SPEECH DISCRIMINATION BASED ON MULTISCALE SPECTRO–TEMPORAL MODULATIONS

Nima Mesgarani, Shihab Shamma

Neural System Labratory
University of Maryland
College Park, MD 20742, USA

Malcolm Slaney

IBM Almaden Research Center
San Jose, CA 95120, USA

ABSTRACT

A novel approach for content based audio classification is presented based on multiscale spectro-temporal modulation features extracted using a model of auditory cortex. The task is to discriminate speech from non-speech which consists of animal vocalizations, music and environmental sounds. Generalization of the system to signals in high level of additive noise and reverberation is evaluated and compared to two existing approaches. The results demonstrate the advantages of the auditory model over the other two systems, especially at low SNRs and high reverberation.

1. INTRODUCTION

Audio segmentation and classification have important applications in audio data retrieval, archive management, modern human–computer interfaces, and in entertainment and security tasks. In speech recognition systems designed for real world conditions, a robust discrimination of speech from other sounds is a crucial step. Speech discrimination can also be used for coding or telecommunication applications where non-speech sounds are not of interest and hence bandwidth is saved by not transmitting them or by assigning them a low resolution code.

Two state-of-the-art systems have been proposed, against which we shall compare our system. The first is proposed by Scheirer and Slaney [1] in which thirteen features in time, Frequency, and cepstrum domain are used to model speech and music. The second system is a speech/non-speech segmentation technique [2] in which frame-by-frame maximum autocorrelation and log-energy features is measured, sorted and then followed by a linear discriminant analysis and a diagonalization transform.

As with other pattern recognition tasks, the first step in this audio classification is to extract and represent the sound by its relevant features so as to capture the discriminative properties of the sound, and to resist distortion under various noisy conditions. The novel aspect of our proposed system is a feature set inspired by investigations of various stages of the auditory system [3]. The features are computed from an auditory model that maps a given sound to a high-dimensional spectro-temporal modulations modeled after the auditory cortical representation. A key component of the approach is a multilinear dimensionality reduction method which by making use of multimodal characteristic of cortical representation, effectively removes redundancies in each subspace separately (section 3).

2. AUDITORY MODEL

The computational auditory model is based on neurophysiological, biophysical, and psychoacoustical investigations at various stages of the auditory system [3][4][5]. It consists of two basic stages. An early stage models the transformation of the acoustic signal into an internal neural representation (auditory spectrogram). A central stage analyzes the spectrogram to estimate the content of its spectral and temporal modulations (Figure 1)[4]. This stage is responsible for extracting the key features upon what the classification is based.

The early stages of auditory processing are modeled as a sequence of three operations [3]. The acoustic signal entering the ear produces a complex spatiotemporal pattern of vibrations along the basilar membrane of the cochlea. The basilar membrane outputs are then converted into inner hair cell intra–cellular potentials. This process is modeled as a 3–step operation: a highpass filter (the fluid–cilia coupling), followed by an instantaneous nonlinear compression (gated ionic channels), and then a lowpass filter (hair cell membrane leakage). Finally, a lateral inhibitory network detects discontinuities in the responses across the tonotopic axis of the auditory nerve array. Higher central auditory stages (especially the primary auditory cortex) further analyze the auditory spectrum into more elaborate representations, interpret them, and separate the different cues and features associated with different sound percepts. Specifically, from a conceptual point of view, these stages estimate the spectral and temporal modulation content of the auditory spectrogram. They do so computationally via a bank

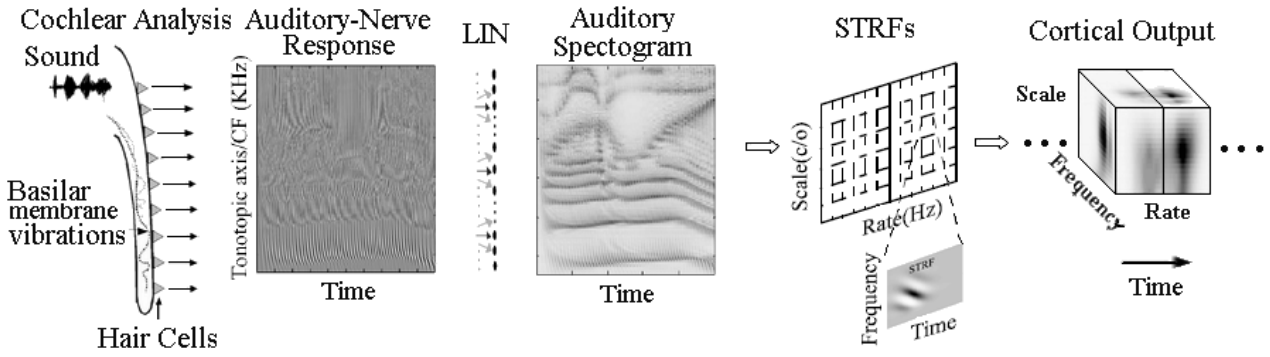

Fig. 1. Schematic of the auditory processing. Sound is analyzed by a model of the cochlea filter bank(left) , each filter output is then half–wave rectified and lowpass filtered by an inner hair cell model. A spatial first–difference operation is then applied mimicking the function of a lateral inhibitory network (LIN). The auditory spectrogram is then analyzed by a bank of spectro-temporal modulation selective filters. The total output as a function of time from the model is therefore indexed by three parameters: scale , rate, and frequency.

of modulation-selective filters centered at each frequency along the tonotopic axis. It has a spectro-temporal impulse response (usually called Spectro-Temporal Response Field – STRF) in the form of a spectro-temporal Gabor function which effectively results is a multi-resolution wavelet analysis of auditory spectrogram. All parameters of this model are derived from physiological data in animals and psychoacoustical data in human subjects [3][6] .

Unlike conventional features, our auditory-based features have multiple scales of time and spectral resolution. Some respond to fast changes while others are tuned to slower modulation patterns; A subset are selective to broadband spectra, and others are more narrowly tuned. For this study, temporal filters (Rate) from 1 to 32Hz and spectral filters (Scale) from 0.5 to 8.00 Cycle/Octave were used to represent the spectro-temporal modulations of sound.

3. MULTILINEAR ANALYSIS OF CORTICAL REPRESENTATION

The output of the auditory model is a multidimensional array. For our purpose here, the time dimension is averaged over a given time window which results in a three mode tensor for each time window with each elements representing the overall modulations at corresponding frequency, rate and scale (128(frequency channels) $\times$26 (rates) $\times$6 (scales). Traditional dimensionality reduction methods like principal component analysis (PCA) are inefficient for multi dimensional data because they treat all the elements of the feature space similarly without considering the varying degree of redundance and discrimination contribution of each mode. Instead, it is possible using multi-dimensional PCA to tailor the amount of reduction in each subspace independently of others based on the relative magnitude of corresponding singular values and discriminative contribution of each mode. It also results in reducing the amount of training samples and computational load significantly since each subspace is considered separately. To generalize the concept of PCA to multidimensional tensors, we consider a generalization of SVD to tensors (Higher Order SVD [7]) . Every $(I_1 \times I_2 \times ... \times I_N)$–tensor A can be written as the product

$$A = S \times_1 U^{(1)} \times_2 U^{(2)} ... \times_N U^{(N)} \quad (1)$$

in which $U^{(n)}$ is a unitary matrix containing left singular vectors of mode–n unfolding of tensor A, and S is a $(I_1 \times I_2 \times ... \times I_N)$ tensor which has the properties of all–orthogonality and ordering. Lathauwer et al. shows [7] that the left singular matrices of the different matrix unfolding of A correspond to unitary transformations that induce the HOSVD structure which in turn ensures that the HOSVD inherits all the classical space properties from the matrix SVD. HOSVD results in a new ordered orthogonal bases for representation of the data in subspaces spanned by each mode of the tensor.

The auditory model transforms a sound signal to its corresponding time-varying cortical representation. Using a comprehensive training set, a new multilinear and mutually orthogonal principal axes can be found that approximate the data in a low–dimensional space. The resulting data tensor D, obtained by stacking all training tensors is decomposed to its mode–n singular vectors:

$$D = S \times_1 U_{frequency} \times_2 U_{rate} \times_3 U_{scale} \times_4 U_{samples} \quad (2)$$

Each singular matrix is then truncated by setting a predetermined threshold. New sound samples are first transformed to their cortical representation, A, and are then projected onto these truncated orthonormal axes $U'_{freq.}$, U'_{rate}, U'_{scale}:

$$Z = A \times_1 U'^{T}_{freq.} \times_2 U'^{T}_{rate} \times_3 U'^{T}_{scale} \quad (3)$$

The resulting tensor Z whose dimension is equal to the total number of retained singular vectors in each mode (7 for

	Auditory Model	*Method one*	*Method two*
Speech	100%	99.3%	94.4%
Non-speech	100%	99.1%	93.05%

Table 1. Percentage of correct classification for time window of one second

	Auditory Model	*Method one*	*Method two*
Speech	100%	97.3%	93.7%
Non-speech	98.9%	98.2%	90.6%

Table 2. Percentage of correct classification for time window of half a second

frequency , 5 for rate and 3 for scale dimensions) is used for classification.

Classification was performed using a Support Vector Machine (SVM) [8]. Radial basis function (RBF) were used as SVM kernel and were adjusted so as to minimize the error mean square and error variance of training set.

4. EXPERIMENTAL RESULTS

4.1. Audio Database

An audio database was assembled from five publicly available corpora. Speech samples were taken from TIMIT Acoustic–Phonetic Continuous Speech Corpus. For training, 300 samples were selected from TIMIT's training subset. For test purpose, 150 different sentences spoken by 50 different speakers (25 male, 25 female) were selected from TIMIT's test subset. Sentences and speakers in training and test sets were different.

To make the non-speech class as comprehensive as possible, animal vocalization from BBC Sound Effects audio CD collection, Music samples from RWC Genre Database [9] and Environmental sounds from Noisex and Aurora databases were assembled together. The training set included 300 speech and 740 non-speech samples and the test set consisted of 150 speech and 450 non-speech samples. The length of each utterance in training and test is equal to the selected time window (e.g. one one–second sample per sound file). [1]

4.2. Comparison and results

To evaluate the robustness and the ability of system to generalize to unseen noisy conditions, we conducted a comparison with two state–of–the–art studies, one from generic–audio analysis community by Scheirer and Slaney [1] and one from automatic–speech–recognition community by Kingsbery et al. [2].

The first system derived thirteen features in time, frequency and cepstrum domain to form two models for speech and music. To eliminate performance differences due to the use of different classifiers, an SVM was used in all comparisons. Our implementation of the system was first evaluated on the original database and comparable, if not better, results were obtained with SVM compared to the original publication. A second system was tested that was based on an audio segmentation algorithm described in [2]. In the proposed technique, the degree of voicing and frame–level log–energy value were used as features. Several frames of these features were sorted in increasing order and concatenated, and was reduced to two dimensions by an LDA followed by MLLT. Our evaluation of the system suggested that direct classification of the original sorted vector with an SVM classifier outperformed the one in reduced dimension. For this reason, the classification was performed in the original feature space.

Our auditory model and the two algorithms from the literature were trained and tested on the same database. One of the important parameters in any such speech detection/discrimination task is the time window or duration of the signal to be classified. Table 1 and 2 shows the effect

[1]The list of files and offsets are available from the authors.

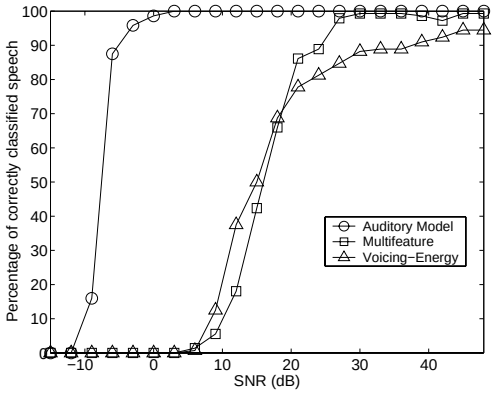

Fig. 2. Percentage of correct classified speech in white noise for auditory model, multi–feature [1] method and voicing–energy [2] method.

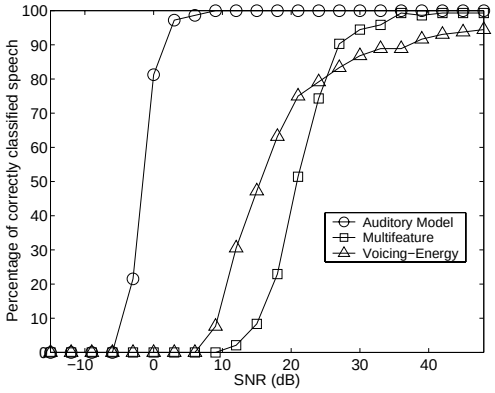

Fig. 3. Percentage of correct classified speech in pink noise for auditory model, multi–feature [1] method and voicing-energy [2] method.

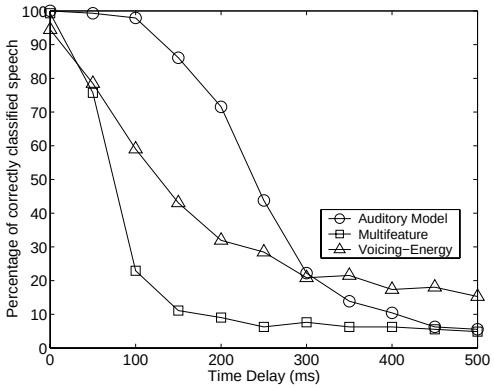

Fig. 4. Percentage of correct classified speech in reverberation for auditory model, multi–feature [1] method and voicing-energy [2] method.

of time window length on performance of the classifiers. Fixing the time window to one second, a series of tests were conducted to evaluate the generalization of the three methods to unseen noisy and reverberant speech. Classifiers were trained solely to discriminate clean speech from non-speech and then tested in three conditions. In each test, the percentage of correctly detected speech was considered as the measure of performance. For the first two tests, white and pink noise were added to speech with specified signal to noise ratio (SNR). The detection results for speech in white noise (Figure 2) demonstrate that while the three systems have comparable performance in clean conditions, the auditory features remain robust down to fairy low SNRs. This pattern is repeated with additive pink noise although performance degradation for all systems occurs at higher SNRs (Figure 3). To examine the effect of different levels of reverberation on the performance , a realistic reverberation condition was simulated by convolving the signal with a random gaussian noise with exponential decay. Figure 4 shows the performance of the systems in this condition.
On the whole, the data demonstrate the significant robustness of the auditory model.

5. SUMMARY AND CONCLUSIONS

An *spectro-temporal Auditory method* for audio classification and segmentation has been described, tested and compared to two state-of-the-art alternative approaches. The method employs features extracted by a biologically inspired auditory model of auditory processing in the cortex. The drawback of such a representation is its high dimensionality, and hence to utilize it, we developed and efficient multi-linear dimensionality reduction algorithm based on HOSVD of the multimodal data. The comparison with alternative systems demonstrate that the proposed system generalizes well to novel situations, an ability that is lacking in many of today's audio and speech recognition and classification systems.

This work is but one in a series of efforts at incorporating multi–scale cortical representations (and more broadly, perceptual insights) in a variety of audio and speech processing applications. For example, the deterioration of the spectro-temporal modulations of speech under any kind of linear or non–linear distortions, can be used as an indicator of predicted speech intelligibility [6]. Similarly, the multi–scale rate–scale–frequency representation can account for the perception of complex sounds and perceptual thresholds in a variety of settings [10], and finally, the auditory model can be readily adapted and expanded for a wide range of applications such as the automatic classification and segmentation of animal sounds [11], or an efficient encoding of speech and music [12].

6. ACKNOWLEDGMENT

The authors are grateful to Brian Zook of the Southwest Research Institute for critical contribution and support of this work. We would also like to thank Telluride Neuromorphic Engineering Workshop and Masataka Goto of the AIST for his help acquiring the RWC music samples. Partial funding for this project was also obtained from the National Science Foundation (ITR, 1150086075).

7. REFERENCES

[1] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", ICASSP'97, 1997.
[2] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system", ICASSP 2002, vol. I, pp. 53–56, 2002.
[3] R. Lyon, S. A. Shamma, "Auditory representation of timbre and pitch", Auditory computation, vol. 6, Springer–Verlag, New York, pp. 221–270, 1996
[4] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system", IEEE Trans. Speech Audio Proc. vol. 3 (5), pp. 382–395, 1995.
[5] X. Yang, K. Wang and S. A. Shamma, "Auditory representation of acoustic signals", IEEE Trans. Inf. Theory, 38 (2), pp. 824–839, (Spec. issue on wavelet trans.), 1992.
[6] M. Elhilali, T. Chi and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility", Speech comm., vol. 41, pp. 331–348, 2003.
[7] L. De Lathauwer, B. De Moor, J. Vandewalle, "A multilinear singular value decomposition", SIAM J. Matrix Anal. Appl., 21, pp. 1253–1278, 2000.
[8] T. Joachims, "Making large–Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning", B. Schlkopf, C. Burges and A. Smola (ed.), MIT–Press, 1999.
[9] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", ISMIR 2003, pp.229–230, 2003.
[10] R. P. Carlyon, S. A. Shamma, "An account of monaural phase sensitivity", J. of Acoust. Soc. Amer. vol. 114(1), pp. 333–48, 2003.
[11] N. Mesgarani and S. A. Shamma, "Bird call classification using a spectro-temporal multiresolution auditory model", International conference on acoustic communication by animals, College Park, MD, July 2003.
[12] L. Atlas, S. A. Shamma, "Joint acoustic and modulation frequency", Eurasip J. on App. signal proc., No. 7, pp. 668–675, June 2003.