/(g_d + g_s + c \cdot s(t)) \tag{8}$$

The time constant of fast adaptation is about 2 ms, which is too short to be significant in the frame-based features where the frame shift interval is typically around 10 ms. When the stimulus is turned off the firing rate recovers to the spontaneous rate with a time constant, T_0, of around 50 ms. From Equation (8), we have:

$$T_0 = 1/(g_d + g_s) \tag{9}$$

Another time constant is associated with the decreasing response to a stimulus which is a general characteristic of auditory neurons. It is reasonable to assume that the time constant, T_{max}, corresponding to the maximum input stimulus level, s_{max}, is around 30 ms. Substituting Equation (9) and the parameters into the Equation (8), yields:

$$c = \frac{1}{s_{max}}\left(\frac{1}{T_{max}} - \frac{1}{T_0}\right) \tag{10}$$

The remaining parameters to be solved are the spontaneous firing constant, g_s, and the decay constant, g_h. By observing the steady-state firing rate $f_s(t)$, we have:

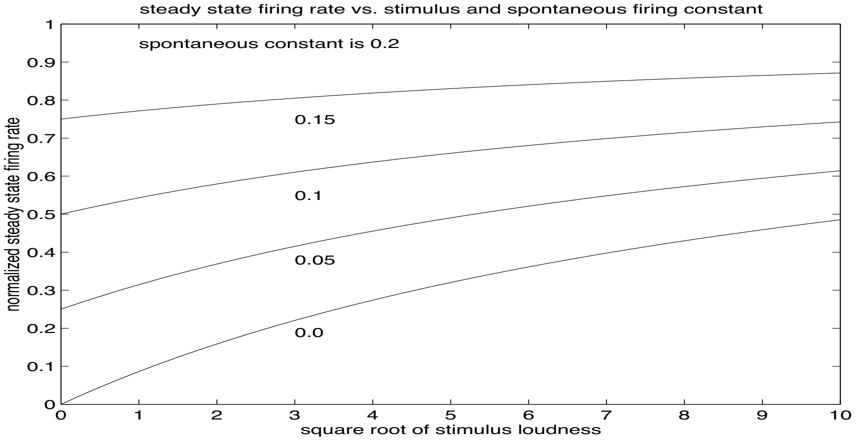


Figure 3 The normalized steady-state firing rate with respect to the square root of stimulus loudness and the spontaneous firing constant, g_s.

$$f_s(t) = r - \frac{r \cdot \frac{1}{T_0} - g_s}{\frac{1}{T_0} + c \cdot s(t)} \tag{11}$$

The dynamic range, β, can be defined as the ratio of the firing rate corresponding to the maximum stimulus and the spontaneous firing rate. We have:

$$\begin{cases} g = \dfrac{c \cdot s_{max} \cdot T_{max}}{\beta \cdot T_0 - T_{max}} \\ g = (1/T_0) - g_s \end{cases} \tag{12}$$

The parameters g_s and g_d can be obtained by setting the appropriate value for β in equation (12).

Figure 3 shows the relationship between the steady-state firing rate, the spontaneous firing constant and the square root of stimulus loudness. The spontaneous firing constant, g_s, can vary between 0 and 0.2. The low spontaneous-rate fibers were simulated with a high steady-state dynamic range when g_s is low, and the high-spontaneous-rate fibers with a narrow, steady-state dynamic range are generated when the high value of g_s was selected.

2.4 Two-Stream Approach

The spectrum of the feature vector components were studied in order to enhance the noise robustness of the auditory front-end. Figure 4 shows the ratio of the averaged spectra between the feature vector component trajectories of clean speech and car noise. A test set containing 110 sentences, spoken by seven male and four female speakers, was chosen from the TIMIT corpus. The ratio was computed by averaging across all 13 feature vector components over all the utterances. We observe that there is high local SNR in the low-frequency

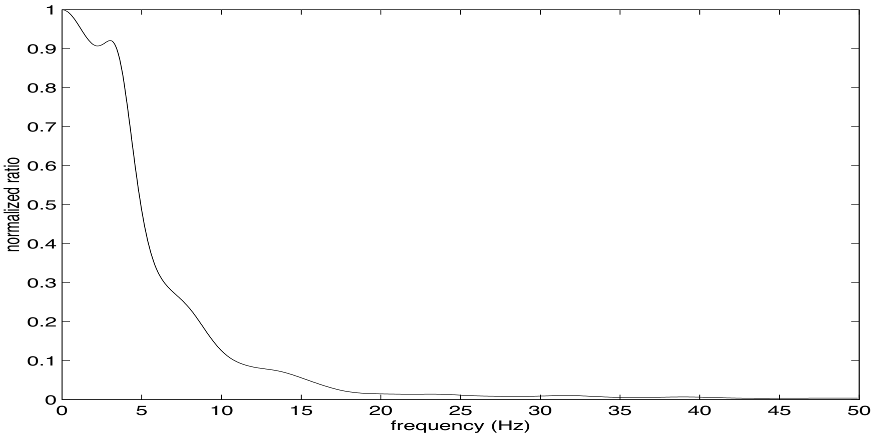


Figure 4 Ratio of the averaged spectra of the feature vector component trajectories between clean speech and car noise normalized to be one at zero frequency.

channels and that the local SNR decreases as the frequency increases. Furthermore, it has been shown that the frequency content beyond a certain frequency value of feature vector component trajectory of the speech contains a significant amount of estimation error [7]. In principle the front-end should be more noise robust if we can utilize information derived from the local SNR. The overall SNR could be increased by weighting the lower band more than the higher one. This approach also reduces the sharp peaks at the transitions produced by the short-term adaptation, resulting in parameter statistics that better fit our HMM framework.

In order to realize this weighting pattern the original feature stream, derived from a discrete cosine transform (DCT), is split into low- and high-frequency channels. These two channels are later recombined by proper weighting and subjected to normalization to form the final feature vector. Figure 5 shows the block diagram of the enhanced front-end. We assume that the transfer functions of the low-pass $H_l(z)$ and high-pass $H_h(z)$ filters are complementary, i.e.

$$H_l(z) + H_h(z) = 1 \tag{13}$$

The equivalent transfer function $H(z)$ of the recombined stream is given by the following equation:

$$w_l \cdot H_l(z) + w_h \cdot H_h(z) = 2\delta \cdot H(z) + (1 - \delta) \tag{14}$$

where δ ($-1 \le \delta \le 1$) is defined as a weighting factor, and the weights are given by $w_l=1+\delta$ and $w_h=1-\delta$. $H(z)$ is a low-pass, all-pass and high-pass filter when δ is 1, 0 and -1, respectively. Figure 6 shows the amplitude response of the low-pass, high-pass and combined filter with δ=0.4. Based on our experiments, the optimum cut-off frequency was found to be around 5 Hz. Figure 6 also shows the amplitude response of the conventional linear regression filter used to generate delta coefficients. It is clear from the figure that the new filter contains more high-frequency information than the conventional filter.

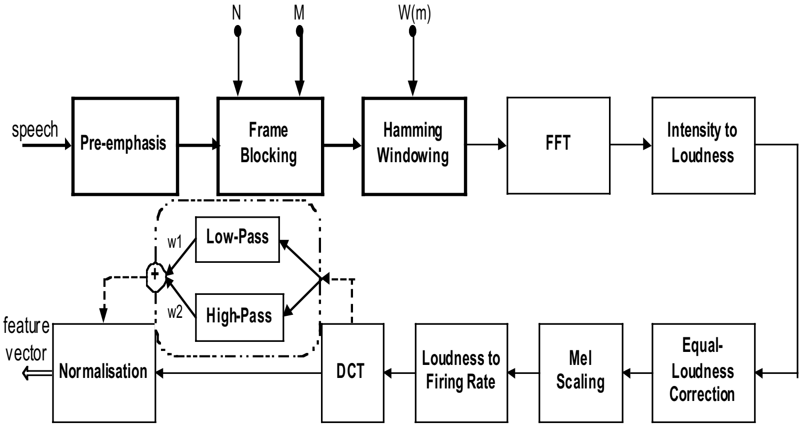


Figure 5 Block diagram of the enhanced auditory front-end. The blocks within the dashed portion show the enhancements.

3. Experiments

3.1 Observations on the TIMIT Database

The TIMIT corpus was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems including the front-end. We have randomly picked an utterance *sa1 ("she has your dark suit in greasy wash water all year")* in order to illustrate some simple comparisons between the MFCC and auditory front-ends.

The trajectories of the first components of the feature vectors generated by MFCC and auditory front-end representations are shown in Figure 7. It appears that the auditory front-end can capture dynamics better than the MFCC front-end. Specifically, the peaks at the transition portions of speech are emphasized and can be clearly observed. Though it looks like the auditory front-end might be capable of bringing some new information, it is not at all clear that it does so. In order to really know, we studied the separability between different phones at the feature level to base our conclusions on more statistical measures. It might be that the peaks do not match the Gaussian density assumption and performance is reduced.

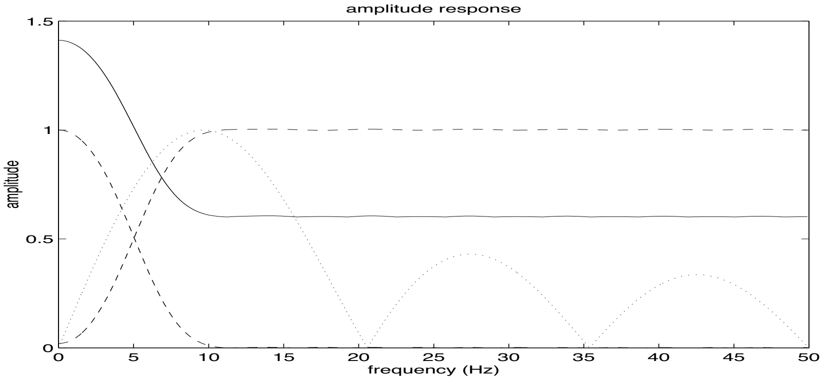


Figure 6 Frequency response of a low-pass. high-pass (both seen as dashed line), combined filter (solid line, d=0.4, cutoff frequency = 5 Hz) and the linear regression filter (dotted line, filter of length 7).

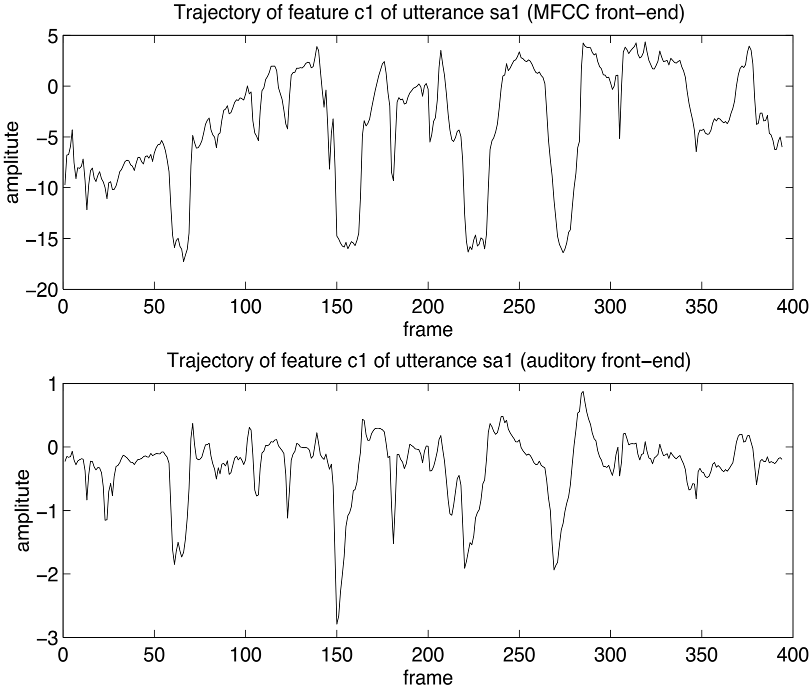


Figure 7 The feature trajectory of an utterance.

The separability of speech units in the feature space is a key indicator for evaluating the front-end. One of the J-measures, defined in equation (15), was used for this purpose.

$$J = \frac{tr(\boldsymbol{B})}{tr(\boldsymbol{W})} \tag{15}$$

where matrix **B** is the between-class covariance, or covariance of class means, and measures how close the speech classes are to each other. Matrix **W** is the within-class covariance, or the average of the class covariance. We applied the J-measure as the phonetic separability indicator to the test set of TIMIT database containing 1680 sentences in both clean and noisy conditions. Table 1 gives the results for both MFCC and auditory front-end.

Table 1 Separability values (J-measure) of phones in the test set of TIMIT database for MFCC and auditory front-ends, without normalization, for clean and noisy speech.

SNR	**clean**	**10 dB**	**0 dB**	**-10 dB**
Aud FE	0.7123	0.6062	0.5016	0.3337
MFCC	1.2215	0.7445	0.4691	0.2380

Based on the separability values for phones, it can be seen that the auditory front-end can provide better discrimination ability in adverse conditions than the MFCC front-end, but is worse in less noisy environments.

While many current speech recognizers provide rather good recognition accuracy in noise-free conditions, their performance degrades rapidly when they are exposed to noisy environments.

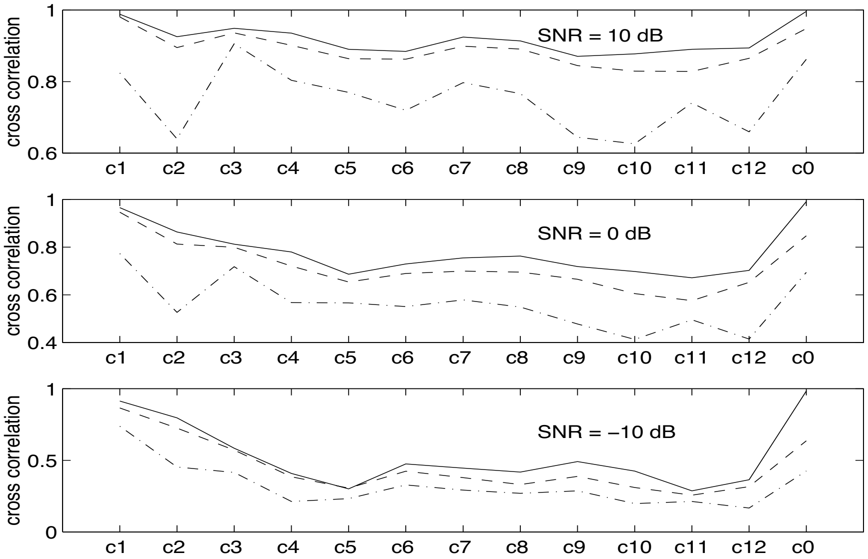


Figure 8 Similarity measure of the features between clean and noisy speech over a range of SNR conditions. The features were generated by the two-stream auditory (solid line), a previously proposed auditory model (dashed line) and MFCC (dash-dot line) front-ends.

In order to get noisy samples, noise from a Volkswagen car traveling at 115 km/h was recorded and further mixed with clean speech (again the utterance *sa1* is used for illustration) to generate noisy speech for different signal-to-noise ratios (SNR).

Figure 8 compares the cross-correlation between the clean and noisy features for the three front-ends at different SNRs. First, each feature was normalized by removing the mean and normalizing the variance to be one. With each SNR (10, 0 and -10 dB), the cross correlation was calculated between the clean and the corresponding noisy speech to measure their similarity. Obviously, if the cross-correlation is low, the features are heavily corrupted by the noise and, if the cross-correlation is high, it means the features are noise robust. In order to optimize the weighting factor for the two-stream approach, recognition experiments were initially carried out at different SNRs. The optimum weighting factor, δ, was found be around 0.4 ~ 0.6.

It is clear that the features produced by the two-stream auditory front-end are more noise robust than the features produced by the other front-ends, and the auditory front-end is more noise-robust than the MFCC front-end. We can also see that the features, c_1 and c_0, are less distorted among all features.

3.2 Isolated-Word Recognition Test

The final goal in any front-end development work is improved speech recognition accuracy. Improvements in the visual representation or in a certain phonetic separability measure are practically worthless without a noticeable difference in the back-end. We decided to test the auditory front-end in an isolated-word, speaker-dependent recognition task. The reason for such a test decision was that we have a high-performance, name-dialing engine which is very difficult to improve.

The test database contained 30 Finnish first names spoken by six male and two female speakers. The recordings were carried out in an office environment during three separate sessions (12 repetitions of each name overall).

Again, noise from a Volkswagen car traveling at 115 km/h was recorded and further mixed with clean speech to generate the noisy speech under certain signal-to-noise ratios (SNR).

Continuous Gaussian-density, left-to-right, state-duration-constrained, hidden Markov models (HMMs) with a global variance vector were estimated with a single training utterance [6]. Table 2 summarizes the results obtained with the auditory and MFCC front-ends without using the normalization block. It should be mentioned that the MFCC based front-end produced 13 cepstral coefficients including the energy value. It can be seen that the auditory front-end provides enhanced noise robustness, though with somewhat lower recognition accuracy in a clean environment.

Table 2 Recognition rates for MFCC and auditory front-ends, without normalization, for different noise conditions.

SNR	**clean**	**5 dB**	**0 dB**	**-5 dB**	**-10 dB**
Aud FE	97.42	91.02	81.63	58.90	27.12
MFCC	98.94	86.25	68.90	37.99	13.49

We have previously proposed a feature-vector normalization (FVN) method to enhance noise robustness of MFCC features [12]. With this normalization, short-term means and variances of each feature vector component are set to zero and one respectively, regardless of environment.

When this normalization is performed on the auditory features we hope that the sharp peaks in the trajectory of each feature vector component are suppressed and that the features become more suitable to the HMM framework (that is, fit better to unimodal Gaussian densities). We also hope that the other advantages of the normalization method, mentioned above, are also present in the auditory modeling case, and not just in the MFCC case. It should be noted that all the remaining experiments were carried out by using auditory or MFCC front-end with normalization.

Table 3 summarizes the results obtained with the auditory and MFCC front-ends with the normalization block enabled (see Figure 4). It can be seen that the auditory front-end still provides enhanced noise robustness. However, there is a marginal decrease in the recognition accuracy for the auditory front-end in the clean environment.

Table 3 Recognition rates for MFCC and auditory front-ends, with normalization, for different noise conditions

SNR	**clean**	**5 dB**	**0 dB**	**-5 dB**	**-10 dB**
Aud FE	99.09	97.12	93.64	84.55	58.98
MFCC	99.43	96.36	91.59	80.42	53.41

Time domain dynamics of speech can be incorporated into the MFCC by adding delta coefficients that are normally calculated with linear regression to estimate instantaneous derivatives (delta) for cepstral coefficients. All calculated delta parameters are appended to the feature vector. We compared the recognition performance between the auditory front-end

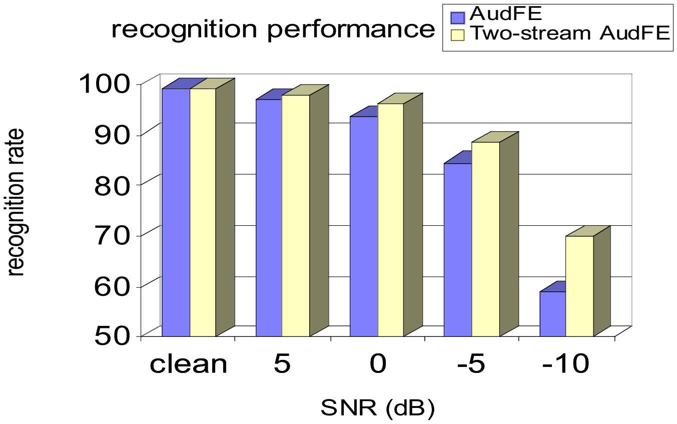


Figure 9 Recognition rates at different SNRs using an earlier form of auditory front-end (AudFE), as well as the two-stream auditory front-end (two-stream AudFE).

and MFCCs with delta information. In Table 4, it is shown that the MFCC with dynamics performs slightly better than the auditory front-end. However, it should be noted that the length of the feature vector using MFCC front-end is 26.

Table 4 Recognition rates for MFCC (with dynamic information, i.e.: MFCC26) and auditory front-ends, with normalization, for different noise conditions.

SNR	clean	5 dB	0 dB	-5 dB	-10 dB
Aud FE	99.09	97.12	93.64	84.55	58.98
MFCC26	99.73	97.58	94.58	86.55	63.11

Recognition tests were carried out between the previously proposed auditory front-end and the two-stream auditory front-end to compare their performance. Both approaches had a feature vector dimension of 13.

Figure 9 shows the recognition results using the two front-ends. It is clearly seen that the two-stream auditory front-end outperforms the previously proposed auditory front-end. The average error rate reduction, over all noise conditions, was found to be around 27%.

We also compared the two-stream auditory front-end approach with the standard MFCC front-end. Figure 10 presents the results for the MFCC front-end with only static features (MFCC13) and also with both static and delta features (MFCC26). Delta-delta coefficients were not used in these speaker-dependent tests, as they produced worse results as compared to MFCCs with statics and deltas. It can be clearly seen that the two-stream auditory approach outperforms the MFCC front-ends in all noisy conditions. There is an average error-rate reduction of 39% and 17% for the new approach over MFCC13 and MFCC26, respectively. However there seems to be a small decrease in the recognition performance in clean conditions

Finally, the superiority of the two-stream approach is demonstrated by comparing it to the previously proposed auditory front-end with delta features, computed using linear regression, thereby having a feature vector dimensionality of 26. It can be seen from Table 5

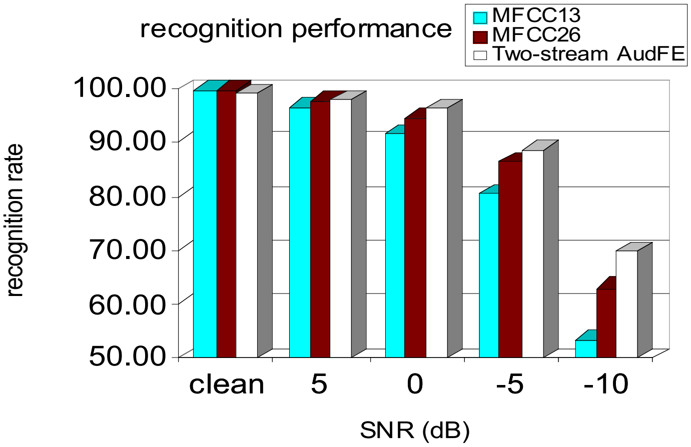


Figure 10 Recognition rates at different SNRs using a MFCC front-end with only static features (MFCC13). An MFCC front-end with static and delta features (MFCC26) and a two-stream auditory front-end.

that the proposed two-stream auditory front-end, with a feature vector dimensionality of 13, produces better recognition accuracy than the auditory front-end with delta features.

Table 5 Recognition rates obtained with different front-ends at different SNRs

SNR	MFCC13	MFCC26	AudFE	AudFE26	tsAudFE
clean	99.43	99.73	99.09	99.02	99.13
5	96.36	97.58	97.12	97.39	97.88
0	91.59	94.58	93.64	95.61	96.14
-5	80.42	86.55	84.55	88.60	88.45
-10	53.41	63.11	58.98	67.99	69.92
Ave.	84.24	88.31	86.67	89.72	90.30

4. Conclusions

The auditory front-end proposed in this chapter incorporates a number of auditory properties pertaining to the inner ear. It captures dynamic features, such as onsets and offsets, that produce observable peaks at the transient components of the signal, thus negating the need for incorporating separate features representing dynamic information as are required for an MFCC front-end. In this chapter it has been shown that the auditory front-end performs better than the MFCC front-end under all conditions except the clean (i.e., highest SNR) environment. In addition, we have shown that the normalization method proposed earlier for the MFCC front-end also improves the performance of the auditory front-end, especially in noisy environments. This performance improvement is due to the fact that the normalization procedure reduces the mismatch between training and testing environments. Applying the normalization to the auditory front-end can suppress the sharpness of the peaks and make the features more suitable to fit within the HMM scheme.

We have proposed a new noise-robust, two-stream auditory feature-extraction method. Speaker-dependent, isolated-word recognition tests performed using the new approach show that this front-end outperforms the previously proposed auditory and conventional MFCC front-end in terms of recognition accuracy in all noisy environments.

Incorporating other auditory phenomena should be considered as a means of improving recognition performance. Our current research has focused on modeling the auditory peripheert. We will henceforth concentrate on trying to incorporate several phenomena associated with the central auditory system. It is also important to ensure that the resulting auditory front-end integrates well within the HMM framework used in automatic speech recognition systems.

References

[1] Cohen, J. "Application of an auditory model to speech recognition." *J. Acoust. Soc. Am.*, 85: 2623–2629, 1989.

[2] Ghitza, O. "Auditory model and human performance in tasks related to speech coding and speech recognition." *IEEE Trans. Speech Audio Proc.*, 2:115–132, 1994.

[3] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech." *J. Acoust. Soc. Am.*, 87:1738–1752, 1990.

[4] Jankowski, C., Vo, H. and Lippmann, R. "A comparison of signal processing front ends for automatic word recognition." *IEEE Trans. Speech Audio Proc.*, 3: 286–293, 1995.

[5] Kates, J. "A time-domain digital cochlear model." *IEEE Trans. Signal Processing*, 39: 2573–2592, 1991.

[6] Laurila, K. "Noise robust speech recognition with state duration constraint." *Proc. Intern. Conf. Acoustics, Speech, Sig. Proc.*, pp. 871–874, 1997.

[7] Nadeu, C., Paches-Leal, P. and Juang, B. "Filtering the time sequences of spectral parameters for speech recognition." *Speech Com.*, 22: 315–332, 1997.

[8] O'Shaughnessy, D. *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley, 1987.

[9] Picone, J. "Signal modeling techniques in speech recognition." *Proc. IEEE*, 81:1215–1247, 1993.

[10] Schroeder, M. and Hall, J. "A model for mechanical to neural transduction in the auditory receptor." *J. Acoust. Soc. Am.*, 55:1055–1060, 1974.

[11] Seneff, S. "A joint synchrony/mean-rate model of auditory speech processing." *J. Phonetics*, 16: 55–76, 1988.

[12] Viikki, O. and Laurila, K. "Cepstral domain segmental feature vector normalization for noise robust speech recognition." *Speech Comm.*, 25:133–147, 1998.

Computational Models of Auditory Function
S. Greenberg and M. Slaney (eds.)
IOS Press, 2001

STRAIGHT: AN EXTREMELY HIGH-QUALITY VOCODER FOR AUDITORY AND SPEECH PERCEPTION RESEARCH

Hideki Kawahara

Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, Wakayama 640-8510, Japan / ATR / CREST

1. Introduction

It would be interesting to develop a system to manipulate the underlying parameters of speech signals without introducing quality degradation due to the manipulations. Such a system would be useful for investigating human speech perception capabilities using close to natural speech stimuli while enabling precise control on relevant parameters. In other words, it would possibly allow us to investigate normal hearing under the operating conditions for which our auditory system is designed.

It is generally difficult to predict the behavior of highly nonlinear processes, such as auditory perception, for very complex stimuli (for example, speech sounds) only from responses to simple elementary stimuli. Investigations in the vicinity of natural speech examples with precisely controlled deviations from the original, however, can provide complementary clues to understanding “hearing” better. It is because, even though speech signals are complex mixtures of numerous components and are very different from usual psychophysical stimuli, the precisely controlled deviations can be designed to be parameterized with a small number of psychophysically meaningful parameters.

A versatile speech manipulation method called STRAIGHT has been developed [2] aiming at fulfilling the requirement outlined above. STRAIGHT is based on a concept introduced by the channel VOCODER [1], which separates spectral envelope information and source information such as periodicity. This separation is useful in designing experiments for investigating the physical correlates of perceptual attributes provided that such a separation does not introduce perceptible degradations.

Speech coding procedures using an analysis and synthesis scheme like the VOCODER have been understood to have a rather low upper limit of speech quality. It is widely believed that the original speech waveform needs to be replicated to produce highly natural speech. The highly natural re-synthesized speech achieved by our STRAIGHT procedures, however, provide a counterexample to these conventional views and make the proposed method a powerful research tool.

2. Basic Concepts

STRAIGHT consists of three key procedures to achieve its goal. The first procedure extracts a smoothed time–frequency representation, which is free from interference due to the source periodicity. This procedure uses pitch-adaptive, time–frequency analysis combined with a surface reconstruction method in the time–frequency region. The second component extracts F_0 and other source related information with high reliability and precision. It extracts the speech F_0 as the instantaneous frequency of the fundamental component of complex sounds like voiced speech, by using a new concept called "fundamentalness." "Fundamentalness" is defined as the negative logarithm of the total amount of AM (amplitude modulation) and FM (frequency modulation) magnitudes of a wavelet transform using an auditory-like analyzing wavelet. The third procedure designs the excitation source for resynthesis using group-delay manipulations, and this enables artificial "naturalness" to be added to the synthetic speech. This procedure takes advantage of the fact that humans are very sensitive to specific group-delay attributes.

2.1 Elimination of Periodicity Interferences

A periodic signal can be represented as the convolution of a unit waveform and a periodic pulse train. For speech a unit waveform can be modeled as the convolution of a single glottal source waveform and the impulse response of the vocal tract and acoustic radiation. Because articulatory organs generally move, the spectral representation of a unit speech waveform will vary with time. As a result, plotting the magnitude of such a spectral representation in the time–frequency plane, can provide a smooth three-dimensional surface *S(w,t)*.

Due to the second convolution with the pulse train, however, the original surface, *S(w,t),* is not directly observable. The time–frequency representations of such signals can suffer from interference caused by the signal periodicity. In other words, only partial information about the surface, *S(w,t),* is available. Therefore, the goal of the current procedures is to recover this hypothetical surface *S(w,t)* using partial information sampled at every τ_0(fundamental period) in the time domain and every $F_0=1/\tau_0$ in the frequency domain.

A spectrogram $|F(w,t)|^2$ calculated using a short-term Fourier transform exhibits a regular structure reflecting the signal periodicity in both the time and frequency domains. If it were possible to derive a spectrogram not having the regular interfering structure in the time domain, the problem of reconstructing the time–frequency surface would be reduced to the problem of eliminating the regular interfering structure in the frequency domain.

2.1.1 Power Spectrum with Reduced Phasic Interference

A practical solution for calculating a temporally stable short-term Fourier transform for a periodic signal has been to use a rectangular pitch synchronous window or to use a relatively long time window spanning three or more pitch periods. However, these alternatives do not work very well for speech signals, because speech is not purely periodic nor stable in time.

One reasonable selection of the time window is to use an adaptive time window, which has comparable time and frequency resolution in terms of the signal periodicity. A spectrogram using such a window generally has regular a "hole" both in the time and frequency domains. This temporal variation of the spectrogram is mainly due to the phasic interference between neighboring harmonic components.

The compensatory time window, $w_c(t;\ \eta,\ \tau_0)$, is designed to exhibit interference characteristics complementary to the original Gaussian window $w_o(t;\ \eta,\ \tau_0)$, where η represents the temporal stretching factor.

$$w_o(t;\eta,\tau) = \frac{1}{\eta\tau_0} e^{-\frac{\pi t^2}{\eta^2 r_0^2}} \tag{1}$$

$$w_o(t;\eta,\tau) = \frac{1}{\eta\tau_0} e^{-\frac{\pi t^2}{\eta^2 r_0^2}} \sin\frac{\pi t}{\tau_0} \tag{2}$$

A spectrogram with reduced phasic interference $|F_r(w,t)|^2$ is represented as the weighted squared sum of spectrograms $|F_c(w,t)|^2$ and $|F_o(w,t)|^2$ using this compensatory window and the original time window, respectively.

$$|F_r(w,t)|^2 = |F_o(w,t)|^2 + \xi(\eta)|F_c(w,t)|^2 \tag{3}$$

Here, $\xi(\eta)$ represents the optimum mixing factor that minimizes the temporal variation of $|F_r(w,t)|^2$.

2.1.2 Pitch-Adaptive Spectral Smoothing

A second-order, cardinal B-spline smoothing function $h(w)$, as defined below, is employed to eliminate the periodicity interference in the frequency domain.

$$h(w;w_0) = \frac{w_0 - |w|}{w_0^2} \tag{4}$$

where $w_0(t)=2\pi f_0(t)$ and $-w_0(t) \geq w \geq w_0(t)$. Since the fundamental angular frequency $w_0(t)$ is a function of time, the smoothing function is adaptive to the fundamental frequency. The smoothed spectrogram at time t is calculated using this smoothing kernel and the spectrogram with reduced phasic interference.

$$|S(w,t)|^2 = g^{-1}\left(\int_{-w_0(t)}^{w_0(t)} h(\lambda) g\left|F_r(w-\lambda,t)\right| d\lambda\right) \tag{5}$$

This smoothing operation is equivalent to piecewise linear interpolation, when the original spectrum is represented as a line spectrum. Note that the proposed operation is local and this makes the procedure less sensitive to F_0 errors and noise.

In Equation 5, $g(\ )$ defines what quantity is to be preserved through the smoothing operation. For example, the identity mapping, $g(x) = x$, preserves the energy of the signal and the 1 /3 power law, $g(x) = x^{1/3}$, preserves the perceived loudness, approximately.

2.2 Reliable and Precise F_0 Extraction

A new algorithm based on the instantaneous frequency has been developed to provide source information in order to guide the STRAIGHT procedures. The proposed method extracts the fundamental frequency as the instantaneous frequency of the fundamental component of a complex sound. This may sound strange, because selecting the fundamental component seems to require a prior knowledge about the fundamental frequency to be extracted. However, a new measure called “fundamentalness” provides a built-in mechanism for selecting the fundamental component without referring to F_0 information.

This “fundamentalness” is designed to have the maximum value when the filter output only consists of the fundamental component. It is possible to use a bank of asymmetric constant-Q band-pass filters; each filter has a gradual slope for the lower cut-off and a steeper slope for the higher cut-off. By defining the "fundamentalness" to be proportional to the negative logarithm of the total amount of FM and AM of the filter output by using the filter bank mentioned above, the desired behaviors can be shown.

An analyzing wavelet, $w_{AG}(t;\eta)$, made from a complex Gabor filter, $w_g(t)$, having a slightly finer resolution in frequency *(i.e.,* $\eta > 1$*)* can form such a filter bank. The input signal, *s(t),* can be divided into a set of filtered complex signals *B(t; τc).*

$$B(t,\tau_c) = |\tau_c|^{\frac{-1}{2}} \int_{-\infty}^{\infty} s(t) w_{AG}\left(\frac{t-u}{\tau_c}\right) du \tag{6}$$

$$w_{AG}(t;\eta) = w_g\left(t - \frac{1}{4};\eta\right) - w_g\left(t + \frac{1}{4};\eta\right) \tag{7}$$

$$w_g(t;\eta) = \frac{1}{\eta} e^{\frac{-\pi t^2}{\eta^2}}$$

The characteristic period of the analyzing wavelet is used to represent the corresponding filter channel.

The "fundamentalness" index, $M_c(t;\tau_c)$, is calculated for each channel (τ_c) based on the output. The definition of the index has been slightly modified from a previous report, because the F_0 trajectories of speech signals normally consist of moving components that carry prosodic information. Removing the contribution of the monotonic F_0 movement reduces artifacts in the "fundamentalness" evaluation caused by prosodic components.

$$\begin{aligned} M_c(t;\tau_c) = & -\log\left[\int_{\Omega}\left(w\frac{(d|B(u)|)}{du} - \mu_{AM}(u)\right)^2 du\right] \\ & -\log\left[\int_{\Omega} w\left(\frac{d^2}{du}\arg((B(u)) - \mu_{FM}(u))\right)^2 du\right] \\ & + \log\left[\int_{\Omega} w|Bu|^2 du\right] + 2\log\tau_c \end{aligned} \tag{8}$$

$$\mu_{AM}(t) = \int_{\Omega} w(u-t;\tau_c)\left(\frac{d|B(u)|}{du}\right) du \tag{9}$$

$$\mu_{AM}(t) = \int_{\Omega} w(u-t;\tau_c)\left(\frac{d^2}{du}\arg((B(u)) - \mu_{FM}(u))\right) du \tag{10}$$

$$w(u-t;\tau_c) = \frac{1}{\sqrt{2}\tau_c} e^{\frac{-\pi t^2}{2\eta^2}} \tag{11}$$

where the integration interval $\Omega=(t-T, t+T)$ is selected to cover the range where the weighting factor $w(u\text{-}t;\tau_c)$ (in Equation(8), the first three terms, $w(u\text{-}t)$, are abbreviated as w) is effectively non-zero. Therefore, index $M_c(t;\tau_c)$ is normalized in terms of the scale. Extracting F_0 simply means finding the maximum index of $M_c(t;\tau_c)$ in terms of τ_c and calculating the average (or more specifically, interpolated) instantaneous frequency using the outputs of the channels neighboring τ_c.

For a discrete time system with f_s as the sampling frequency, the instantaneous frequency $f_i(\mathrm{t}, \tau_c)$ of the output of channel, τ_c is calculated using the following equation:

$$f_i(t;\tau_c) = \frac{f_s}{\pi} arc \sin\frac{|\Delta B(t;\tau_c)|}{2|B(t;\tau_c)|} \tag{12}$$

where Δ represents the differentiation operator.

2.3 Excitation Source Design

There are two different ways of resynthesizing speech using the extracted smoothed spectrogram and fundamental frequency information. They are a source-filter implementation and a sinusoidal representation.

In this chapter, the former is used to resynthesize the speech signal from the time–frequency representation and periodicity information. In this implementation, the minimum phase impulse responses calculated from spectral slices of the time–frequency representation are used as a time-varying filter.

In a source-filter model, the extracted f_0 (in fine resolution) is used to re-synthesize the speech signal, *y(t),* using the following equation:

$$\begin{aligned} y(t) &= \sum_{t_i \in Q} \frac{1}{\sqrt{G(f_0(t_i))}} v_{ti}(t - T(t_i)) \\ v_{ti}(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} V(w, t_i)\Phi(w)e^{jw(t)}dw \\ \text{where} \quad T(t_i) &= \sum_{t_i \in Q, k < i} \frac{1}{\sqrt{G(f_0(t_k))}} \end{aligned} \tag{13}$$

where Q represents a set of positions in the excitation for the synthesis, and $G(\ )$ represents the pitch modification. The all-pass filter function, $\Phi(w)$, is used to control the fine pitch and the temporal structure of the source signal and is described in the next section. In the discrete-time system the range of integration in the equation to derive $v_t(\tau)$ becomes $[-\pi, \pi]$ using the normalized angular frequency $w = 2\pi f/f_s$, where f_s represents the sampling frequency.

$V(w, t_i)$ represents the Fourier transform of the minimum phase impulse response, which is calculated from the modified amplitude spectrum, $A(S(u(w), r(t)), u(w), r(t))$, where $A(\ )$, $u(\ )$, and $r(\ )$ represent manipulations in the amplitude, frequency, and time axes, respectively.

$$\begin{aligned} V(w, t) &= \exp\left(\frac{1}{\sqrt{2\pi}}\int_0^{\infty} h_t(q)e^{jwq}dq\right) \\ h(q) &= \begin{cases} 0 & (q < 0) \\ c_t(0) & (q = 0) \\ 2c_t(q) & (q > 0) \end{cases} \\ \text{and} \quad c_t(q) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{jwq}\log A dw \end{aligned} \tag{14}$$

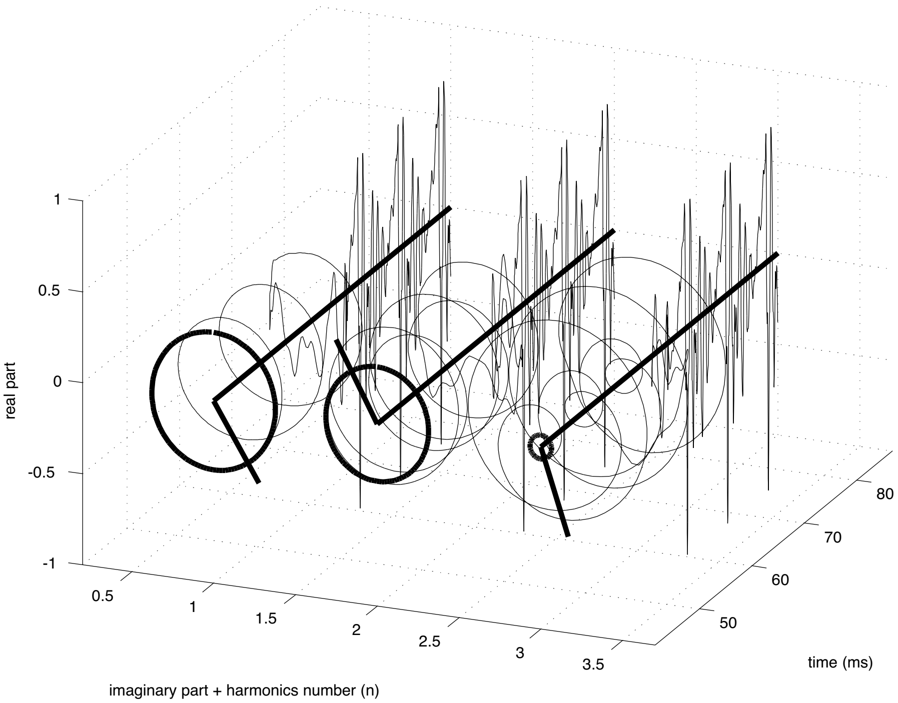


Figure 1 Illustration of the method for the vowel [a] pronounced by a male speaker. Shown are channels centered on F_0, $2F_0$ and $3F_0$. For each, the waveform input to the filter is plotted in perspective, followed by a polar representation of the complex output. The radius of the thick circle represents the instantaneous amplitude, which is constant for the filter centered on F0 and pulsating for those centered on $2F_0$ and $3F_0$.

$$v_t(\tau) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} V(w,t)\Phi(w)e^{jwq}dw \qquad (15)$$

where q represents the quefrency.

The all-pass filter function *Φ(w)* allows the control of the excitation source waveform by group delay manipulation. This is necessary because while there is no degradation in speech quality caused by parameter manipulations using the STRAIGHT procedures, there is still some initial degradation in quality under headphone listening when no temporal fine structure control is employed.

The all-pass filters used here have the following form:

$$\Phi(w) = \exp\left(-j\int_0^w \rho(\lambda)\tau_g(\lambda)d(\lambda + \alpha w)\right) \qquad (16)$$

where random group delay $\tau_g(\ )$ is made from band-limited gaussian white noise. The αw term is introduced for fine pitch control.

The following equation is used to shape the group-delay frequency characteristics, where w_c represents the lower-boundary frequency where the group delay dispersion starts. The parameter, b_w, defines the width of the transitional area:

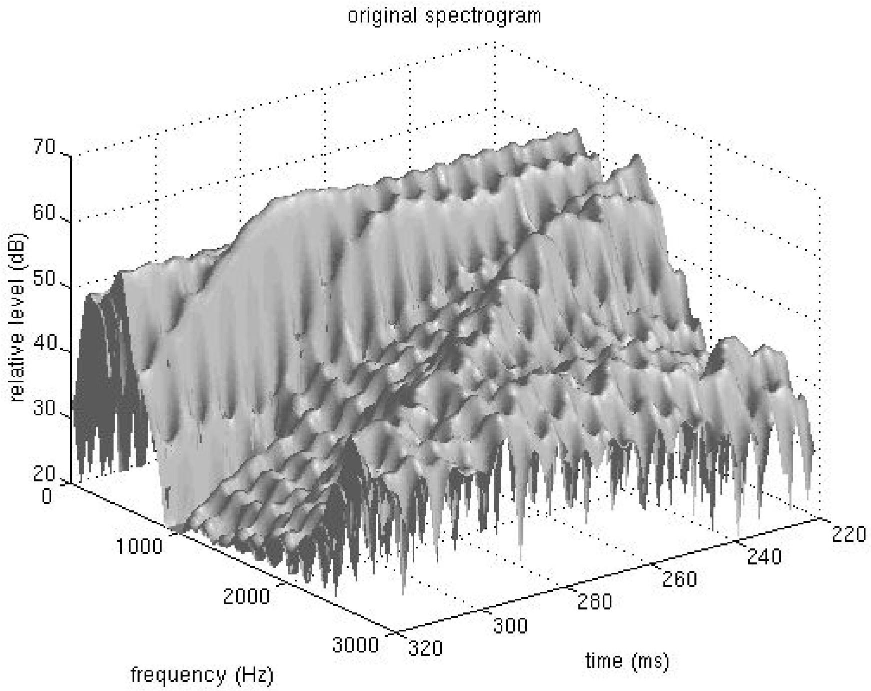


Figure 2 Spectrogram of a female's pronunciation of "right" using a pitch-adaptive window.

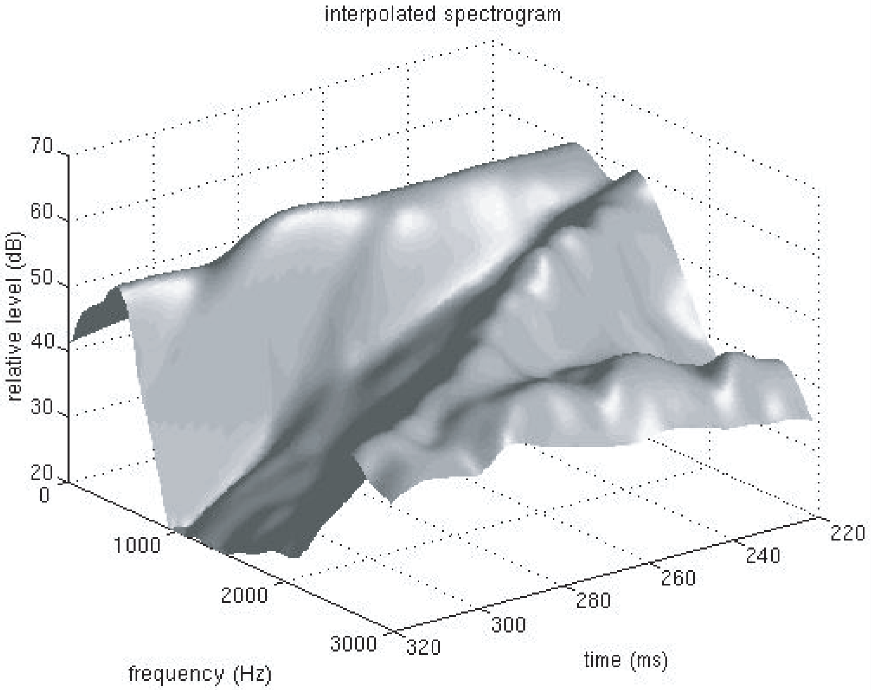


Figure 3 Smoothed spectrogram of a female's pronunciation of "right" using a pitch-adaptive smoothing operation.

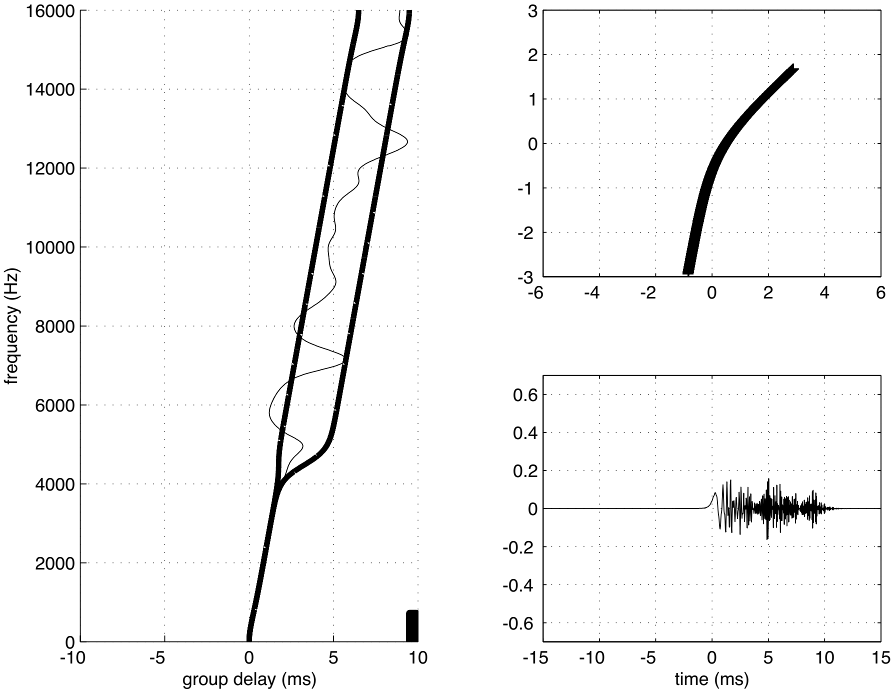


Figure 4 All-pass filter example. The left plot represents a group-delay function designed from random numbers with frequency weighting (thick lines). The top right plot shows a mapping function to introduce group delay asymmetry. The bottom right plot shows the corresponding impulse response.

$$\rho(w) = \frac{1}{1 + \exp\left(\frac{- w - w_c}{b_w}\right)} \tag{17}$$

Detailed discussions of group delay is given in the literature. In addition, in a preliminary test using temporally asymmetric group-delay functions, it was revealed that humans are sensitive to the identity of the asymmetry and that the asymmetry introduces an interesting timbre [3].

2.4 Numerical Examples

Figure 1 illustrates how the outputs of auditory-like filters behave. Only the filter output corresponding to F_0 exhibits a stable behavior without AM and FM.

Figure 2 shows a 3-dimensional plot of a pitch-adaptive spectrogram with interference due to periodicity. Figure 3 shows a smoothed spectrogram for the same speech material. The interference caused by the signal periodicity is removed while the general spectral shape is kept intact. This illustrates that the basic concept of the time–frequency smoothing of STRAIGHT is effective.

Table 1 Test conditions for re-synthesis.

Symbol	Condition
AD	Original sound
Prev	Control with additional √40dB white noise
SYN	Optimal smoothing
EH	Temporal processing 1
EH2	Temporal processing 2

3. Improving the Reproduction Quality

The original implementation described above provides reasonably high-quality reproduction via loudspeakers; however, significant degradation is still perceived when headphones are used. This section introduces several post-processing methods for improving reproduction quality.

3.1 Post-processing of the Spectral Envelope

The smoothed time–frequency representation obtained using the second-order cardinal B-spline smoothing kernel was found to introduce speech quality degradation due to over-smoothing, because the smoothing effect of time windowing was not taken into account in the original implementation. It was additionally found that the degradation due to this over-smoothing was more salient in female speech.

Two steps were introduced into the reproduction procedure to solve this over-smoothing problem. First, we used a second-order cardinal B-spline for the smoothing, so it was possible to design an optimal smoothing function that, at least for knot points, could compensate for any over-smoothing. This reduced the problem to an inverse filtering problem (i.e., recovering the original impulse from the smoothed impulse).

The optimum smoothing coefficient vector, **c**, is calculated from the target unit impulse vector **u** and the coefficients matrix ***H*** made from the over-smoothed unit impulse (i.e., $\{\mathrm{H}\}_{kl} = v_{k+l}$, where v is the over-smoothed vector):

$$\mathbf{c} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{u} \qquad (18)$$

$$\mathbf{u} = [u_{-M}, u_{-M+1}, \ldots, u_{0}, \ldots, u_{M-1}, u_{M}]'$$

$$\mathbf{c} = [c_{-N}, c_{-N+1}, \ldots, c_{0}, \ldots, c_{N-1}, c_{N}]'$$

The second step for recovering from the over smoothing is to compensate for any excessive decay in the vicinity of the point of excitation. Based on an approximation of the decay effect using a Taylor expansion up to the t^2 term, the following weighting function is introduced to compensate for the total decay effect caused by the time windowing and the optimum smoothing function represented in an alternative form using parameters $b_1, \ldots, b_k$. These parameters are recursively calculated from the elements of **c**:

$$q_{w0}(t) \approx 1 + \zeta\left(\frac{\pi^2}{3}(b_1 + 4b_2 + \ldots + k^2 b_k) + \frac{\pi}{\eta^2}\right) f_0^2\ t^2 \qquad (19)$$

where ζ represents an additional controlling factor. This operation in the time domain is equivalent to enhancing the spectral envelope by adding its second-order derivative to the original in the frequency domain. However, the time-domain procedure is preferable,

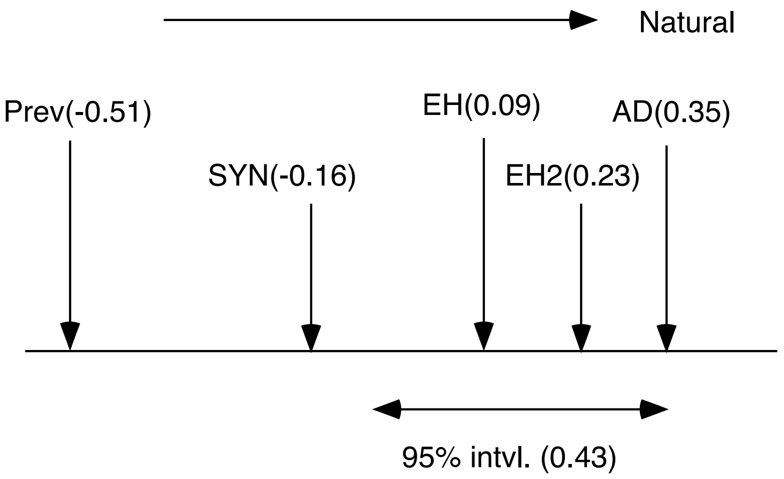


Figure 5 Effects of time-domain processing. The horizontal axis represents the psychometric scale calculated based on Thurstone's case V.

because it makes it easier to implement pitch synchronous variation of the vocal tract transfer function in the time domain.

3.2 Experiment

A preliminary experiment was conducted to test the effects of spectral and temporal post-processing. Table 3.2 represents the resynthesis conditions tested. Because the reproduced speech quality was almost equivalent to the original natural speech, the 2AFC procedure (paired comparison with two alternative forced choice) was used. The test material was a female's utterance of "kousyou ni oware te imasu" (busy in making negotiations), sampled at 24 kHz with 16-bit resolution.

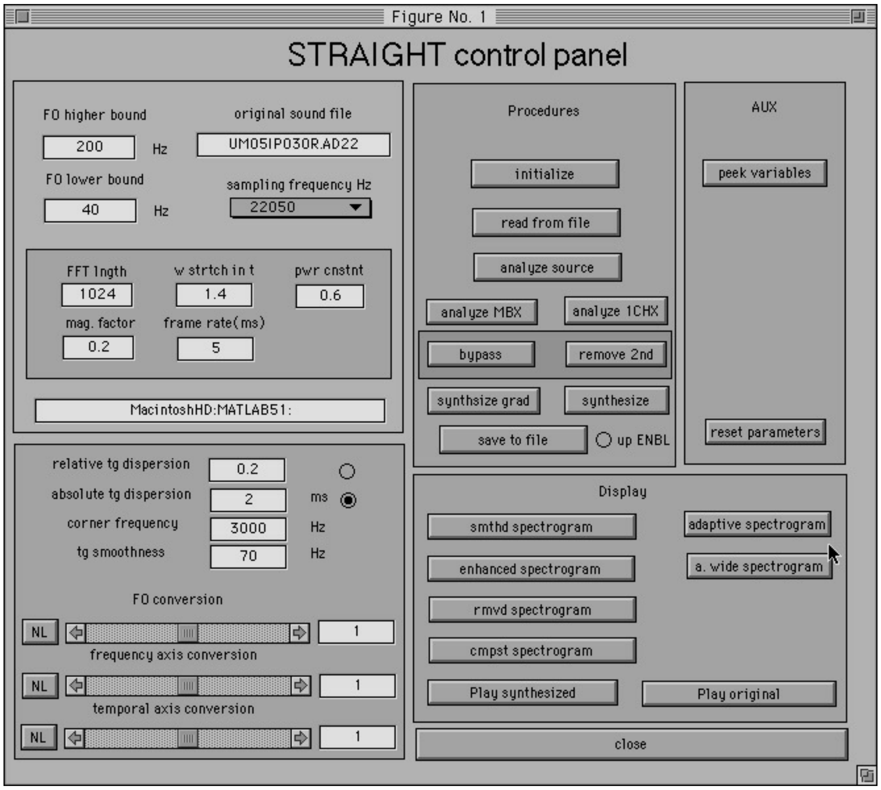
Figure 5 shows the results obtained on the psychometric scale calculated by Thurstone's case V procedure. The total number of subjects was 10.

The best re-synthesized speech is almost indistinguishable from the original in terms of "naturalness." Note that a similar synthetic sound with some (40 dB S/N) additional white noise was rated the worst.

4. GUI and Implementation

All the STRAIGHT procedures (one of them is described in the previous section) are implemented in MATLAB and its signal processing toolbox. MATLAB provides portability among various platforms and also provides accessibility to internal variables. These features make STRAIGHT a flexible tool. Furthermore, the recent introduction of GUI has made it easy to use STRAIGHT.

Figure 6 shows the GUI-based control panel of STRAIGHT. The top center sub-panel is for general procedures. The usual way of using STRAIGHT is to click buttons from top to bottom in this center panel. This represents the standard ordering of constituent procedures for manipulating speech. The bottom right sub-panel has a collection of buttons to display information about the speech sample under inspection. The sub-panel also provides audio monitoring of the signal. Parameters mainly used in the analysis stage are accessible from the top left sub-panel and are controllable using the "edit" and "menu selection" GUI-primi-



F0 higher bound 200 Hz

original sound file UM05IP030R.AD22

F0 lower bound 40 Hz

sampling frequency Hz 22050

FFT lngth 1024

w strtch in t 1.4

pwr cnstnt 0.6

mag. factor 0.2

frame rate(ms) 5

MacintoshHD:MATLAB51:

relative tg dispersion 0.2 ○

absolute tg dispersion 2 ms ●

corner frequency 3000 Hz

tg smoothness 70 Hz

F0 conversion 1

frequency axis conversion 1

temporal axis conversion 1

Figure 6 STRAIGHT control panel.

tives of MATLAB. Parameters used in the synthesis stage are accessible from the bottom left sub-panel and are controllable using the “edit,” “slider,” and “radio button” GUI-primitives of MATLAB. Any nonlinear arbitrary mapping of an original and the synthesis parameters can be controlled using a direct manipulation controller.

Users with MATLAB knowledge can use component procedures of STRAIGHT to construct programs for their specific purpose, because thy are implemented as function subprograms for general use.

5. Summary and Conclusions

A new set of simple procedures called STRAIGHT has been introduced to provide tools for speech perception research using artificial stimuli that sound highly natural. The proposed method is a speech analysis, modification and synthesis system that provides naturally sounding manipulated speech even with a large number of parameter modifications. The proposed method uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time–frequency region, and an excitation source design based on group-delay manipulation. It also consists of a pitch-extraction method using instantaneous frequency calculation based on a new concept called “fundamentalness.”

STRAIGHT has revived the underlying concept of the channel VOCODER to implement a powerful research tool for speech perception, primarily because, it has revealed that it is not necessary to replicate waveforms to re-synthesize highly natural speech. This method allows researchers to control spectral envelope characteristics and source related parameters independently and precisely without introducing artificial degradation. Experiments using STRAIGHT will provide complimentary clues to existing psychophysical data in analyzing the properties of the human auditory system pertinent for speech processing.

Acknowledgements

This work is a part of the "Computational Auditory Brain" project supported by CREST (Core Research for Evolutional Science and Technology) of the Japan Science and Technology Corporation, a subsidiary of the Science and Technology Agency of the Japanese government. The author would like to thank Ms. Ikuyo Masuda-Katsuse and Dr. Alain de Cheveigne for their contributing discussions.

References

[1] Dudley, H. "Remaking speech." *J. Acoust. Soc. Am.*, 11: 169–177, 1939.

[2] Kawahara, H. "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited." *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc.*, pp. 1303–1306, 1997.

[3] Kawahara, H., Tsuzaki, M. and Patterson, R. D. "A method to shape a class of all-pass filters and their perceptual correlates." *Tech. Com. Psycho. Physio., Acoust. Soc. Jpn.*, H-96-79:1–8, 1996. [Japanese].

[4] Kawahara, H., Katsuse-Masuda, I. and de Cheveigne, A. "Restructuring speech representations using STRAIGHT: The possible role of a repetitive structure in sounds." *Speech Comm.*, 27: 187–207, 1999.

Subject Index

Author Index