Introduction to the Special Issue on Music Signal Processing

MUSIC is enjoyed by billions of people worldwide, and today many listeners enjoy ubiquitous access to practically unlimited music collections due to the proliferation of portable music players. The field of music signal processing, while young in comparison with more established areas such as speech processing, has been steadily growing in past decade. It now encompasses a wide range of topics in computer-based music analysis, processing, and retrieval. In earlier decades, music research using computers relied primarily on symbolic representations such as musical notation or MIDI, but thanks to the increased availability of digitized audio material and the explosion of available computing power, the focus of research efforts is now the processing and analysis of music audio signals, which vastly increases the amount and variety of relevant material.

This special issue is devoted to this emerging field of music signal processing, which has been growing both within and beyond the traditional signal processing community. Our goals in producing this issue are two-fold: First, we want to spur progress in the core techniques needed for the future signal processing systems that will enable users to access and explore music in all its different aspects. Our second goal is to introduce this vibrant and exciting new field to a wider signal processing readership.

The problem of extracting meaningful information from audio waveforms is well suited to the techniques of digital signal processing, but when dealing with specific audio domains such as speech or music, it is crucial to properly understand and apply the appropriate domain-specific properties, be they acoustic, linguistic, or musical. In this special issue, we have gathered contributions that take into account key properties of music signals, such as the presence of multiple, coordinated sources, the existence of structure at multiple temporal levels, and the peculiar kinds of information being carried.

The first paper is written by the guest editors and provides an overview of some signal analysis techniques that specifically address musical dimensions such as melody, harmony, rhythm, and timbre. We examine how particular characteristics of music signals impact and determine these techniques, and we highlight a number of novel music analysis and retrieval tasks that such processing makes possible.

Most musical instruments are explicitly constructed to allow performers to produce sounds with easily controlled, locally stable fundamental periods. Such a signal is well described as a series of frequency components at multiples of a fundamental frequency, and results in the percept of a musical note at a clearly defined pitch in the mind of the listener. The first set of papers of this special issue is related to *pitch analysis* of polyphonic music recordings and aims at estimating the fundamental frequencies (F0s) of several concurrent sounds. Multiple-F0 estimation is a major step towards decomposing a complex signal into its constituent sound sources, and as such forms the basis for a number of music analysis applications. Benetos and Dixon employ joint estimation of multiple F0s and a novel spectral envelope estimation procedure to perform polyphonic music transcription. In the contribution by Wu *et al.*, pitch estimation is jointly performed with instrument identification by considering the sustain as well as the attack portion of each note's sound within a single model. Peeling and Godsill model harmonically related spectral peaks in a music signal with a non-uniform Poisson process and use this to improve on polyphonic pitch estimation.

Playing even a single musical note with a clearly defined pitch on a real instrument usually produces a complex mixture spectrum with multiple partials—in addition to percussive and transient components. The energy distribution among the partials is closely related to the *timbre* of the instrument. As a consequence, determining multiple pitches from a sound mixture also requires some sort of modeling the timbre of the involved instruments, and vice versa. Carabias *et al.* explicitly represent an instrument's timbral characteristics using a source-filter model, where parameters are tied across different pitches. Grindlay and Ellis go a different way by learning a probabilistic model they refer to as "Hierarchical Eigeninstruments" from audio examples. Both contributions are evaluated in the context of a music transcription application. A different perspective on the musical aspect of timbre is given by Chen and Huang in the context of sound synthesis, where physical modeling techniques are applied to simulate the sound of a pipa, a traditional Chinese plucked string instrument.

As with many real-world audio scenarios, the task of extracting the individual sources from a mixed signal—*source separation*—is of central importance in music processing. In the musical context, the individual sources often correspond to individual instruments including singing voice or drums, and relate to musical voices such as the bass line or the melody. What makes this task difficult is that musical sources are far from being independent. Actually, quite the opposite is true: musicians perform together, interact with each other, and play consonant notes contributing to a single musical impression. As a consequence, the sources are typically highly correlated, share many of their harmonics, and follow the same rhythmic patterns. To deal with such complex scenarios, one has to exploit strong prior information on the sources or voices. Durrieu *et al.* introduce a mid-level representation for discriminating the dominant pitched source from other sources and apply their methods to lead instrument separation. Kim *et al.* address the problem of separating drums from single-channel polyphonic music utilizing either prior knowledge of the drum sources under analysis or the repetitive nature of drum sources in the target signals. Duan and Pardo exploit even stronger cues by employing a score-informed source separation strategy, where a musical score is used as additional input to guide the separation process.

Digital Object Identifier 10.1109/JSTSP.2011.2165109

Finally, Jo *et al.* tackle the problem of extracting the melody line from a polyphonic audio mixture by exploiting the fact that the melody typically corresponds to the most salient pitch sequence.

The final set of papers addresses various music analysis and feature extraction tasks, where further musical aspects related to *harmony*, *rhythm*, and the presence of *lyrics* play an important role. Degara *et al.* show that rhythmic information related to tempo and beat can be exploited to improve the extraction of note onsets from music recordings. Weiss and Bello address the problem of finding recurrent harmonic patterns in audio recordings by employing a chroma-based representation that correlates with the harmonic progression of music. In the contribution by Hiromasa *et al.*, the authors describe a system that automatically aligns lyrics to popular music, employing phoneme models to temporally align available lyrics to a corresponding recording of the song. Finally, the contribution by Schuller *et al.* addresses the higher-level problem of extracting audio features directly from compressed audio material, thus significantly reducing the computational complexity.

This special issue is the result of the work by many people over the last two years. First of all, we thank the authors for their contributions and their efforts in the revision process. The final papers were drawn from a pool of more than 40 submissions. We highly appreciate and wish to thank the reviewers for their hard work and valuable feedback, which has contributed significantly to the quality of this issue. Last but not least, both Vikram Krishnamurthy, the Editor in Chief, and Rebecca Wollman, the responsible IEEE Publications Coordinator, were extremely helpful with their advice and active support in the review and editing process. Thank you very much.

MEINARD MÜLLER, *Lead Guest Editor*
Saarland University and MPI Informatik
66123 Saarbrücken, Germany

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

ANSSI KLAPURI, *Guest Editor*
Queen Mary University of London
London E1 4NS, U.K.

GAËL RICHARD, *Guest Editor*
Institut Télécom, Télécom ParisTech, CNRS-LTCI
75014 Paris, France

SHIGEKI SAGAYAMA, *Guest Editor*
The University of Tokyo
Tokyo 113-8656, Japan

Meinard Müller received the Diplom degree in mathematics, the Ph.D. degree in computer science, and the Habilitation degree in the field of multimedia retrieval from Bonn University, Bonn, Germany.

In 2002/2003, he conducted postdoctoral research in combinatorics in the Mathematical Department, Keio University, Tokyo, Japan. In 2007, he finished his Habilitation at Bonn University in the field of multimedia retrieval writing a book titled *Information Retrieval for Music and Motion*, which appeared as Springer monograph. Currently, he is a member of the Saarland University and the Max-Planck Institut für Informatik, Saarbrücken, Germany, where he leads the research group *Multimedia Information Retrieval and Music Processing* within the Cluster of Excellence on *Multimodal Computing and Interaction*. His recent research interests include content-based multimedia retrieval, audio signal processing, music processing, music information retrieval, and motion processing.

Dr. Müller is a member of the IEEE AASP technical committee as well as a member of the Board of Directors of the International Society of Music Information Retrieval (ISMIR).

Daniel P. W. Ellis (M'96–SM'04) received the Ph.D. degree from the Massachusetts Institute of Technology (MIT), Cambridge.

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 all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound. He worked at MIT, where he was a Research Assistant at the Media Lab, and he spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA, where he remains an external fellow.

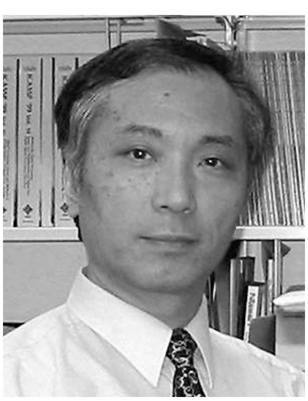
Anssi Klapuri received the Ph.D. degree from Tampere University of Technology (TUT), Tampere, Finland, in 2004.

He visited as a Post-Doctoral Researcher at Ecole Centrale de Lille, France, and Cambridge University, U.K., in 2005 and 2006, respectively. He worked until 2009 as a Professor at TUT. In 2009, he joined Queen Mary, University of London, as a Lecturer in sound and music processing. His research interests include audio signal processing, auditory modeling, and machine learning.

Gaël Richard (M'02–SM'06) received the State Engineering degree from TELECOM ParisTech (formerly ENST), Paris, France, in 1990, the Ph.D. degree in speech synthesis from LIMSI-CNRS, University of Paris-XI, in 1994, and the Habilitation à Diriger des Recherches degree from the University of Paris XI in September 2001.

After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches to speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications, Bois d'Arcy, France, and for Philips Consumer Communications, Montrouge, France. In particular, he was the Project Manager of several large-scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, TELECOM ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves Research Group. He is coauthor of over 80 papers, inventor of a number of patents, and one of the experts of the European commission in the field of speech and audio signal processing

Prof. Richard is a member of EURASIP and an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.



Shigeki Sagayama (M'82) was born in Hyogo, Japan, in 1948. He received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998, respectively, all in mathematical engineering and information physics.

He joined Nippon Telegraph and Telephone Public Corporation (currently, NTT) in 1974 and started his career in speech analysis, synthesis, and recognition at NTT Laboratories, Musashino, Japan. From 1990 to 1993, he was Head of Speech Processing Department, ATR Interpreting Telephony Laboratories, Kyoto, Japan, pursuing an automatic speech translation project. From 1993 to 1998, he was responsible for speech recognition, synthesis, and dialog systems at NTT Human Interface Laboratories, Yokosuka, Japan. In 1998, he became a Professor of the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan. In 2000, he was appointed Professor of the Graduate School of Information Science and Technology (formerly Graduate School of Engineering), University of Tokyo. His major research interests include processing and recognition of speech, music, acoustic signals, handwriting, and images. He was the leader of anthropomorphic spoken dialog agent project (Galatea Project) from 2000 to 2003.

Prof. Sagayama is a member of the Acoustical Society of Japan (ASJ), IE-ICEJ, and IPSJ. He received the National Invention Award from the Institute of Invention of Japan in 1991, the Chief Official's award for Research Achievement from the Science and Technology Agency of Japan in 1996, and other academic awards including Paper Awards from the Institute of Electronics, Information, and Communications Engineers (IEICEJ) Japan, in 1996 and from the Information Processing Society of Japan (IPSJ) in 1995.