Editorial

Scalable Audio-Content Analysis

Bhiksha Raj,[1] Paris Smaragdis,[2] Malcolm Slaney,[3, 4] Chung-Hsien Wu,[5] Liming Chen,[6] and Hyoung-Gook Kim[7]

[1] *Carnegie Mellon University, PA 15213, USA*

[2] *Advanced Technology Laboratories, Adobe Systems Inc., Newton, MA 02466, USA*

[3] *Yahoo! Research, Santa Clara, CA 95054, USA*

[4] *Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, CA 94305-8180, USA*

[5] *Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan*

[6] *Department of Mathematics and Informatics, Ecole Centrale de Lyon, University of Lyon, 69006 Lyon, France*

[7] *Intelligent Multimedia Signal Processing Laboratory, Kwangwoon University, Seoul 139-701, Republic of Korea*

Correspondence should be addressed to Bhiksha Raj, bhiksha@cs.cmu.edu

The rapid increase in the amount of easily accessible audio, in the form of streaming audio content, recordings on social media sites such as Facebook and Youtube, public and personal song collections, and so on has raised new technical challenges. In order to make effective use of these recordings, we require smart techniques for storage and organization of these data, as well as for analyzing and retrieving them based on their content. Moreover, these techniques must be *scalable*, in order to deal with the volume of data.

The six papers in this special issue address some of these topics.

The first problem to be addressed in dealing with large volumes of audio data is that of storage. Ideally, we must compress the data such that they require fewer bits to store while not compromising audio quality. Current coding schemes provide a variety of tradeoffs between compression, audio quality, and latency. P. Motlicek et al. contribute their investigations into this area in their paper titled "*Wide-band audio coding based on frequency-domain linear prediction.*" They take advantage of the fact that latency is not a constraint for storage and propose an audio coding scheme that is based on linear prediction of the spectra of fairly long segments of the audio. They achieve compression rates comparable to MPEG4, while yet retaining the perceptual quality of the audio.

The papers by N. Misdariis et al., B. Schuller et al., and X. Ma et al. investigate content-based *description* of various types of audio data.

In their paper titled "*Environmental sound perception: metadescription and modeling based on independent primary studies*" N. Misdariis et al. apply methodologies usually used to study *timbre* of music to analyze various car sounds, with the goal of finding descriptors (obtained by application of multidimensional scaling) that might be useful for content-based indexing and retrieval of such sounds.

B. Schuller et al. study ways of modeling the *mood* of musical recordings using a discretized emotional model. They propose to determine nonprototypical valence and arousal in popular music, using features derived both from the acoustics of the recordings and, where available, song lyrics. Another major contribution of this work is the constitution of a dataset of annotated music of significant size, having more than 2000 titles and covering different representative genres. The annotations are made available to the research community.

X. Ma et al. explore semantic labeling of generic audio content in their paper titled "*Semantic labeling of nonspeech clips.*" They obtain semantic annotations of a large corpus of audio recordings by analyzing their descriptions by human subjects. In the process they also determine, perhaps not surprisingly, that descriptions by subjects are more likely to agree at coarse levels than at fine levels.

The papers by M. Rouvier et al., M. Helén, and T. Virtanen deal with *retrieval* of stored data.

In audio recordings containing speech, it is useful, or even important, to detect key words and phrases that could

be used to index or retrieve the recordings or tag them for further analysis. In large and continuously expanding corpora, this must be done fast, yet effectively. In their paper titled "*Query-driven strategy for on-the-fly term spotting in spontaneous speech*," Rouvier et al. propose a fast two-level architecture for detecting key words in spontaneous speech recordings. The first level performs a fast detection of speech segments that are likely to contain the desired terms. The second level refines the detection further using a speech recognizer and a query-driven decoding algorithm.

In their paper, titled "*Audio query by example using similarity measures between probability density functions of features*," M. Helén and T. Virtanen address an alternate problem: retrieval of *generic* (i.e., not necessarily speech-containing) audio. In particular, they consider the problem of *query by example*—retrieving other instances of audio that are similar to a given example. They investigate a number of different approaches and find that similarity measures based on distances between probability distribution functions computed from audio recordings result in the best retrieval.

No single issue of any journal can reasonably expect to cover even a small fraction of the problem space we address, and we do not strive to do so in this issue. Rather, it is our hope to provide a selection of good-quality papers that touch upon various aspects of the problem, that are both informative and enjoyable to read, and that present novel approaches or provide new insights that might be of use to the research community. We believe that the selection we have provided reflects these goals, and we hope you agree.

Bhiksha Raj
Paris Smaragdis
Malcolm Slaney
Chung-Hsien Wu
Liming Chen
Hyoung-Gook Kim