Introduction to the Special Section on Voice Transformation

VOICE Transformation aims at the control of non-linguistic information of speech signals such as voice quality and voice individuality. It refers to the various modifications one may apply to the sound produced by a person, speaking or singing. Voice Transformation covers a wide area of research from speech production modeling and understanding to perception of speech, from natural language processing, modeling, and control of speaking style, to pattern recognition and statistical signal processing.

While there are common research interests with speaker-dependent technologies (e.g., speaker recognition/verification), Voice Transformation goes beyond these technologies; not only the cues that are relevant to voice individuality should be detected but the corresponding features need to be modified in a way that the transformed speech signal sounds natural. Speech models suggested for Voice Transformation should be able to manipulate efficiently these features. Building high-quality Voice Transformation systems requires to take into account phenomena that are usually ignored in other speech research and technology areas. This includes the nonlinear nature of speech, and the interaction of vocal tract and source characteristics. Modulation phenomena should be treated carefully during transformation. For Voice Transformation, a fusion of prosodic features at levels higher than that of the speech wave should be performed in order to define the speaking style of a speaker, to recognize characteristic patterns, and then suggest techniques to map one speaking style to another or just conduct convincing transformations on a style. In a few words, Voice Transformation subsumes speech understanding.

Voice Transformation was considered as a hot, novel, and fast-growing topic in the 1990s, having as a potential application the concatenated speech synthesis systems where new (virtual or target) voices could be created without requiring to pass through the extremely expensive process of developing new voices. By that time, it was widely accepted that Voice Transformation systems were far from providing the required performance. With the recent developments in speech synthesis, this need is more pronounced. There is an increasing demand for high-quality Voice Transformation methods not only for creating target or virtual voices, but also to model various effects (e.g., Lombard effect), to convey expressiveness or emotions, to make more natural the dialog systems which use speech synthesis, etc. Besides speech synthesis, Voice Transformation has other potential applications in areas like entertainment, film, and music industry, toys, chat rooms and games, dialog systems, security and speaker individuality for interpreting telephony, high-end hearing aids, vocal pathology, and voice restoration.

Furthermore, we have witnessed emergence of research programs in the area such as the European Union-funded efforts in Emerging Technologies and Infrastructures (FET) and Information Society Technologies (IST) in general, as well as in U.S. Air Force-funding programs, NIH, and NSF.

Digital Object Identifier 10.1109/TASL.2010.2051826

An inherent characteristic in Voice Transformation is the mapping function which describes the way that the speech signal or its features should be transformed for a given transformation/modification task. For example in a specific Voice Transformation application, that of Voice Conversion, there is a mapping function which transforms the features of the source speaker to the features of the target speaker. The paper by Helander *et al.* introduces the use of Partial Least Squares-based transforms in voice conversion, in order to prevent overfitting during the mapping stage. A different approach for voice conversion is suggested by Erro *et al.* based on a weighting frequency warping approach. This is combined with a GMM-based approach for constructing an energy correction filter.

Most of the approaches suggested in the literature for Voice Transformation and Voice Conversion are based on the assumption that parallel databases are available for training purposes. However, there are many applications where parallel databases is impossible, like in cross-lingual voice conversion, where the source and target speakers speak different languages. There are two papers in this special issue that address the problem of voice conversion using nonparallel databases, which is also referred to as text-independent voice conversion. The paper by Zheng *et al.* suggests a new data alignment method for text-independent voice conversion considering both the phonetic accuracy and preservation of the internal topology of parameter space. To ensure accuracy, they use phonetic labels of the training data as supervisory information, and phonetic restriction is considered for supervised alignment. The investigation of Erro *et al.* is based on existing voice conversion techniques, without the requirement of any phonetic or linguistic information. They suggest a new iterative alignment method that allows pairing phonetically equivalent acoustic vectors from nonparallel utterances from different speakers, even under cross-lingual conditions. Desai *et al.* go even beyond the requirement of having some recordings from the source and target speakers (in a text-independent context). They suggest a voice conversion system based on Artificial Neural Networks (ANNs) which is able to capture speaker-specific characteristics of a target speaker while it avoids the need for speech data from a source speaker during training.

Voice Transformation is one way to increase the range of expressiveness of current speech synthesis systems without major requirements in the design and size of their databases. The paper by Türk and Schröder investigates the use of voice conversion and speech modification techniques for transforming neutral synthetic speech into aggressive, cheerful, and depressed speech. The paper by Erro *et al.* deviates from the above paper in that they propose an emotion conversion system based on prosodic unit selection strategies. They also use voice conversion algorithms for the modification of voice quality since these modifications are considered essential for a reasonable perception of the emotions. Two papers on HMM-based speech synthesis by Yamagishi *et al.* and Watts *et al.* show how an "average voice model," plus model adaptation can be

effciently used to create many new voices for speech synthesis. More specifically, Watts *et al.* uses HMM adaptation and compares this with GMM-based voice conversion techniques to synthesize child speech from an existing synthesizer using a gender-mixed average adult voice. The paper by Bedenbaugh *et al.* describes a rather special application of voice transformation to circumvent some of the limitations of brain mapping in the study of the central neuronal representation of speech and other natural sounds.

The paper by Bedenbaugh *et al.* describes a rather special application of voice transformation to circumvent some of the limitations of brain mapping in the study of the central neuronal representation of speech and other natural sounds.

Developing methods for the evaluation of voice transformation results is an important issue in developing more efficient algorithms in the future. Mainly, listening tests have been suggested for this purpose. The paper by Felps and Gutierrez-Osuna suggests the use of three objective measures for evaluating the effectiveness of accent conversion methods; transforming foreign-accented speech into its native-accented counterpart.

To the best of our knowledge, this is the second special issue on Voice Transformation in any scientific journal. We hope that the readers will find this issue inspiring and it will provide an excellent rostrum to researchers for more advancements in the area. We would like to thank the authors of all submitted papers for their contributions and express our appreciation to the reviewers for their help. Further, we wish to offer our warm thanks to the Editor-in-Chief, Prof. Helen Meng, the past Editor-in-Chief, Prof. Mari Ostendorf, and the Publications Coordinator, Ms. Kathy Jackson, for their support and assistance during the preparation of this special issue.

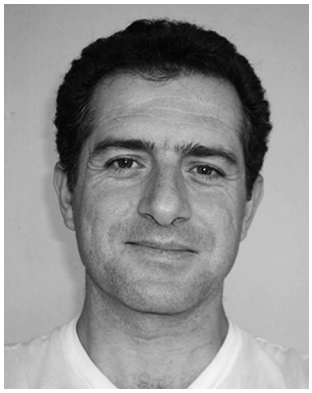
YANNIS STYLIANOU, *Lead Guest Editor*
Computer Science Department
University of Crete
GR-71409 Heraklion, Crete, Greece

TOMOKI TODA, *Guest Editor*
Graduate School of Information Science
Nara Institute of Science and Technology
Nara 630-0192, Japan

CHUNG-HSIEN WU, *Guest Editor*
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan 701, Taiwan

ALEXANDER KAIN, *Guest Editor*
Computer Science and Electrical Engineering Department
Oregon Health and Science University
Beaverton, OR 97006 USA

OLIVIER ROSEC, *Guest Editor*
Orange Labs, France Telecom R&D
22307 Lannion, Cedex, France



Yannis Stylianou (M'95) received Ph.D. degree in signal processing from the Ecole National Supérieure des Télécommunications, ENST, Paris, France, in 1996.

From 1996 until 2001, he was with AT&T Labs Research and in 2001 he joined Bell-Labs Lucent Technologies. Since 2002, he has been an Associate Professor at the Department of Computer Science, University of Crete, Heraklion, Crete, Greece. He currently participates in the EU FET-Open project 245491 LISTA—The Listening Talker (2010–2013). During Interspeech 2007 (Antwerp, Belgium), he gave a Tutorial on Voice Transformation. His research interests are in speech signal processing algorithms for speech analysis and statistical signal processing.

Tomoki Toda (M'95) received the Ph.D. degree in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2003.

He was a Research Fellow of the Japan Society for the Promotion of Science in the Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Japan, from 2003 to 2005. He is currently an Assistant Professor in the Graduate School of Information Science, NAIST. His research interests include speech transformation, speech synthesis, speech analysis, and speech recognition.

Dr. Toda received the 2009 Young Author Best Paper Award from the IEEE Signal Processing Society. During Fifth ISCA Speech Synthesis Workshop (SSW5), Pittsburgh, PA, he gave a Tutorial on Voice Transformation.

Chung-Hsien Wu (M'88–SM'03) received the Ph.D. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 1991.

Since August 1991, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He became Professor and Distinguished Professor in August 1997 and August 2004, respectively. His research interests include speech recognition, text-to-speech, multimedia information retrieval, spoken language processing, and sign language processing for the hearing-impaired.

Alexander Kain (M'02) received the Ph.D. degree in computer science and engineering from the Oregon Graduate Institute, Portland OR, in 2001.

He is a Research Assistant Professor at the Center for Spoken Language Understanding (CSLU), Computer Science and Electrical Engineering Department (CSEE), Oregon Health and Science University (OHSU), Beaverton, and a Senior Research Scientist at BioSpeech, Inc., Portland, OR. His current research focuses on speech signal processing algorithms for the quantitative assessment and transformation of clear and conversational speech, the transformation of aphonic and dysarthric speech to improve intelligibility and perceived voice quality, and improving text-to-speech synthesis quality and exibility.

Olivier Rosec (M'07) received the Ph.D. degree in signal processing from the Université de Bretagne Occidentale, Brest, France, in 2000.

From 1996 until 1999, he was with the Acoustics and Seismics Department, IFREMER, Brest, as a Research Engineer. In 2000, he joined the speech synthesis team of Orange Labs, Lannion, France, as a Senior Technical Staff Member. His current research interests include speech modeling and analysis, speech modification and coding for concatenative synthesis, voice transformation, and voice conversion.