Introduction to the Issue on Statistical Parametric Speech Synthesis

STATISTICAL parametric speech synthesis has been a hot topic for some time. Classic statistical parametric speech synthesizers are able to produce fairly natural-sounding and flexible voices, needing only a relatively small training database, and can be more easily adapted to a new voice or speaking style than concatenative systems.

Although this approach has proved very successful in recent decades, there are still several open problems in this rich field. For instance, (a) how to build a statistical parametric speech synthesizer that can generate emotional or expressive speech, and how to use it in natural dialog applications; (b) whether, and how, to integrate physical speech production models into statistical parametric speech synthesis to improve the speech quality; (c) how to build high quality statistical parametric speech synthesizers using only a *very* small training database; (d) how to build parametric synthesizers or apply speaker adaptation for cross-language or multilingual applications. Motivated by these open problems, this special issue includes papers on expressive speech synthesis, multilingual techniques, excitation/generation models, adaptation and also introduces some new applications of statistical parametric speech synthesis.

The first group of papers suggests innovations in the area of statistical modeling. Koriyama *et al.* propose a statistical parametric speech synthesis technique based on Gaussian process regression (GRP) instead of the Hidden Markov Model (HMM). There is no question that the HMM is the most successful statistical models in spoken language technology and has been applied to almost every imaginable application, from speech and speaker recognition to speech synthesis. However, there are other statistical models which may be suitable for speech technology, including speech synthesis, such as GRP.

The second group of papers considers the parametric representation of speech for statistical parametric speech synthesis. Erro *et al.* explore the potential of the harmonic-plus-noise model of speech in the development of a high-quality vocoder applicable to statistical parametric speech synthesis. Cabral *et al.* propose an analysis method to separate the glottal source and vocal tract components of speech, which can produce high-quality synthetic speech using an acoustic glottal source model. Csapó *et al.* propose two alternative extensions of excitation models in order to model irregular phonation, which typically occurs phrase-finally. From these papers, we conclude that to continue to improve the speech quality of statistical parametric speech synthesis, further attention needs to be paid to the parametric representation of speech. These papers make the following contributions to that effort:

1) Improve the existing parametric representation of speech, Erro, *et al.*;
2) Propose a new parametric representation of speech, Cabral, *et al.*;
3) Parametric representation of irregular phonation, Csapó, *et al.*

and together they suggest a variety of future directions in this area.

The third group of papers focuses on the speech parameter generation stage of statistical parametric speech synthesis. The speech generated from statistical parametric speech synthesis is often said to suffer from over-smoothing caused by statistical modeling. The standard parameter generation algorithm, through which the speech parameters are generated from trained HMMs, produces a sequence of observation vectors according to a maximum likelihood optimization criterion. In the papers in this issue, two ways to overcome the supposed over-smoothing problem are presented. One is to adjust the optimization criterion of the parameter generation algorithm. Nose *et al.* introduce a parameter generation algorithm using a local variance (LV) model and provide experimental results that indicated better results than using global variance (GV). The other way is to improve the accuracy of HMMs themselves. Takaki *et al.* examine the use of a spectral modeling technique based on an additive structure of context dependencies.

Although improvements to the parametric representation of speech and the parameter generation algorithm are both shown to improve synthetic speech quality, still the speech signal generated from the vocoder is of lower quality than the original speech waveform. Is it therefore natural to think about ways to directly use original speech waveforms in statistical parametric synthesis. So, the fourth group of papers in this issue concern a hybrid method in which the parameters generated from a statistical model are used to guide concatenative speech synthesis using a stored speech corpus. Takamichi *et al.* use rich context models to construct a hybrid speech synthesis system that combines HMM-based speech synthesis and unit selection waveform generation. This hybrid method combines some of the advantages of both techniques, such as flexibility, rapid system construction and high-quality speech. A key problem in this hybrid method is how to define the target and join costs for the unit selection synthesis, based on the HMM-generated speech parameters. It seems that further work on this aspect of hybrid speech synthesis system is still needed.

The fifth group of papers is about adaptation. In these papers, two new adaptation methods are proposed. Sung *et al.* introduce factored maximum penalized likelihood kernel regression (FMLKR) and Saheer *et al.* combine vocal tract length normalization (VTLN) with linear transforms in a hierarchical Bayesian framework. Urbain *et al.* exploit the adaptation of HMM-based speech synthesis to generate laughter. Karhila *et al.* analyze the robustness to noise of HMM adaptation. The

Digital Object Identifier 10.1109/JSTSP.2014.2309416

flexibility of speech adaptation in statistical parametric speech synthesis is also covered: Wan *et al.* describe the application of average voice models (AVMs) and a novel application of cluster adaptive training (CAT) with multiple context-dependent decision trees to create HMM-TTS voices using diverse data. Together, this group of papers investigate a number of important issues in constructing adaptation system, namely the adaptation method, corpus construction, and the application of adaptation during synthesis.

The final group of papers is about applications of statistical parametric speech synthesis. Picart *et al.* focus on the automatic modification of the degree of articulation of an existing standard neutral voice in the framework of HMM-based speech synthesis. Chen *et al.* explore an integrated approach in order to construct a synthesizer with more expressions. Schabus *et al.* use a joint audiovisual hidden semi-Markov model for speech synthesis. These papers all showcase the effectiveness of statistical parametric speech synthesis in expressive speech synthesis. What we don't see in the papers in this issue are cross-language or multi-lingual applications. So, it would seem that more attention should be paid to how to use the technology of statistical parametric speech synthesis in those applications.

Together, the papers included in this special issue cover a broad range of current issues in statistical parametric speech synthesis, each of them offering novel ideas and improvements to the state of the art. Nevertheless, the speech generated from statistical parametric speech synthesis still has inadequate quality and naturalness, and it is likely this is a combined effect of the oversimplified parametric representation of speech used by the vocoder and the subsequent statistical modeling. The first four groups of papers each focus on different ways to improve the quality of synthesized speech by addressing one or other of those components. The last two groups tell us about the flexibility of statistical parametric speech synthesis, which is its key 'selling point' and the reason that continued work to improve the speech quality is warranted.

The editors would like to sincerely thank both the EiC for recommending this special issue be published and for his patience in guiding us—the guest editors—through the process, the publication assistants who provided invaluable assistance in navigating both the process and the Manuscript Central website. We hope that you—the reader—will be stimulated by the interesting work represented in this special issue and that it helps spur additional research within this exciting and growing field.

JIANHUA TAO, *Lead Guest Editor*
National Laboratory of Pattern Recognition,
Institute of Automation,
Chinese Academy of Sciences
Beijing 100190, China
jhtao@nlpr.ia.ac.cn

KEIKICHI HIROSE, *Guest Editor*
University of Tokyo
Tokyo 113-8654, Japan
hirose@gavo.t.u-tokyo.ac.jp

KEIICHI TOKUDA, *Guest Editor*
Nagoya Institute of Technology
Nagoya 466-8555, Japan
tokuda@nitech.ac.jp

ALAN W. BLACK, *Guest Editor*
Carnegie Mellon University
Pittsburgh, PA 15213 USA
awb@cs.cmu.edu

SIMON KING, *Guest Editor*
The University of Edinburgh
Edinburgh EH8 9YL, U.K.
Simon.King@ed.ac.uk

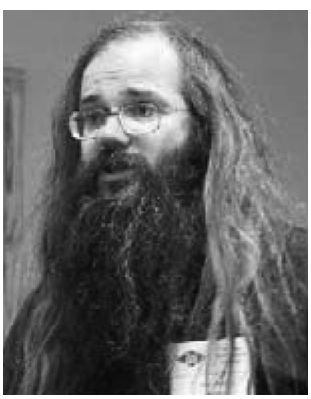
Jianhua Tao (M'98) received the M.S. degree from Nanjing University, Nanjing, China, in 1996 and the Ph.D. degree from Tsinghua University, Beijing, China, in 2001.

He is the Professor and Deputy Director at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China. His interests include speech synthesis and recognition, human–computer interaction, and emotional information processing. He has published more than 100 papers in major journals and proceedings, such as the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, ICASSP, Interspeech, ICME, and ICPR. He is the Executive Committee member of the AAAC association. He is the Editorial Board Member for the *Journal on Multimodal User Interfaces* (JMUI), the *International Journal of Synthetic Emotions* (IJSE), the Steering Committee Member for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and the Subject Editor for *Speech Communication*.

Keikichi Hirose (M'78–SM'12) received Ph.D. degree in electronic engineering in 1977 from the University of Tokyo. He has been a Professor at the University of Tokyo since 1994. He has been engaged in a wide range of research on spoken language processing, including analysis, synthesis, recognition, dialogue systems, and computer-assisted language learning.

From 2000 to 2004, he was Principal Investigator of the national project "Realization of advanced spoken language information processing utilizing prosodic features," supported by the Japanese Government. He served as the general chair for INTERSPEECH 2010, Makuhari, Japan. Since 2010, he has served as the Chair of ISCA Special Interest Group on Speech Prosody (SProSIG). He serves as a Board member for International Speech Communication Association. He published more than 100 journal papers and more than 340 conference papers.

Keiichi Tokuda is the director of the Speech Processing Laboratory and a Professor in the Department of Computer Science at the Nagoya Institute of Technology. He has been working on HMM-based speech synthesis after he proposed an algorithm for speech parameter generation from HMM in 1995. He is also the principal designer of opensource software packages: HTS (http://hts.sp.nitech.ac.jp/) and SPTK (http://sp-tk.sourceforge.net/). In 2005, Dr. Alan Black (CMU) and Keiichi Tokuda organized the largest ever evaluation of corpus-based speech synthesis techniques, the Blizzard Challenge, which has progressed to an annual event. He is an IEEE Fellow and ISCA Fellow. He published over 80 journal papers and over 200 conference papers, and received six paper awards and two achievement awards.



Alan W. Black is a Professor in the Language Technologies Institute at Carnegie Mellon University. Before joining the faculty at CMU in 1999, he worked in the Centre for Speech Technology Research at the University of Edinburgh, and before that at ATR in Japan. He is one of the principal authors of the free software Festival Speech Synthesis System, the FestVox voice building tools and CMU Flite, a small footprint speech synthesis engine, that is the basis for many research and commercial systems around the world. He also works in spoken dialog systems, the LetsGo Bus Information project and mobile speech-to-speech translation systems. Prof Black is an elected member of ISCA board (2007–2015). He has over 200 refereed publications and is one of the highest cited authors in his field.

Simon King (M'95–SM'08) holds M.A. and M.Phil. degrees from Cambridge and a Ph.D. from the University of Edinburgh. He has been with the Centre for Speech Technology Research at the University of Edinburgh since 1993, where he is now Professor of Speech Processing and the director of the center. His interests include speech synthesis, recognition and signal processing and he has around 170 publications in these areas. He has served on the ISCA SynSIG board and co-organizes the Blizzard Challenge. He previously served on the IEEE SLTC and as an associate editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and is a current associate editor of *Computer Speech and Language*.