Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 12687, 2 pages
doi:10.1155/2007/12687

Editorial

Perceptual Models for Speech, Audio, and Music Processing

Jont B. Allen,[1] Wai-Yip Geoffrey Chan,[2] and Stephen Voran[3]

[1] *Beckman Institute, University of Illinois, 405 North Mathews Avenue, Urbana, IL 61801, USA*
[2] *Electrical and Computer Engineering Department, Queen's University, 99 University Avenue, Kingston, ON, Canada K7L 3N6*
[3] *Institute for Telecommunication Sciences, 325 Broadway, Boulder, CO 80305, USA*

Received 22 November 2007; Accepted 22 November 2007

New understandings of human auditory perception have recently contributed to advances in numerous areas related to audio, speech, and music processing. These include coding, speech and speaker recognition, synthesis, signal separation, signal enhancement, automatic content identification and retrieval, and quality estimation. Researchers continue to seek more detailed, accurate, and robust characterizations of human auditory perception, from the periphery to the auditory cortex, and in some cases whole brain inventories.

This special issue on Perceptual Models for Speech, Audio, and Music Processing contains seven papers that exemplify the breadth and depth of current work in perceptual modeling and its applications.

The issue opens with "Practical gammatone-like filters for auditory processing" by A. G. Katsiamis et al.which contains a nice review on how to make cochlear-like filters using classical signal processing methods. As described in the paper, the human cochlea is nonlinear. The nonlinearity in the cochlea is believed to control for dynamic range issues, perhaps due to the small dynamic range of neurons. Having a time domain version of the cochlea with a built in nonlinearity is an important tool in many signal processing applications. This paper shows one way this might be accomplished using a cascade of second-order sections. While we do not know how the human cochlea accomplishes this task of nonlinear filtering, the technique described here is one reasonable method for solving this very difficult problem.

B. Raj et al.apply perceptual modeling to the automatic speech recognition problem in "An FFT-based companding front end for noise-robust automatic speech recognition." These authors describe efficient FFT-based processing that mimics two-tone suppression, which is a key attribute of simultaneous masking. This processing involves a bank of relatively wide filters, followed by a compressive nonlinearity, then relatively narrow filters, and finally an expansion stage. The net result is that strong spectral components tend to reduce the level of weaker neighboring spectral components, and this is a form of spectral peak enhancement. The authors apply this work as a preprocessor for a mel-cepstrum HMM-based automatic speech recognition algorithm and they demonstrate improved performance for a variety of low-SNR background noise conditions.

"Wideband speech recovery using psychoacoustic criteria" describes how a perceptual loudness criterion can be used advantageously in wideband speech coding. Authors V. Berisha and A. Spanias propose enhancing a narrowband speech coder by sending a few samples of the high band (4–8 kHz) spectral envelope, and these samples are selected according to a loudness criterion. They apply this perception-based technique to the standardized narrowband adaptive multirate (AMR-NB) speech coder and evaluate the results through subjective testing. One test compares this bandwidth extended AMR-NB speech (total bitrate 9.1 kbps) to conventional AMR-NB speech (total bitrate of 10.2 kbps). In spite of the lower total bit-rate, listeners show a clear preference for the bandwidth extended speech.

Next is "Denoising in the domain of spectrotemporal modulations" where N. Mesgarani and S. Shamma examine the effectiveness of denoising speech signals using a spectrotemporal modulation decomposition proposed earlier by Chi, Ru, and Shamma. The decomposition is performed over two stages. First, the "early auditory system" maps the input speech signal to an auditory spectrogram. Then, the "central auditory system" decomposes the spectrogram into spectral and temporal modulations. N. Mesgarani and S. Shamma demonstrate that speech and different types of noise are well separated in the spectrotemporal modulation domain. Their denoising experiment, based on Wiener filtering in the modulation domain, shows their scheme to provide distinctively better speech quality than a conventional Wiener filtering scheme when the noise is stationary.

In “Perceptual continuity and naturalness of expressive strength in singing voice based on speech morphing,” T. Yonezawa et al.address the synthesis of expression in a singing voice with a specific focus on creating natural, continuous transitions between expressive strengths. They employ a speech morphing algorithm and subjective tests to identify a nonlinear morphing pattern that results in a nearly linear progression of perceived strength of expression. In additional subjective testing the authors verify that this perceived linear progression does indeed equate to a natural sound.

Next comes a very unusual article titled “Electrophysiological study of algorithmically processed metric/rhythmic variations in language and music.” Here S. Ystad et al. use event-related potentials (ERPs), which are small voltages recorded from the skin of the scalp, to study questions of meter, rhythm, semantics, and harmony in language and music. The key potential is called N400, which is known to relate to speech perception. They find that “language ERP analyses indicate that semantically incongruous words are processed independently of the subject’s attention.” This argues for automatic semantic processing. For the case of music they find that their ERP analyses show that “rhythmic incongruities are processed independently of attention.” Again, this argues for an “automatic processing of rhythm.”

Finally, A. Ikeno and J. H. L. Hansen consider a different form of perception in “The effect of listener accent background on accent perception and comprehension.” Their paper describes experiments where three classes of English speakers (US, British, and nonnative) perform an accent classification task and a transcription task using English speech recordings that include three different regional accents (Belfast, Cambridge, and Cardiff). For both tasks, significant effects are seen for listener accent background, speaker accent type, and the interaction of these two factors as well. In light of this and other experimental results, they conclude that accent perception must involve both speech perception and language processing.

We hope that this diverse collection of works serves to inform readers about current successes and also to inspire them in further efforts to model the various attributes of human auditory perception, or apply such models to the important open problems in speech, audio, and music processing.

ACKNOWLEDGMENTS

The guest editors extend thanks to all of the authors and reviewers who have made this special issue possible.

Jont B. Allen
Wai-Yip Geoffrey Chan
Stephen Voran