IEEE TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.ieee.org/sp/index.html

JULY 2006 VOLUME 14 NUMBER 4 ITASD8 (ISSN 1558-7916)

SPECIAL SECTION ON EXPRESSIVE SPEECH SYNTHESIS

(Contents Continued on Back Cover)

IEEE

Editorial
Special Section on Expressive Speech Synthesis

EXPRESSIVE speech synthesis (ESS) is a multidisciplinary research area that addresses some of the most complex problems in speech and language processing. The challenges posed by ESS have been the subject of several collaborative research projects across universities and research institutes around the world. Over the last decade, ESS has benefited from advances in speech and language processing as well as from the availability of large conversational-speech databases. These advances have spurred research on the expressiveness of speech and on conveying paralinguistic information including emotion, speaker-state, and speaker-listener relationships. There have also been substantial efforts towards automating database creation and evaluating the quality of speech synthesized for a variety of tasks that require not just the transmission of information, but also the expression of affect.

ESS should thus address the problem of WHAT paralinguistic information is encoded and HOW this information is encoded in signals. Several proposals are made in this section to cope with the mapping between substance and content. This should not obscure the fact that the dual problem of information representation and signal characterization is still a largely open question. ESS is a technology that should remain in close contact with theoretical debates on the cognitive representations and social aspects of expressions and emotions, as well as with the latest developments in signal manipulation techniques.

ESS opens up numerous new applications and challenges for speech synthesis where listeners and the situation of communication can be known and individuated. Paralinguistic dimensions should obviously be identified and synthesized in speech-to-speech translation. Being able to carry paralinguistic information through voice and to simulate emotions is also of particular interest in situated human–computer interaction, where the system should not only give information but also signal empathy and maintain attention as well as handle back-channeling and turn-taking. These paralinguistic dimensions of human face-to-face communication are crucial for giving presence and humanness to embodied conversational agents and humanoid robots.

For this special section, we solicited original theoretical and practical work offering new and broad views of the latest research in expressive speech synthesis, and the reaction has been excellent. We received 20 submissions spanning a variety of topics ranging from signal processing to prosody prediction, addressing rule-based, HMM, and concatenative approaches. All of the eight papers that were accepted appear in this section.

The first paper, by Pitrelli *et al.*, describes the IBM Expressive Text-to-Speech Synthesis System for American English and discusses concatenative versus rule-based approaches to the synthesis of four expressive styles of speech. This is followed by Wu *et al.*'s proposal for a new method of "Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis." Generating expressive speech by converting "neutral" voices is a challenging task. The paper describes a method to transform neutral speech to expressive styles including happiness, sadness, anger, confusion, apology and question.

Navas *et al.*, in their paper entitled "An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS," studied the value of semantic content in recording emotional databases. Their findings suggest the possibility for using semantically unspecified content to record databases for emotional speech synthesis at the cost of exaggerated emotional speech output. These findings help to advance research in expressive speech synthesis for new expressive states especially for languages lacking linguistic and speech resources such as Basque, the main language used in the study.

In the paper entitled "Expressing Degree of Activation in Synthetic Speech" by Schröder, the author formulates a set of prosody rules linking speech prosody to a three-dimensional emotional space that is common in the psychological literature, namely, activation, evaluation, and power. Implementing those rules, the author shows a correlation between the prosody parameters used in the study and the activation dimension using a multimodal perception test.

The special characteristics of expressive speech are particularly addressed in the description by Theune *et al.* of the modifications made to the prosodic module of a text-to-speech synthesizer for "Generating Expressive Speech for Storytelling Applications." Rules have been identified for maintaining interest and creating suspense at key turns of the story.

The importance of prosody is also highlighted in Tao's "Prosody conversion from neutral speech to emotional speech," and Gu *et al.* "Modeling the Effects of Emphasis and Question on Fundamental Frequency Contours of Cantonese Utterances." The first paper introduces three different models, a linear modification model (LMM), a Gaussian mixture model (GMM), and a classification and regression tree (CART) model for prosody conversion of "neutral" speech to emotional speech, and compares evaluation results for each method. The second paper investigates the prosodic features of spoken Cantonese in the context of a command-response model where the focus is on F0 prediction with careful considerations regarding the effects of questions and emphasis.

Digital Object Identifier 10.1109/TASL.2006.878306

The section closes with a position paper from the lead editor addressing the special needs of speech synthesis in conversational applications and illustrating the use of nonverbal speech utterances, such as laughs, backchannels, and grunts, many of which have similar phonetic structures, but signal their different meanings through variations in prosody and voice qualities.

As a final word, we would like to thank the Editor-in-Chief, Mari Ostendorf, and her predecessor, Isabel Trancoso, for their immense help and guidance throughout the process. Our thanks also go to the more than 50 reviewers who did a wonderful job of filtering and improving the papers, and especially to Kathy Jackson from the IEEE Signal Processing Society for her kind assistance in assembling the section.

NICK CAMPBELL, *Guest Editor*
Nat'l Inst. of Information & Communications Technology and ATR Spoken Language Communications Laboratories
Kyoto 619-0288, Japan
nick@nict.go.jp

WAEL HAMZA, *Guest Editor*
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598 USA
hamzaw@us.ibm.com

HARALD HÖGE, *Guest Editor*
Siemens AG Corporate Technology
Munich D-81730, Germany
harald.hoege@siemens.com

JIANHUA TAO, *Guest Editor*
National Laboratory of Pattern Recognition,
Chinese Academy of Sciences
Beijing 100080, China
jhtao@nlpr.ia.ac.cn

GÉRARD BAILLY, *Guest Editor*
Institut de la Communication Parlée
Grenoble 38031, France
bailly@icp.inpg.fr

Nick Campbell received the Ph.D. degree in Experimental psychology from the University of Sussex, Sussex, U.K.

He is currently engaged as a Chief Researcher in the Department of Acoustics and Speech Research, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan, where he also serves as Research Director for the JST/CREST Expressive Speech Processing and the SCOPE "Robot's Ears" projects. He was first invited as a Research Fellow at the IBM U.K. Scientific Centre, where he developed algorithms for speech synthesis, and later at the AT&T Bell Laboratories, where he worked on the synthesis of Japanese. He served as Senior Linguist at the Edinburgh University Centre for Speech Technology Research before joining ATR in 1990. His research interests are based on large speech databases, and include nonverbal speech processing, concatenative speech synthesis, and prosodic information modeling. He spends his spare time working with postgraduate students as Visiting Professor at the Nara Institute of Science and Technology (NAIST), Nara, Japan, and at Kobe University, Kobe, Japan.

Wael Hamza was born in Cairo, Egypt, in 1969. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Faculty of Engineering, Cairo University, Cairo, Egypt in 1991, 1995, and 2001 respectively.

During these years, he served as a Researcher in the Research and Development International (RDI), Giza, Egypt, developing algorithms for speech and language processing. Since 2001, he has been a member of the Speech Synthesis Group, IBM T. J. Watson Research Center, Yorktown Heights, NY.

Harald Höge received the Ph.D. degree in physics from the University of Frankfurt/Main, Frankfurt, Germany, in 1974.

In 1970, he joined Siemens AG, Munich, Germany, working on echo compensation and speech coding. Since 1978, he has been leading a research group focusing on speech recognition, speech synthesis, and speaker characterization. He initiated and was involved in many national, European, and international project such as SPICOS, C-STAR, Verbmobil, SpeechDat, SALA, SPEECON LC-STAR, ECESS, and TC-STAR. He gives lectures at the Universität der Bundeswehr München on speech and image processing, where he was honored in 2001 with the title "Honorary Professor." He is author and coauthor of 80 papers and holds 25 patents.

Jianhua Tao (M'98) received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001.

He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include speech synthesis, speaker simulation, affective computing, and multimedia integration. He has published more than 60 papers in major journals and proceedings.

Dr. Tao received several awards from the important conferences, such as Eurospeech, NCMMSC, etc. He was elected as the chair or program committee member for several major conferences. Currently, he is the Secretary of Speech Information Processing Committee of Chinese Information Processing Society and Secretary of the special interest group of Chinese Spoken Language Processing in ISCA.

Gérard Bailly joined the Institut de la Communication Parlée (ICP), Grenoble, France, in 1986 as Chargeé de Recherche with the French National Center for Scientific Research (CNRS) after two years as a Postdoctoral Fellow with INRS Telecommunications in Montréal, Canada. As Directeur de Recherches since 2002, he is head of the ICP Talking Machines team and Joint Director of the research federation ELESA. He coorganized the first speech synthesis conference in 1991 and cochaired the first international conferences on smart objects and ambient intelligence in 2003 and 2005. He is the editor of three books, author of more than 20 papers in international journals, 15 book chapters, and 150 papers in international conferences. His current interest is audiovisual synthesis and multimodal interaction with conversational agents in face-to-face situated communication.