Guest Editors' Introduction: Special Section on Emergent Systems, Algorithms, and Architectures for Speech-Based Human-Machine Interaction

Rodrigo Capobianco Guido, *Member*, *IEEE*, Li Deng, *Fellow*, *IEEE*, and Shoji Makino, *Fellow*, *IEEE*

WELCOME to this special section of the *IEEE Transactions on Computers* (*TC*) on emergent systems, algorithms, and architectures for speech-based human-machine interaction.

Recently, we have witnessed significant progress on software and hardware techniques for human-machine interaction, especially for speech-based ones. While highly successful at the core, with many commercial applications resulting from this progress, the techniques still require a great deal of improvement, especially in the areas of speech signal processing, robustness of speech systems, speech modeling, learning algorithms, real-time decoding algorithms, speech quality enhancement, speech synthesis, speaker identification, speech source separation, and so on, to further impact speech-centric human-machine interaction. This special section concentrates on the state-of-the-art of such techniques. We received many submissions for consideration. We finally selected nine papers among them, which have been critically peer-reviewed with top-quality research and/or with more timely information than the papers that were not selected. The selected papers have addressed a considerable number of new ideas and applications in the field of speech-based human-machine interaction.

The first paper we present in this issue is entitled "Architecture, User Interface, and Enabling Technology in Windows Vista's Speech Systems," by Julian Odell and Kunal Mukerjee from Microsoft Corporation. In this paper, they describe the architecture, user interface, and key technologies that make up the speech system incorporated into Microsoft Windows Vista, which allows the combination of high accuracy and high usability for the end-to-end speech experience. The key elements of the speech user interface and how they maintain the user's ability to control the system despite limitations in the underlying recognition technology are clearly and carefully explained.

The second paper of this issue, by Maycel-Isaac Faraj and Josef Bigun, is entitled "Synergy of Lip Motion and Acoustic Features in Biometric Speech and Speaker Recognition." In this work, the authors present a novel and robust audio-visual digit and speaker-recognition system using lip-motion and speech biometrics, where a verification barrier based on a person's lip movement is added to the system to guard against advanced spoofing attempts. The system is based on a Support Vector Machine (SVM) and a Gaussian Mixture Model (GMM) and is tested with a database of about 300 different identities as well as the XM2VTS database.

Roger Moore is the author of the third paper of this issue, "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction," where a novel architecture for speech-based human-machine interaction is presented. It is inspired by recent findings in the neurobiology of living systems. PRESENCE, which is the acronym for *PREdictive SENsorimotor Control and Emulation*, blurs the distinction between the core components of a traditional spoken language dialogue system, focusing on a recursive hierarchical feedback control structure. It is founded on a model of interaction in which the system has in mind the needs and intentions of a user and vice versa.

"Speaker Verification via High-Level Feature-Based Phonetic-Class Pronunciation Modeling" is the fourth paper of this issue, authored by Shi-Xiong Zhang, Man-Wai Mak, and Helen Meng. The focus of this paper is to improve the results obtained with the articulatory feature-based conditional pronunciation models (AFCPMs) for speaker verification by grouping similar phonemes into phonetic classes and representing background and speaker models as phonetic-class dependent density functions. Evaluations based on the 2000 NIST SRE show that this phonetic-class approach effectively alleviates the data spareness problem encountered in conventional AFCPM, which results in better performance when fused with acoustic features.

The next paper of this issue is "Incorporating Knowledge Sources into a Statistical Acoustic Model for Spoken Language Communication Systems," where Sakriani Sakti, Konstantin Markov, and Satoshi Nakamura introduce a

- *R.C. Guido is with the SpeechLab, University of São Paulo (USP), Institute of Physics at São Carlos (IFSC), Department of Physics and Informatics (FFI), Avenida Trabalhador São Carlense 400, São Carlos, SP 13560-970, Brazil. E-mail: guido@ifsc.usp.br.*
- *L. Deng is with Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399. E-mail: deng@microsoft.com.*
- *S. Makino is with NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Kyoto 619-0237 Japan. E-mail: maki@cslab.kecl.ntt.co.jp.*

For information on obtaining reprints of this article, please send e-mail to: tc@computer.org.

0018-9340/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

general framework to incorporate additional sources of knowledge into an Hidden Markov Model (HMM)-based statistical acoustic model. Their method has the advantages of allowing the probabilistic relationship between information sources to be learned and of facilitating the decomposition of the joint probability density function (PDF) into a linked set of local conditional PDFs. The results demonstrated by the authors revealed that it improved the word accuracy with respect to standard HMMs, with or without additional sources of knowledge.

The field of rich transcription research, which includes speaker diarization together with the annotation of sentence boundaries and the elimination of speaker disfluencies, is explored by José M. Pardo, Xavier Anguera, and Chuck Wooters in their paper entitled "Speaker Diarization for Multiple-Distant-Microphone Meetings Using Several Sources of Information." The authors analyze the correlation between signals coming from multiple microphones and propose an improved method for carrying out speaker diarization for meetings with multiple distant microphones. In fact, they achieve state-of-the-art performance on the NIST Spring 2006 rich transcription evaluation database, improving the Diarization Error Rate (DER) by 15 percent to 20 percent relative to previous systems.

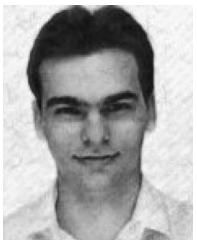
Chien-Lin Huang and Chung-Hsien Wu are the authors of our next paper, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis." In this work, a novel approach to generating phonetic units in order to recognize mixed-language or multilingual speech is presented, where the Hyperspace Analog to Language (HAL) model is adopted for contextual modeling and contextual similarity estimation and the Multidimensional Scaling (MDS) method is applied for reducing dimensionality. The created phonetic set provides a compact and robust set that considers acoustic and contextual information for mixed-language or multilingual speech recognition in systems that utilize speech-based human-machine interaction.

"Unsupervised Speaker Change Detection Using SVM Training Misclassification Rate" is the next paper, authored by Po-Chuan Lin, Jia-Ching Wang, Jhing-Fa Wang, and Hao-Ching Sung. They present an unsupervised speaker change detection algorithm based on Support Vector Machines (SVMs) to detect speaker change in a speech stream. The proposed algorithm can identify speaker changes with less speech data collection, making it capable of detecting speaker segments with short durations. According to the experiments on the NIST Rich Transcription 2005 Spring Evaluation (RT-05S) corpus, their system gives a lowest missed detection rate.

Chi-Chun Hsia, Chung-Hsien Wu, and Jian-Qi Wu are the authors of the last paper, which is entitled "Conversion Function Clustering and Selection Using Linguistic and Spectral Information for Emotional Voice Conversion." They use a Gaussian Mixture Bi-gram Model (GMBM) as the conversion function to characterize the temporal and spectral evolution of speech signals and use it for emotional voice conversion, incorporating linguistic and spectral information for conversion function clustering and selection. Subjective and objective evaluations with statistical hypothesis testing are conducted to evaluate the quality of the converted speech, which compares favorably with previous methods in conversion-based emotional speech synthesis.

Last, we would like to thank the more than 40 referees who spent a considerable amount of time reviewing the papers and also the authors and coauthors of all of the papers submitted to our special section, including those whose papers could not be published in this issue due to extensive revision requests as well as the limited page budget. Special thanks are due to the Editor-in-Chief, Dr. Fabrizio Lombardi, for hosting our special issue and to Ms. Joyce Arnold for her excellent support during the submission and peer review process. We hope you appreciate the reading.

Warmest Regards,
Rodrigo Capobianco Guido
Li Deng
Shoji Makino
Guest Editors



Rodrigo Capobianco Guido received the BSc degrees in computer science and in computer engineering, the MSc degree in electrical engineering, and the PhD degree in computational applied physics, respectively, from São Paulo State University (UNESP) at São José do Rio Preto, Brazil, from the Educational Foundation at Votuporanga (FEV), Brazil, in 1998, from Campinas State University (UNICAMP), Brazil, in 2000, and from the University of São Paulo at São Carlos (USP), Brazil, in 2003, all focusing on signal processing. He has worked with signal processing since 1995. He has already participated in two postdoctoral programs in speech processing and wavelets at USP, from 2003 to 2007. He has taught wavelets, speech and audio processing, programming languages, and electronics since 1999 and has published scientific articles in IEEE and Elsevier journals and magazines plus papers in conferences. He has served or is serving as as a guest editor for many special issues of IEEE and Elsevier journals and as an organizer and chairman for IEEE conferences, including IEEE ISM '05, ISM '06, ISM '07, and ICSC '07. He is a member of the editorial board of scientific journals, including *Pattern Recognition Letters*, *Neurocomputing*, and the *International Journal of Semantic Computing*, and has also served as a reviewer for many IEEE and Elsevier journals and conferences (~140 reviewed papers). He received several grants and awards from the State of São Paulo Research Foundation (FAPESP), from the National Council of Research and Development (CNpQ), and the Ministry of Education and Culture (MEC) of Brazil. He has supervised many theses in his field and is a member of the IEEE. The main objective of his group's work is concentrated in digital speech and audio processing (analysis, synthesis, pattern matching and recognition, voice morphing, compression, and so on), especially based on wavelets, neural networks, plus other statistical techniques.

Li Deng received the PhD degree from the University of Wisconsin-Madison in electrical engineering. In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada, as an assistant professor and became a full professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and, from 1997-1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, Washington, as a senior researcher, where he is currently a principal researcher. He is also an affiliate professor in the Department of Electrical Engineering at the University of Washington, Seattle. His research areas include automatic speech and speaker recognition, statistical methods and machine learning, neural information processing, machine intelligence, statistical signal processing, digital communication, human speech production and perception, acoustic phonetics, auditory speech processing, noise robust speech processing, speech synthesis and enhancement, spoken language understanding and systems, multimedia signal processing, and multimodal human-computer interaction. In these areas, he has published more than 250 refereed papers in leading international conferences and journals, 12 book chapters, and has given keynotes, tutorials, and lectures worldwide. He has been granted more than 20 US or international patents in acoustics, speech and language technology, and signal processing. He has authored two books: *Speech Processing—A Dynamic and Optimization-Oriented Approach* (Marcel Dekker Publishers, 2003) and *Dynamic Speech Models—Theory, Algorithms, and Applications* (Morgan & Claypool Publishers, 2006). He served on the Education Committee and the Speech Processing Technical Committee of the IEEE Signal Processing Society (1996-2000, and was associate editor for the *IEEE Transactions on Speech and Audio Processing* (2002-2005). He currently serves on the Society's Multimedia Signal Processing Technical Committee and is an area editor of *IEEE Signal Processing Magazine*. He was a technical chair of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04) and is the general chair of the IEEE Workshop on Multimedia Signal Processing, 2006. He is a fellow of the Acoustical Society of America and a fellow of the IEEE.

Shoji Makino received the BE, ME, and PhD degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981 and he is now an executive manager at the NTT Communication Science Laboratories. He is also a guest professor at Hokkaido University. His research interests include adaptive filtering technologies and realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech. He received the ICA Unsupervised Learning Pioneer Award in 2006, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002, the TELECOM System Technology Award of the TAF in 2004, the Best Paper Award of the IWAENC in 2003, the Achievement Award of the IEICE in 1997, and the Outstanding Technological Development Award of the ASJ in 1995. He is the author or coauthor of more than 200 articles in journals and conference proceedings and is responsible for more than 150 patents. He is a tutorial speaker at ICASSP '07 and was a panelist at HSCMA '05. He is a member of both the Awards Board and the Conference Board of the IEEE Signal Processing Society. He is an associate editor of the *IEEE Transactions on Speech and Audio Processing* and an associate editor of the *EURASIP Journal on Applied Signal Processing*. He is a guest editor of a special issue of the *IEEE Transactions on Audio, Speech and Language Processing*. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE Signal Processing Society and the chair-elect of the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society. He is the chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ. He is a member of the International IWAENC standing committee and a member of the International ICA steering committee. He is the general chair of WASPAA '07 in Mohonk, was the general chair of IWAENC '03 in Kyoto and the organizing chair of ICA '03 in Nara. He is an IEEE fellow, a council member of the ASJ, a member of EURASIP, and a member of the IEICE.