Meinard Müller, Bryan A. Pardo, Gautham J. Mysore,
and Vesa Välimäki

Recent Advances in Music Signal Processing

Music is a ubiquitous and vital part of our lives. Thanks to the digital revolution in music distribution and storage, music has become one of the most popular categories of multimedia content. In general terms, music processing research contributes concepts, models, and algorithms that extend our capabilities of accessing, analyzing, manipulating, and creating music. Given the complexity and diversity of music, researchers must account for various aspects, such as the genre, instrumentation, dynamics, tempo, rhythm, and timbre. Music signals typically comprise a wide range and large number of different sound sources. Postprocessing and the use of audio effects in the mixing and mastering stages may further complicate the analysis of recorded musical material. Furthermore, music is inherently multimodal, incorporating speech-like signals (singing), video (of live performances), and still images (scanned music scores). This wealth of data makes music processing a challenging field of research and closely connected to areas such as audio and acoustic signal processing, multimedia signal processing, and machine learning.

Compared with speech processing, a research field with a long tradition, music processing is still a relatively young discipline, but it is rapidly growing. In recent decades, the music processing community has come together by organizing major annual conferences on topics including music information retrieval, sound and music computing, audio-effects processing, computer music, and applications in audio engineering. Although computer-based music research has traditionally been conducted using symbolic representations, the research focus has shifted to other types of music-related data including audio recordings, digitized images, music videos, and other types of sensor data. As a consequence, digital signal processing has found its way into many research communities dealing with music-related data.

In this special issue of *IEEE Signal Processing Magazine (SPM)*, we survey recent advances in music processing with a focus on audio signals. Eleven articles cover topics including music analysis, retrieval, source separation, singing-voice processing, musical sound synthesis, and user interfaces, to name a few. The tutorial-style articles provide an overview of theory and applications and discuss main advances. Although music processing has benefited a lot from traditional fields, such as signal processing, we hope to convince the reader that the rich and challenging problem domain of music also has many things to offer to signal processing and other research disciplines.

We start the special issue with the classic problem of music transcription, where the objective is to extract note events, key signature, time signature, instrumentation, and other score parameters from a given music recording. "Automatic Music Transcription" by Benetos et al. gives an overview of computational algorithms for converting a music signal to written music notation, with an emphasis on approaches for transcribing polyphonic music produced by pitched instruments and voice. The article details the methodology used in the two main families of approaches: those based on deep learning and those based on nonnegative matrix factorization (NMF). It also links automatic music transcription to other problems found in the broader field of digital signal processing, including multiple-F0 estimation, instrument recognition, and source separation.

Music signals contain complex mixtures of different yet highly correlated sound sources, which creates a challenge for processing. For example, a singer, a guitarist, a keyboard player, and a drummer may be active at the same time, following the same rhythmic pattern, and playing notes that are harmonically related. In "Musical Source Separation," Cano et al. review the problem of recovering the individual tracks as if they had been played in isolation. The article discusses sound characteristics of music signals and how these characteristics can be exploited to develop appropriate source-separation algorithms. Furthermore, it covers various key techniques including kernel additive models, NMF, sinusoidal models, and deep neural networks. Although music source separation has significantly progressed over the last decade,

Digital Object Identifier 10.1109/MSP.2018.2876190
Date of publication: 24 December 2018

there are still numerous open problems, including the quality assessment of separated sources.

In the last ten years, music streaming has become the predominant way of accessing and consuming music. Along with providing content, music streaming services also give personalized recommendations based on play history or collaborative filtering. In this context, song descriptors, such as genre, mood, instrumentation, and vocal quality, are needed. Obtaining these descriptors by manual annotation is a costly and time-consuming process that does not scale to huge music collections. Therefore, computational approaches for music content understanding—including music genre classification, music mood classification, and music autotagging—have become a major strand of research. The article "Deep Learning for Audio-Based Music Classification and Tagging" by Nam et al. provides an up-to-date survey of the deep network designs tailored for music classification and tagging tasks. It covers best practices and applications to music services and discusses the limitations of current approaches and open issues in this area of research.

Compared with speech processing, a research field with a long tradition, music processing is still a relatively young discipline, but it is rapidly growing.

The rapid growth of digitally available music data goes beyond audio recordings and includes other modalities, such as digitized images of sheet music, album covers, liner notes, and video clips. The following three articles are all concerned with exploiting, linking, and jointly analyzing this wealth of data. "Cross-Modal Music Retrieval and Applications" by Müller et al. reviews several cross-modal retrieval scenarios, with a particular focus on sheet music (visual domain) and audio (acoustic domain). Given a query in one modality, the task is to retrieve semantically corresponding documents in some other modality. By bridging the gap between various music representations, this technology enables music navigation and browsing applications, including the classic problem of automated score following. Besides traditional approaches based on musically motivated features, this article also discusses generalized audio fingerprinting and recent data embedding techniques based on deep learning.

The article "Audiovisual Analysis of Music Performances" by Duan et al. shows the growing significance of jointly analyzing audio and visual data for music, with a specific focus on music performance. This cross-modal approach is of particular relevance for tasks, such as tracking a musician's fingering or a conductor's gestures, for which the video provides information that is complementary to the audio signal. The article provides an overview of analyzing performances in which the audio and video have both static correspondences (such as musicians in an orchestra whose relative positions do not change) and dynamic correspondences (such as vibrato analysis).

Goto and Dannenberg's article, "Music Interfaces Based on Automatic Music Signal Analysis," shows how music processing techniques open up new possibilities for developing interactive music systems and Web services. Such interfaces include audiovisual elements based on structural segments, beats, melody line, and chords to simplify audio navigation or enrich listening experience. As for music production, intelligent audio editors allow users to identify and rearrange drums patterns and other score-based sound events.

The human voice plays an integral part in nearly all music cultures. Therefore, it is not surprising that the analysis, synthesis, and classification of music signals that involve the human voice is at the core of music signal processing. Singing not only differs fundamentally from spoken language but often occurs in a polyphonic context and in combination with other instruments. Thus, it requires computational approaches that are different from those used in speech processing. The article "An Introduction to Signal Processing for Singing-Voice Analysis" by Humphrey et al. covers fundamentals of the human voice and vocalization techniques. Based on time–frequency patterns specific to singing, various music processing tasks, such as singer activity detection, melody estimation, genre classification, and intonation estimation, are covered. Furthermore, the article highlights the unique difficulties that are faced when dealing with music, where tasks such as language identification, audio–lyrics alignment, and lyrics transcription (tasks that are also well known in speech processing) become extremely challenging.

While the article by Humphrey et al. treats singing-voice processing from an analysis perspective, the contribution "Speech-to-Singing Voice Conversion" by Vijayan et al. approaches this topic from a synthesis perspective. It covers the problem of converting a speaking voice into a singing one while preserving the linguistic content and the speaker's vocal identity. While discussing template- and model-based key techniques for singing synthesis, the article highlights the differences between singing and speaking in terms of dynamic range, pitch variations, and duration of the linguistic content.

Continuing with music synthesis, the article "Model-Based Digital Pianos" by Bank and Chabassier addresses the task of developing digital sounds that mimic the sound of an acoustic piano. The authors review the physics of the piano and then introduce a comprehensive physical model considering hammer motions, string vibrations, soundboard, and sound radiation. Such models can become prohibitively expensive to compute. Therefore, when developing real-time sound-synthesis applications, one requires model simplifications that do not sacrifice perceptual quality. The article reviews optimization approaches that neglect inaudible phenomena and enhance the physical models with perceptually appropriate modifications.

Nogueira et al. shine another light on music signals and their properties in their article "Making Music More Accessible for Cochlear Implant Listeners" by adopting the perspective of

hearing-impaired listeners. Although cochlear implants (CIs) enable users to understand continuously spoken speech to a high degree, music features, such as pitch and timbre, are still poorly transmitted. The contribution discusses various signal processing methods (e.g., the amplification of a song's melody or the simplification of the music content) that make music more accessible and enjoyable for CI users. Finally, the authors address the problem of objective and subjective evaluation measures used to assess the quality of the transmitted music signal.

In applied research areas, such as music signal processing, theoretical concepts need to be implemented and evaluated using real-world data in concrete application scenarios. Implementation details can have a substantial impact on the overall performance of a given system. In practice, it is often nearly impossible to specify all relevant factors of an experiment, including design choices in signal processing and machine-learning approaches, the exact composition of the data collection used, and implicit assumptions in the evaluation process. In "Open-Source Practices for Music Signal Processing Research," McFee et al. give recommendations on best practices for open-source software development in the context of music information retrieval applications. Furthermore, they lay out future directions for incorporating open-source and open-science methodology—a topic that pervades all data-driven areas of signal processing.

Many people contributed to compiling this special issue, which is the result of a team effort throughout the whole research community. First, we thank *SPM*'s Special Issues Area Editor Douglas O'Shaughnessy, IEEE Signal Processing Society Publications Administrator Rebecca Wollman, and Managing Editor Jessica Welsh for helping with the organization and production of this issue. Furthermore, we thank the many reviewers for their detailed and constructive comments throughout several rounds of revision. Last but not least, we thank the authors for writing tutorial-style articles and finding a way to be comprehensive and compact, instructive, and entertaining, while covering basic principles and state-of-the-art techniques.

Music signal processing is an exciting and challenging area of research. Music not only connects people but also relates to many different research disciplines. This interdisciplinary perspective is becoming commonplace in music technology research labs and enabling novel algorithm development. We hope that this special issue provides examples of such work and inspires new ideas that cross boundaries of the IEEE Signal Processing Society and other fields.

Meet the guest editors

Meinard Müller (meinard.mueller@audiolabs-erlangen.de) received his diploma degree in mathematics in 1997 and his Ph.D. degree in computer science in 2001, both from the University of Bonn, Germany. Since 2012, he has held a professorship for semantic audio signal processing at the International Audio Laboratories Erlangen, Germany. His recent research interests include music processing, music information retrieval, and audio signal processing. He has coauthored more than 100 peer-reviewed scientific articles and has written a monograph, *Information Retrieval for Music and Motion* (Springer, 2007), and a textbook, *Fundamentals of Music Processing* (Springer, 2015).

Bryan A. Pardo (pardo@northwestern.edu) received his M.Mus. degree in jazz studies in 2001 and his Ph.D. degree in computer science in 2005, both from the University of Michigan, Ann Arbor. He is an associate professor of computer science and music at Northwestern University, Evanston, Illinois. He has authored more than 100 peer-reviewed publications in the areas of machine learning, signal processing, and music information retrieval. When he is not programming, writing, or teaching, he performs throughout North America on saxophone and clarinet.

Gautham J. Mysore (gmysore@adobe.com) received his M.A. degree in music, science, and technology in 2005, his M.S. degree in electrical engineering in 2008, and his Ph.D. degree in computer-based music theory and acoustics in 2010, all from Stanford University, California. He was previously a visiting researcher at the Gatsby Computational Neuroscience Unit at University College London, United Kingdom. He is a principal scientist and head of the Audio Research Group at Adobe Research in San Francisco, California, and an adjunct professor in the Center for Computer Research in Music and Acoustics at Stanford University. His research involves developing new machine-learning and signal processing algorithms for a wide variety of real-world audio applications.

Vesa Välimäki (vesa.valimaki@aalto.fi) received his M.Sc. degree in 1992 and his doctor of science in technology degree in 1995, both in electrical engineering from the Helsinki University of Technology, Espoo, Finland. In 2008–2009, he was a visiting scholar at Stanford University, California. He was the chair of the 2017 International Conference on Sound and Music Computing. He is a professor of audio signal processing and vice dean for research in electrical engineering at the Aalto University, Espoo, Finland. His research interests include headphone and loudspeaker signal processing, audio-effects processing, and sound synthesis. He is a fellow of the IEEE Audio Engineering Society.

SP