Introduction to the Special Issue on Music Information Retrieval

MUSIC listening stands at an unprecedented point in history. Never before have so many listeners had access to such large collections of music. Large online music providers offer consumers millions of songs from a catalog of material extending over many decades and genres. No other signal processing technologies touch us as often or as personally as those related to music listening, whether in the guise of MP3 files, or recommendation systems, or new ways to explore music or enhance music playback.

However, this rich experience comes with a drawback. How do you choose what to listen to when you have instant access to millions of different songs? How can music librarians fulfill their roles given the explosion of content? And what can we do to keep track of new music, when technology makes it possible for almost anyone to produce professional-quality music, and many more types of musical activity have become viable thanks to markets enabled by the Internet?

The papers in this special issue were selected to represent the range of current problems in music information retrieval (MIR). These papers describe the state of the art in music signal analysis, from research that explores how to pull apart a musical signal into its constituents, to research aimed at finding specific kinds of matching items in large databases, to research into broader notions of musical similarity and retrieval including emotions and other semantic attributes. This is truly a wide range of research with substantial overlap between its various subdisciplines, but for the purpose of organization we have grouped the papers in this special issue into these three broad categories of signal analysis, retrieval, and classification; we discuss each type of work in turn.

Signal Analysis: At the lowest level, music audio presents a wide range of challenges in building machine systems that can recover the kind of music information (notes, instruments, phrases, lyrics, etc.) that form a listener's understanding of the music. Eight of the papers in this collection address these issues: Three papers look at analyzing music audio into the distinct note sequences arising from the different instruments, specifically considering the recovery of the respective pitches (Klapuri), the problems of organizing pitches according to instrument (Every), or the identification of a single, main melody (Lagrange *et al.*). A fourth paper by Lee and Slaney sidesteps the difficulties of identifying individual notes and instead analyzes the audio into a sequence of chords, i.e., the musical units built from simultaneous notes. Finally, a fifth paper by Doets and Lagendijk look for clues to the extent of compression in a musical signal based on audio fingerprinting.

At a broader level, three papers consider the segmentation of signals into the phrases or segments from which the often repetitive structure of music is built. Levy and Sandler describe a scheme for clustering similar segments, whereas Dubnov looks specifically at the notion of anticipation and its interaction with musical repetition. Kan *et al.* take a different approach by using the known lyrics of a musical piece including singing, then aligning the text to the audio based on various effective musical constraints.

Retrieval: As we have explained, the core problem in MIR is finding any particular item in the enormous archives that now exist, which can contain millions of items, each hundreds of seconds long. Here, six papers provide solutions for different aspects and formulations of this problem. One recurrent paradigm is popularly known as query-by-humming (QBH). Jang and Lee present a high-performance QBH system as an example of a particular approach to optimizing search time in such huge tasks. Unal *et al.* present a particular technique for the matching of queries that are likely to contain errors or ambiguities, and show its effectiveness in improving results for queries sung by users without musical training. Suyoto *et al.* present another technique for matching inexact queries that addresses the efficiency of looking for matching subranges.

In a more general task of matching music or other complex audio excerpts to one another, the challenge again is to relax the match in a way that manages to capture what a user considers similar, and at the same time preserve the efficiency necessary to search enormous collections. Kurth and Müller present methods for quantizing time segments according to their harmonic content, and for efficiently indexing such a representation to permit the retrieval of repeating phrases and alternative performances. Kimura *et al.* present a technique for efficiently finding approximate matches that is also applicable to signals other than music audio. Pampalk *et al.* look specifically at the significant subtask of drum sound retrieval (e.g., for drum loop collections) and develop a method for quantifying the similarity between members of this diverse family of sounds that resembles subjective similarity judgments.

Classification and Recommendation: At a higher level, listeners can describe entire pieces of music with a small number of terms that reflect internal categorizations based on similarity of musical style, or emotion, or other more obscure points of contact. Much of the work in MIR is oriented towards giving machines the ability to predict or understand these categories, with a view to automatically helping users choose what to listen to. Holzapfel and Stylianou look specifically at classifying music by genre, and propose a novel feature extraction approach with significant benefits. Yoshii *et al.* describe a complete music recommendation system that neatly integrates both similarity information derived from the collections of other listeners—widely known to be a very reliable basis—with purely audio-based measures in the case where the music occurs in too few collections (perhaps because it is a new release).

Digital Object Identifier 10.1109/TASL.2007.915221

Three papers look at connecting the statistics of features derived from audio with much more abstract, semantic attributes. Yang *et al.* make a careful regression between specific audio features and a database of music for which emotional labeling has been collected. Mion and De Poli look at describing broader aspects of musical expression. Finally, Turnbull *et al.* present techniques for building classifiers that associate audio with a wide range of words, based on specific labeling performed by human informants.

A special issue such as this one requires an extraordinary effort from the community. Planning for this special issue started in January 2006. The papers in this special issue were drawn from a pool of 40 submissions. More than 100 reviewers evaluated and contributed to the quality of these papers. We appreciate their efforts. Finally, both Mari Ostendorf and Kathy Jackson were immensely helpful in helping us understand the editorial process, getting these papers through the system, and assembling the final issue. Thank you to all.

We hope you learn from and enjoy each of these contributions.

MALCOLM SLANEY, *Guest Editor*
Yahoo! Research Labs and Stanford University CCRMA
Santa Clara, CA 95054

DANIEL P. W. ELLIS, *Guest Editor*
Columbia University
New York, NY 10027-6902

MARK SANDLER, *Guest Editor*
Queen Mary, University of London
London, E1 4NS U.K.

MASATAKA GOTO, *Guest Editor*
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, 305-8568 Japan

MICHAEL M. GOODWIN, *Guest Editor*
Creative Advanced Technology Center
Scotts Valley, CA 95066

Malcolm Slaney (SM'01) received the Ph.D. degree from Purdue University, West Lafayette, IN, for his work on diffraction tomography.

Since the start of his career, he has been a Researcher at Bell Labs, Schlumberger Palo Alto Research, Apple's Advanced Technology Lab, Interval Research, IBM Almaden Research Center, and most recently at Yahoo! Research, Santa Clara, CA. Since 1990, he has organized the Stanford CCRMA Hearing Seminar, where he now holds the title (Consulting) Professor. He is a coauthor (with A. C. Kak) of the book *Principles of Computerized Tomographic Imaging*, which has been republished as a Classics in Applied Mathematics by SIAM Press. He is a coeditor of the book *Computational Models of Hearing* (IOS Press, 2001).

Daniel P. W. Ellis (SM'04) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996.

He was a Research Assistant at the Media Lab, MIT. He is an Associate Professor in the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He is an External Fellow of the International Computer Science Institute, Berkeley, CA. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound.

Mark Sandler (SM'98) was born in 1955. He received the B.Sc. and Ph.D. degrees from the University of Essex, Essex, U.K., in 1978 and 1984, respectively.

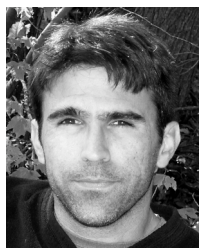
He is a Professor of Signal Processing at Queen Mary, University of London, London, U.K., and Director of the Center for Digital Music. He has published over 300 papers in journals and conferences.

Dr. Sandler is a Fellow of the Institute of Electronic Engineers (IEE) and a Fellow of the Audio Engineering Society. He is a two-time recipient of the IEE A. H. Reeves Premium Prize.

Masataka Goto received the Doctor of Engineering degree from Waseda University, Tokyo, Japan, in 1998.

He then joined the Electrotechnical Laboratory (ETL), which was reorganized as the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan, in 2001, where he has been a Senior Research Scientist since 2005. He served concurrently as a Researcher in Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST), from 2000 to 2003, and has been an Associate Professor in the Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba, since 2005.

Dr. Goto received 21 awards, including the IPSJ Best Paper Award, IPSJ Yamashita SIG Research Awards, and Interaction 2003 Best Paper Award. He is a member of the IPSJ, ASJ, JSMPC, IEICE, and ISCA.



Michael M. Goodwin (M'98) received the B.S. and M.S. degrees in electrical engineering and computer science (EECS) from the Massachusetts Institute of Technology, Cambridge, in 1992 and the Ph.D. degree in EECS from the University of California, Berkeley, in 1997.

Since 1999, he has been with the Audio Research Department at the Creative Advanced Technology Center, Scotts Valley, CA. His current research interests include signal modeling and enhancement, audio coding, spatial audio, and array processing. He is the author of a book entitled *Adaptive Signal Models: Theory, Algorithms, and Audio Applications* (Kluwer, 1998).

Dr. Goodwin served as chair of the IEEE Signal Processing Society's Technical Committee on Audio and Electroacoustics from 2005 to 2007. He is also a member of the Audio Engineering Society (AES) and the AES technical committees on signal processing and audio coding.