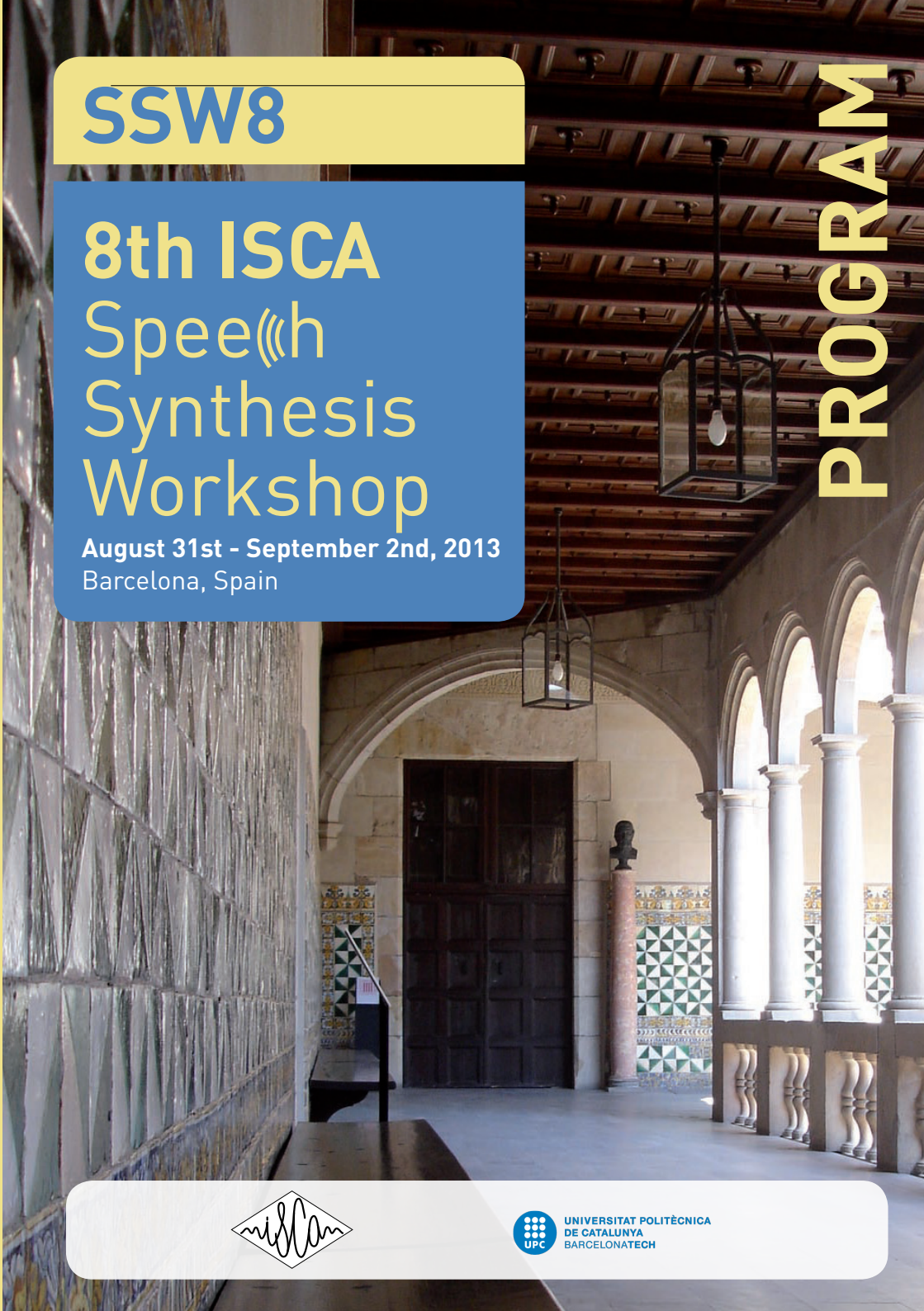

SSW8

8th ISCA Speech Synthesis Workshop

August 31st - September 2nd, 2013
Barcelona, Spain

PROGRAM

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

8th ISCA Workshop on Speech Synthesis

Program and Book of Abstracts

Barcelona • August 31 – September 2, 2013

Edited by Antonio Bonafonte

At the time of release, the proceedings can be downloaded from the website of the SSW8: ssw8.talp.cat

Cover Picture: Arxiu de l'Institut d'Estudis Catalans. Photographer: Santi Muxach

Contents

Message from the Chair

Welcome to the eight edition of the ISCA Speech Synthesis Workshop. Twenty-three years have passed since the first edition, in Autrans, France, 1990. It is a considerable period of time, and we have seen several generations of systems: synthesis by rule, diphone concatenation, unit selection synthesis and, last years, parametric/statistical speech synthesis. Each generation implies significant improvements on aspects as naturalness, intelligibility, flexibility, range of application or robustness. However, we have to recognize that many of the objectives of a decade ago, have still not been reached. For instance, while synthetic image and synthetic music is in some cases preferred to the real ones by the entertainment industry, synthetic speech is still just a defective copy of human speech. Human voices include a lot of information which we still are not able to code in the synthetic voices.

Each speech generation synthesis system has started by some innovative pioneer works followed by an explosion of research contributions that consolidate, extend and exploit the new technology;and finally, a stabilization where a lot of effort produces limited improvements. From my point of view, speech synthesis using Hidden Markov models is reaching this stabilization phase. At that phase, new revolutionary ideas are needed to significantly push the technology beyond of today limits. While there are still many contributions to be done in this field of parametric statistical synthesis, we need open minds to explore new approaches to speech synthesis that may become new paradigms.

The workshop program includes three invited conferences that we thing can be inspiring talks. On the first day, Dr. Heiga Zen will present recent application of deep learning to statistical parametric speech synthesis. In the last years, it has been proved that deep neural networks are better than traditional Gaussian Mixtures for acoustic Modeling in Speech Recognition. Sev-

eral papers have already been published applying the same ideas to speech synthesis. This seems a promising direction. On the second day, Prof. Nigel Ward will give a talk about prosodic patterns in dialog. It is widely accepted that we need to include information of higher level (as semantic/pragmatic) to produce conversational engines. Prof. N. Ward has analyzed the prosody patterns on real dialogs from several perspectives. Finally, on Monday 2nd, Prof. Xavier Serra will explain the evolution of synthetic singing voices and the research of his lab on this field. Synthetic Music has achieved very high quality and can be used to create synthetic songs as natural and expressive as songs produced by real singers. Although synthetic singing is a close area, traditionally these are developed by two different research communities and this is an opportunity to learn their approach and their perspectives to this topic.

The technical program includes 51 regular papers. Each paper has been reviewed by three reviewers. The program is composed of 6 oral sessions for a total of 20 papers and 3 poster sessions for 31 papers. Furthermore, there are 4 additional presentations in a demo session. I am very grateful to all the contributors for their interest and efforts that have made possible to organize the workshop.

The Blizzard Challenge 2013 Workshop, September 3, is organized by Simon King, Alan W Black, Keiichi Tokuda and Kishore Prahallad. The contribution of these evaluation campaigns to the progress of the area is remarkable. It is a pleasure that this edition can be hosted by Universitat Politècnica de Catalunya.

During the preparation of the workshop I have collaborated with many people and I am sincerely grateful to all of them. First of all, I want to mention the advisory committee for their helpful advises. I would like to thank all members of the Program Commitee for their excellent work of reviewing the papers in a very tight schedule. It has been a pleasure to work with the program co-chairs, Daniel Erro and David Escudero.

I am most grateful to Olga Nuñez and Yolanda López, from the Technical Secretariat, for all the support organizing the workshop and for their constant work.

Welcome to SSW8! I am looking forward to an exciting workshop and rapid progress in the field. I wish you a wonderful workshop time in Barcelona.

Barcelona, August 2013
Antonio Bonafonte, Workshop Chair

Organizing Committee

Chair:
Antonio Bonafonte, Universitat Politècnica de Catalunya

Members:
Daniel Erro, Ikerbasque – University of the Basque Country
David Escudero, Universidad de Valladolid
Asunción Moreno, Universitat Politècnica de Catalunya
Jordi Adell, PAL Robotics (Barcelona)
Ignasi Esquerra, Universitat Politècnica de Catalunya
Francesc Alías, La Salle – Universitat Ramon Llull

Advisory Committee

Gerard Bailly, CNRS/INPG, France.
Alan Black, CMU, USA.
Nick Campbell, Univ. of Dublin, Ireland.
Rolf Carlson, KTH, Sweden.
Thierry Dutoit, Faculté Polytechnique de Mons, Belgium.
Wolfgang Hess, Univ. of Bonn, Germany.
Julia Hirschberg, Columbia Univ., USA.
Simon King, Edinburgh University, UK.
Bernd Möbius, Saarland University, Germany.
Jan van Santen, OHSU, USA.
Juergen Schroeter, AT&T Labs, USA.
Paul Taylor, Google, UK.
Keiichi Tokuda, Nagoya Institute of Technology, Japan.

Program Committee

Program Chairs:
Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain
Daniel Erro, Ikerbasque – University of the Basque Country, Spain
David Escudero, Universidad de Valladolid, Spain

Members & board of reviewers:
Jordi Adell, Pal-Robotics S.L., Barcelona, Spain
Francesc Alías-Pujol, La Salle – Universitat Ramon Llull, Barcelona, Spain
Gerard Bailly, GIPSA-Lab, Grenoble, France
Roberto Barra-Chicote, Universidad Politécnica de Madrid, Spain
Jerome Bellegarda, Apple Inc., USA
Alan Black, Carniege Mellon University, Pittsburgh, USA
Andrew Breen, Nuance Communications, UK
Nick Campbell, University of Dublin, Ireland
Rolf Carlson, KTH, Stockholm, Sweden
Alistair Conkie, AT&T Labs, NJ, USA
Ricardo Cordoba, Universidad Politécnica de Madrid, Spain
Christophe D'Alessandro, CNRS-LIMSI, Orsay, France
Thierry Dutoit, TCTS Lab., Numediart Institute, University of Mons,Belgium
Raul Fernandez, IBM Research, Yorktown Heights, NY, USA
Juan-María Garrido, Universitat Pompeu Fabra, Barcelona, Spain
Xavi Gonzalvo, Google, UK
Inma Hernáez, University of the Basque Country, Spain
Wolfgang Hess, IfK Universität Bonn, Germany

Julia Hirschberg, Columbia University, NY, USA
Ignasi Iriondo, La Salle – Universitat Ramon Llull, Barcelona, Spain
Simon King, University of Edinburgh, UK
Esther Klabbers, Oregon Health & Science University, USA
Javier Latorre, Toshiba Research Europe, UK
Zhenhua Ling, University of Science and Technology of China (USTC)
Bernd Moebius, Saarland University, Germany
Juan Montero, Universidad Politécnica de Madrid, Spain
Asunción Moreno, Universitat Politècnica de Catalunya, Spain
Eva Navas, University of the Basque Country, Spain
Kishore Prahallad, International Institute of Information Technology
Eduardo Rodríguez Banga, University of Vigo, Spain
Juergen Schroeter, AT&T Labs, NJ, USA
Joan Claudi Socoró, La Salle – Universitat Ramon Llull, Barcelona, Spain
Frank Soong, Microsoft Research Asia, China
David Suendermann, DHBW Stuttgart, Germany
Jianhua Tao, Chinese Academy of Sciences, Beijing
Paul Taylor, Google, UK
Tomoki Toda, Nara Institute of Science and Technology, Japan
Keiichi Tokuda, Nagoya Institute of Technology, Japan
Jan Van Santen, Oregon Health & Science University
Junichi Yamagishi, University of Edinburgh, UK
Heiga Zen, Google, UK

Contributing Reviewers:
Enrico Bocchieri, AT&T Labs, NJ, USA
Kei Hashimoto, Nagoya Institute of Technology, Japan
Andrej Ljolje, AT&T Labs, NJ, USA
Taniya Mishra, AT&T Labs, NJ, USA
Hansjörg Mixdorff, BTH Berlin University of Applied Sciences, Germany
Keiichiro Oura, Nagoya Institute of Technology, Japan

Technical Program

Workshop Program at a Glance

Saturday, August 31

08:00	Registration opens
09:10 – 09:20	Opening
09:20 – 11:00	Oral Session 1: Prosody and pausing
11:00 – 12:40	Poster Session 1 & Coffee
12:45 – 14:45	Lunch Break
14:45 – 16:00	Oral Session 2: Open Challenges in speech synthesis
16:00 – 16:30	Coffee Break
16:30 – 17:20	Keynote Session 1: Deep Learning in Speech Synthesis
17:20 – 17:30	SynSIG Message
17.30 – 18.30	"Jam" Music Session
18.30 – 20.00	Reception at Institut de Estudis Catalans (SSW8 Venue)

Sunday, September 1

08:00	Registration opens
09:00 – 10:15	Oral Session 3: Robustness in synthetic speech
10:15 – 10:45	Coffee Break
10:45 – 12:25	Oral Session 4: Issues in HMM-based speech synthesis
12:30 – 14:00	Lunch Break
14:00 – 14:50	Keynote Session 2: Prosodic Patterns in Dialog
14:50 – 16:30	Poster Session 2 & Coffee
16:30 – 20:00	Guided visit to Sagrada Familia & Park Güell
20:00 – 22:00	Dinner at the Restaurant of the Royal Barcelona Maritim Club

Monday, September 2

08:00	Registration opens
09:00 – 09:50	Keynote Session 3: Singing voice synthesis
09:50 – 10:40	Oral Session 5: Synthetic singing voices
10:40 – 11:10	Coffee Break
11:10 – 12:50	Oral Session 6: Expressive speech synthesis
12:50 – 14:20	Lunch Break
14:20 – 15:20	Demo Session
15:20 – 17:00	Poster Session 3 & Coffee
17:00 – 17:10	Closing

Saturday, August 31

Oral Session 1: Prosody and pausing.

Saturday, August 31, 9:20 – 11:00

Chair: Alan Black

OS1-1 9:20 – 9:45	Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS *Norbert Braunschweiler, Langzhou Chen*	*14*
OS1-2 9:45 – 10:10	Role of Pausing in Text-to-Speech Synthesis for Simultaneous Interpretation *Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Alistair Conkie*	*14*
OS1-3 10:10 – 10:35	Minimum Error Rate Training for Phrasing in Speech Synthesis *Alok Parlikar, Alan Black*	*15*
OS1-4 10:35 – 11:00	HMM-based Speech Synthesis of Live Sports Commentaries: Integration of a Two-Layer Prosody Annotation *Benjamin Picart, Sandrine Brognaux, Thomas Drugman*	*15*

Poster Session 1

Saturday, August 31, 11:00 – 12:40

Chair: Eduardo Rodríguez Banga

PS1-1 11:00 – 12:40	Parametric model for vocal effort interpolation with Harmonics Plus Noise Models *Àngel Calzada Defez, Joan Claudi Socoró Carrié, Robert Clark*	*17*
PS1-2 11:00 – 12:40	Vietnamese HMM-based Speech Synthesis with prosody information *Anh-Tuan Dinh, Thanh-Son Phan, Tat-Thang Vu, Chi Mai Luong*	*18*
PS1-3 11:00 – 12:40	Context labels based on "bunsetsu" for HMM-based speech synthesis of Japanese	*18*

Oral Session 2: Open Challenges in speech synthesis.

Saturday, August 31, 14:45 – 16:00

Chair: Simon King

Keynote Session 1

Saturday, August 31, 16:30 – 17:20

Chair: Keiichi Tokuda

Sunday, September 1

Oral Session 3: Robustness in synthetic speech.

Sunday, September 1, 9:00 – 10:15

Chair: Frank Soong

OS3-1 9:00 – 9:25	A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices *Mauro Nicolao, Fabio Tesser, Roger K. Moore*	*27*
OS3-2 9:25 – 9:50	Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise *Cassia Valentini-Botinhao, Mirjam Wester, Junichi Yamagishi, Simon King*	*28*
OS3-3 9:50 – 10:15	Noise Robustness in HMM-TTS Speaker Adaptation *Kayoko Yanagisawa, Javier Latorre, Vincent Wan, Mark J. F. Gales, Simon King*	*28*

Oral Session 4: Issues in HMM-based speech synthesis.

Sunday, September 1, 10:45 – 12:25

Chair: Tomoki Toda

OS4-1 10:45 – 11:10	New Method for Rapid Vocal Tract Length Adaptation in HMM-based Speech Synthesis *Daniel Erro, Agustin Alonso, Luis Serrano, Eva Navas, Inma Hernaez*	*30*
OS4-2 11:10 – 11:35	Text-to-speech synthesizer based on combination of composite wavelet and hidden Markov models *Nobukatsu Hojo, Kota Yoshizato, Hirokazu Kameoka, Daisuke Saito, Shigeki Sagayama*	*30*
OS4-3 11:35 – 12:00	An experimental comparison of multiple vocoder types *Qiong Hu, Korin Richmond, Junichi Yamagishi, Javier Latorre*	*31*
OS4-4 12:00 – 12:25	Statistical Model Training Technique for Speech Synthesis Based on Speaker Class *Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno*	*32*

Keynote Session 2

Sunday, September 1, 14:00 – 14:50

Chair: Nick Campbell

KN2 14:00 – 14:50 — Prosodic Patterns in Dialog *33*
Nigel Ward

Poster Session 2

Sunday, September 1, 14:50 – 16:30

Chair: Junichi Yamagishi

PS2-1 14:50 – 16:30 — Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech *34*
Florian Hinterleitner, Christoph Norrenbrock, Sebastian Möller

PS2-2 14:50 – 16:30 — Evaluation of contextual descriptors for HMM-based speech synthesis in French *34*
Sébastien Le Maguer, Nelly Barbot, Olivier Boeffard

PS2-3 14:50 – 16:30 — Towards Speaking Style Transplantation in Speech Synthesis *35*
Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Junichi Yamagishi, Oliver Watts, Juan M. Montero

PS2-4 14:50 – 16:30 — Investigating the shortcomings of HMM synthesis *36*
Thomas Merritt, Simon King

PS2-5 14:50 – 16:30 — Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis *36*
Raúl Montaño, Francesc Alías, Josep Ferrer

PS2-6 14:50 – 16:30 — Objective evaluation measures for speaker-adaptive HMM-TTS systems *37*
Ulpu Remes, Reima Karhila, Mikko Kurimo

PS2-7 14:50 – 16:30 — Experiments with Signal-Driven Symbolic Prosody for Statistical Parametric Speech Synthesis *37*

Monday, September 2

Keynote Session 3

Monday, September 2, 9:00 – 9:50

Chair: Asunción Moreno

KN3 9:00 – 9:50 Singing voice synthesis in the context of music technology research *40*
Xavier Serra

Oral Session 5: Synthetic singing voices.

Monday, September 2, 9:50 – 10:40

Chair: Xavier Serra

OS5-1 9:50 – 10:15 Mage - Reactive articulatory feature control of HMM-based parametric speech synthesis *41*
Maria Astrinaki, Alexis Moinet, Junichi Yamagishi, Korin Richmond, Zhen-Hua Ling, Simon King, Thierry Dutoit

OS5-2 10:15 – 10:40 Systematic database creation for expressive singing voice synthesis control *41*
Marti Umbert, Jordi Bonada, Merlijn Blaauw

Oral Session 6: Expressive speech synthesis.

Monday, September 2, 11:10 – 12:50

Chair: Paul Taylor

Demo Session

Monday, September 2, 14:20 – 15:20

Chair: Javier Latorre

Poster Session 3

Monday, September 2, 15:20 – 17:00

Chair: Keikichi Hirose

Jindřich Matoušek, Daniel Tihelka, Milan Legát

PS3-5 15:20 – 17:00	Development of Electrolarynx with Hands-Free Prosody Control *Kenji Matsui, Kenta Kimura, Yoshihisa Nakatoh, Yumiko O. Kato*	*51*
PS3-6 15:20 – 17:00	A Hybrid TTS between Unit Selection and HMM-based TTS under limited data conditions *Trung-Nghia Phung, Chi Mai Luong, Masato Akagi*	*51*
PS3-7 15:20 – 17:00	Wavelets for intonation modeling in HMM speech synthesis *Antti Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, Martti Vainio*	*52*
PS3-8 15:20 – 17:00	A Common Attribute based Unified HTS framework for Speech Synthesis in Indian Languages *Ramani B, S Lilly Christina, G Anushiya Rachel, Sherlin Solomi V, Mahesh Kumar Nandwana, Anusha Prakash, Aswin Shanmugam S, Raghava Krishnan, S Kishore Prahalad, K Samudravijaya, P Vijayalakshmi, T Nagarajan, Hema Murthy*	*53*
PS3-9 15:20 – 17:00	Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for HMM-based speech synthesis *Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda*	*54*
PS3-10 15:20 – 17:00	Residual Compensation based on Articulatory Feature-based Phone Clustering for Hybrid Mandarin Speech Synthesis *Yi-Chin Huang, Chung-Hsien Wu, Shih-Lun Lin*	*54*

Abstracts

Oral Session 1: Prosody and pausing.

Saturday, August 31
Chair: Alan Black
Carnegie Mellon University, United States

OS1-1 **Saturday 9:20 – 9:45**

Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS

Norbert Braunschweiler, Langzhou Chen
Toshiba Research Europe Ltd., United Kingdom

The presence of inhalation breaths in speech pauses has recently attracted more attention especially since the focus of speech synthesis research has shifted to prosodic aspects beyond a single sentence, as, for instance in the synthesis of audiobooks. Inhalation breath pauses are usually not an issue in traditional speech synthesis corpora because they typically use single sentences of limited length and therefore pauses including inhalation breaths rarely occur or they are deliberately avoided during recording. However, in readings of large coherent texts like audiobooks, there are often inhalation breaths, particularly in publicly available audiobooks. These inhalation breaths are relevant for the modelling of pauses in audiobook synthesis and can cause a reduction in naturalness when un-modelled. Therefore this paper presents a method to automatically classify pauses into one of four classes (silent pause, inhalation breath pause, noisy pause, no pause) for improved pause modelling in HMM-TTS.

OS1-2 **Saturday 9:45 – 10:10**

Role of Pausing in Text-to-Speech Synthesis for Simultaneous Interpretation

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Alistair

Conkie
AT&T, United States

The goal of simultaneous speech-to-speech (S2S) translation is to translate source language speech into target language with low latency. While conventional speech-to-speech (S2S) translation systems typically ignore the source language acoustic-prosodic information such as pausing, exploiting such information for simultaneous S2S translation can potentially aid in the chunking of source text into short phrases that can be subsequently translated incrementally with low latency. Such an approach is often used by human interpreters in simultaneous interpretation. In this work we investigate the phenomena of pausing in simultaneous interpretation and study the impact of utilizing such information for target language text-to-speech synthesis in a simultaneous S2S system. On one hand, we superimpose the source language pause information obtained through forced alignment (or decoding) in an isomorphic manner on the target side while on the other hand, we use a classifier to predict the pause information for the target text by exploiting features from the target language, source language or both. We contrast our approach with the baseline that does not use any pauses. We perform our investigation on a simultaneous interpretation corpus of Parliamentary speeches and present subjective evaluation results based on the quality of synthesized target speech.

OS1-3 **Saturday 10:10 – 10:35**

Minimum Error Rate Training for Phrasing in Speech Synthesis

Alok Parlikar, Alan Black
Carnegie Mellon University, United States

Phrase break prediction models in speech synthesis are classifiers that predict whether or not each word boundary is a prosodic break. These classifiers are generally trained to optimize the likelihood of prediction, and their performance is evaluated in terms of classification accuracy. We propose a minimum error rate training method for phrase break prediction. We combine multiple phrasing models into a log-linear framework and optimize the system directly to the quality of break prediction, as measured by the F-measure. We show that this method significantly improves our phrasing models. We also show how this framework allows us to design a knob that can be tweaked to increase or decrease the number of phrase breaks at synthesis time.

OS1-4 **Saturday 10:35 – 11:00**

HMM-based Speech Synthesis of Live Sports Commentaries: Integration of a Two-Layer Prosody Annotation

Benjamin Picart[1], *Sandrine Brognaux*[2], *Thomas Drugman*[1]

[1]Faculté Polytechnique (FPMs) - University of Mons (UMons), Belgium; [2]Cental, ICTEAM - Université Catholique de Louvain, Belgium

This paper proposes the integration of a two-layer prosody annotation specific to live sports commentaries into HMM-based speech synthesis. Local labels are assigned to all syllables and refer to accentual phenomena. Global labels categorize sequences of words into five distinct speaking styles, defined in terms of valence and arousal. Two stages of the synthesis process are analyzed. First, the integration of global labels (i.e. speaking styles) is carried out either using speaker-dependent training or adaptation methods. Secondly, a comprehensive study allows evaluating the effects achieved by each prosody annotation layer on the generated speech. The evaluation process is based on three subjective criteria: intelligibility, expressivity and segmental quality. Our experiments indicate that: (i) for the integration of global labels, adaptation techniques outperform speaking style-dependent models both in terms of intelligibility and segmental quality; (ii) the integration of local labels results in an enhanced expressivity, while it provides slightly higher intelligibility and segmental quality performance; (iii) combining the two levels of annotation (local and global) leads to the best results. It is indeed shown that it obtains better levels of expressivity and intelligibility.

Poster Session 1

Saturday, August 31

Chair: Eduardo Rodríguez Banga
Universidad de Vigo, Spain

PS1-1 **Saturday 11:00 – 12:40**

Parametric model for vocal effort interpolation with Harmonics Plus Noise Models

Àngel Calzada Defez[1], *Joan Claudi Socoró Carrié*[1], *Robert Clark*[2]

[1]La Salle (Universitat Ramon Llull), Spain; [2]University of Edinburgh, United Kingdom

It is known that voice quality plays an important role in expressive speech. In this paper, we present a methodology for modifying vocal effort level, which can be applied by text-to-speech (TTS) systems to provide the flexibility needed to improve the naturalness of synthesized speech. This extends previous work using low order Linear Prediction Coefficients (LPC) where the flexibility was constrained by the amount of vocal effort levels available in the corpora. The proposed methodology overcomes these limitations by replacing the low order LPC by ninth order polynomials to allow not only vocal effort to be modified towards the available templates, but also to allow the generation of intermediate vocal effort levels between levels available in training data. This flexibility comes from the combination of Harmonics plus Noise Models and using a parametric model to represent the spectral envelope. The conducted perceptual tests demonstrate the effectiveness of the proposed technique in performing vocal effort interpolations while maintaining the signal quality in the final synthesis. The proposed technique can be used in unit-selection TTS systems to reduce corpus size while increasing its flexibility, and the techniques could potentially be employed by HMM based speech synthesis systems if appropriate acoustic features are being used.

PS1-2 **Saturday 11:00 – 12:40**

Vietnamese HMM-based Speech Synthesis with prosody information

Anh-Tuan Dinh[1], *Thanh-Son Phan*[2], *Tat-Thang Vu*[1], *Chi Mai Luong*[1]
[1]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam; [2]Faculty of Information Technology, Le Qui Don Technical University, Hanoi, Vietnam

Generating natural-sounding synthetic voice is an aim of all text to speech system. To meet the goal, many prosody features have been used in full-context labels of an HMM-based Vietnamese synthesizer. In the prosody specification, POS and Intonation information are considered not as important as positional information. The paper investigates the impact of POS and Intonation tagging on the naturalness of HMM-based voice. It was discovered that, the POS and Intonation tags help reconstruct the duration and emotion in synthesized voice.

PS1-3 **Saturday 11:00 – 12:40**

Context labels based on "bunsetsu" for HMM-based speech synthesis of Japanese

Hiroya Hashimoto, Keikichi Hirose, Nobuaki Minematsu
The University of Tokyo, Japan

A new set of context labels was developed for HMM-based speech synthesis of Japanese. The conventional labels include those directly related to sentence length, such as number of "mora" and order of breath group in a sentence. When reading a sentence, it is unlikely that we count its total length before utterance. Also a set of increased number of labels is required to handle sentences with various lengths, resulting in a less efficient clustering process. Furthermore, labels related to prosody are mostly designed based on the unit "accent phrase," whose definition is somewhat unclear; it is not uniquely defined for a given sentence, but also is affected by other factors such as speaker identity, speaking rate, and utterance style. Accent phrase boundaries may be labeled differently for utterances of the same content, and this situation affects other labels, because of numerical labeling scheme counted from the sentence/breath-group initial. In the proposed labels, "bunsetsu" is used instead. Also, we only view its relations with preceding and following "bunsetsu's." Thus labels not related to the sentence lengths are obtained, with easier automatic prediction only from sentence representations.

Validity of the proposed labels was shown through speech synthesis experiments.

PS1-4 **Saturday 11:00 – 12:40**

Using Adaptation to Improve Speech Transcription Alignment in Noisy and Reverberant Environments

Yoshitaka Mamiya[1], *Adriana Stan*[2], *Junichi Yamagishi*[1], *Peter Bell*[1], *Oliver Watts*[1], *Robert Clark*[1], *Simon King*[1]

[1]University of Edinburgh, United Kingdom; [2]Communications Department, Technical University of Cluj-Napoca, Romania

When using data retrieved from the internet to create new speech databases, the recording conditions can often be highly variable within and between sessions. This variance influences the overall performance of any automatic speech and text alignment techniques used to process this data. In this paper we discuss the use of speaker adaptation methods to address this issue. Starting from a baseline system for automatic sentence-level segmentation and speech and text alignment based on GMMs and grapheme HMMs, respectively, we employ Maximum A Posteriori (MAP) and Constrained Maximum Likelihood Linear Regression (CMLLR) techniques to model the variation in the data in order to increase the amount of confidently aligned speech. We tested 29 different scenarios, which include reverberation, 8 talker babble noise and white noise, each in various combinations and SNRs. Results show that the MAP-based segmentation's performance is very much influenced by the noise type, as well as the presence or absence of reverberation. On the other hand, the CMLLR adaptation of the acoustic models gives an average 20% increase in the aligned data percentage for the majority of the studied scenarios.

PS1-5 **Saturday 11:00 – 12:40**

Speech synthesis using a maximally decimated pseudo QMF bank for embedded devices

Nobuyuki Nishizawa, Tsuneo Kato

KDDI R&D Laboratories, Inc., Japan

A fast speech waveform generation method using a maximally decimated pseudo quadrature mirror filter (QMF) bank is proposed. The method is based on subband coding with pseudo QMF banks, which is also used in MPEG Audio. In the method, subband code vectors for speech sounds are synthesized from magnitudes

of spectral envelope and fundamental frequencies for periodic frames, and then waveforms are generated by decoding of the vectors. Since the synthesizing of the vectors is performed at the reduced sampling rate by the maximal decimation and the decoding is processed with fast discrete cosine transformation algorithms, faster speech waveform generation is achieved totally. Although pre-encoded vectors for noise components were used to reduce the computational costs in our former studies, in this study, all code vectors for noise components are made with a noise generator at run time for small footprint systems. In contrast, a subjective test for synthetic sounds by HMM-based speech synthesis using mel-cepstrum showed the proposed method was comparable to our former method and also the conventional method using a mel log spectrum approximation (MLSA) filter in quality of sounds.

PS1-6 **Saturday 11:00 – 12:40**

HMM-based sCost quality control for unit selection speech synthesis

Sathish Pammi[1]*, Marcela Charfuelan*[2]
[1]ISIR, Universit Pierre et Marie CURIE (UPMC), France; [2]DFKI GmbH, Germany

This paper describes the implementation of a unit selection text-to-speech system that incorporates a statistical model Cost (sCost), in addition to target and join costs, for controlling the selection of unit candidates. sCost, a quality control measure, is calculated off-line for each unit by comparing HMM based synthesis and recorded speech with their corresponding unit segment labels. Dynamic time warping (DTW) is used to perform such comparison at level of spectrum, pitch and voice strengths. The method has been tested on unit selection voices created using audio book data. Preliminary results indicate that the use of sCost based only on spectrum introduce more variety on style pronunciation but affects quality; whereas using sCost based on spectrum, pitch and voicing strengths improves significantly the quality, maintaining a more stable narrative style.

PS1-7 **Saturday 11:00 – 12:40**

Understanding Factors in Emotion Perception

Lakshmi Saheer, Blaise Potard
Idiap Research Institute, Switzerland

Emotion in speech is an important and challenging research area. Addition or

understanding of emotions from speech is challenging. But, an equally difficult task is to identify the intended emotion from an audio or speech. Understanding emotions is important not only in itself as a research area, but also, for adding emotions to synthesised speech. Evaluating synthesised speech with emotions can be simplified if the correct factors in emotion perception can be first identified. To this end, this work explores various factors that could influence the perception of emotions. These factors include semantic information of the text, contextual information, language understanding and knowledge. This work also investigates the right framework for a subjective perceptual evaluation by providing different options to the listeners and checking which are the most effective response to evaluate the perception of the emotion.

PS1-8 **Saturday 11:00 – 12:40**

Multilingual Number Transcription for Text-to-Speech Conversion

Rubén San-Segundo[1], *Juan Manuel Montero*[1], *Mircea Giurgiu*[2], *Ioana Muresan*[2], *Simon King*[3]

[1]Speech Technology Group, ETSI Telecomunicación, UPM, Spain; [2]Dept. of Telecommun., Tech. Univ. of Cluj-Napoca, Cluj-Napoca, Romania., Romania; [3]University of Edinburgh, United Kingdom

This paper describes the text normalization module of a text to speech fully-trainable conversion system and its application to number transcription. The main target is to generate a language independent text normalization module, based on data instead of on expert rules. This paper proposes a general architecture based on statistical machine translation techniques. This proposal is composed of three main modules: a tokenizer for splitting the text input into a token graph, a phrase-based translation module for token translation, and a post-processing module for removing some tokens. This architecture has been evaluated for number transcription in several languages: English, Spanish and Romanian. Number transcription is an important aspect in the text normalization problem.

PS1-9 **Saturday 11:00 – 12:40**

Noise-Robust Voice Conversion Based on Spectral Mapping on

Sparse Space

Ryoichi Takashima, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki
Kobe University, Japan

This paper presents a voice conversion (VC) technique for noisy environments based on a sparse representation of speech. In our previous work, we discussed an exemplar-based VC technique for noisy environments. In that report, source exemplars and target exemplars are extracted from the parallel training data, having the same texts uttered by the source and target speakers. The input source signal is represented using the source exemplars and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. However, this exemplar-based approach needs to hold all training exemplars (frames) and it requires high computation times to obtain the weights of the source exemplars. In this paper, we propose a framework to train the basis matrices of source and target exemplars so that they have a common weight matrix. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. The effectiveness of this method was confirmed by comparing its effectiveness, in speaker conversion experiments using noise-added speech data, with the effectiveness of an exemplar-based method and a conventional Gaussian mixture model (GMM)-based method.

PS1-10 **Saturday 11:00 – 12:40**

Cross-variety speaker transformation in HSMM-based speech synthesis

Markus Toman, Michael Pucher, Dietmar Schabus
Telecommunications Research Center (FTW), Vienna, Austria

We present and compare different approaches for cross-variety speaker transformation in Hidden Semi-Markov Model (HSMM) based speech synthesis that allow for a transformation of an arbitrary speaker's voice from one variety to another one. The methods developed are applied to three different varieties, namely standard Austrian German, one Middle Bavarian (Upper Austria, Bad Goisern) and one South Bavarian (East Tyrol, Innervillgraten) dialect. For data mapping of HSMM-states we use Kullback-Leibler divergence, transfer probability density functions to the decision tree of the other variety and perform speaker adaptation. We investigate an existing data mapping method and a method that constrains the mappings for common phones and show that both methods can retain speaker

similarity and variety similarity. Furthermore we show that in some cases the constrained mapping method gives better results than the standard method.

PS1-11 **Saturday 11:00 – 12:40**

Multi-variety adaptive acoustic modeling in HSMM-based speech synthesis

Markus Toman, Michael Pucher, Dietmar Schabus

Telecommunications Research Center (FTW), Vienna, Austria

In this paper we apply adaptive modeling methods in Hidden Semi-Markov Model (HSMM) based speech synthesis to the modeling of three different varieties, namely standard Austrian German, one Middle Bavarian (Upper Austria, Bad Goisern), and one South Bavarian (East Tyrol, Innervillgraten) dialect. We investigate different adaptation methods like dialect-adaptive training and dialect clustering that can exploit the common phone sets of dialects and standard, as well as speaker-dependent modeling. We show that most adaptive and speaker-dependent methods achieve a good score on overall (speaker and variety) similarity. Concerning overall quality there is no significant difference between adaptive methods and speaker-dependent methods in general for the present data set.

Oral Session 2: Open Challenges in speech synthesis.

Saturday, August 31
Chair: Simon King
University of Edinburgh, UK

OS2-1 **Saturday 14:45 – 15:10**

Investigation of intra-speaker spectral parameter variation and its prediction towards improvement of spectral conversion metric

Tatsuo Inukai, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Nara Institute of Science and Technology, Japan

In statistical voice conversion, distance measure between the converted and target spectral parameters are often used as evalu-ation/training metrics. However, even if same speaker utters the same sentence several times, the spectral parameters of those utterances vary, and therefore, a spectral distance between them still exists. Moreover during real-time conversion procedure, converted speech keeping original prosodic features of input speech is often generated because converting prosodic feature with complex method is essentially difficult. In such a case, an ideal sample of converted speech will be a utterance uttered by a target speaker imitating prosody of the input speech. How-ever a spectral variation caused by such a prosodic change is not considered in the current evaluation/training metrics. In this study, we investigate an intra-speaker spectral variation between utterances of the same sentence focusing on mel-cepstral coeffi-cients as a spectral parameter. Moreover, we propose a method for predicting it from prosodic parameter differences between those utterances and conduct experimental evaluations to show its effectiveness.

OS2-2 **Saturday 15:10 – 15:35**

Text to Speech in New Languages without a Standardized Orthography

Sunayana Sitaram, Gopala Anumanchipalli, Justin Chiu, Alok Parlikar, Alan Black

Carnegie Mellon University, United States

Many spoken languages do not have a standardized writing system. Building text to speech voices for them, without accurate transcripts of speech data is difficult. Our language independent method to bootstrap synthetic voices using only speech data relies upon crosslingual phonetic decoding of speech. In this paper, we describe novel additions to our bootstrapping method. We present results on eight different languages—English, Dari, Pashto, Iraqi, Thai, Konkani, Inupiaq and Ojibwe, from different language families and show that our phonetic voices can be made understandable with as little as an hour of speech data that never had transcriptions, and without many resources in the target language available. We also present purely acoustic techniques that can help induce syllable and word level information that can further improve the intelligibility of these voices.

OS2-3 **Saturday 15:35 – 16:00**

Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis

Oliver Watts[1]*, Adriana Stan*[2]*, Rob Clark*[1]*, Yoshitaka Mamiya*[1]*, Mircea Giurgiu*[2]*, Junichi Yamagishi*[1]*, Simon King*[1]

[1]University of Edinburgh, United Kingdom; [2]Technical University of Cluj-Napoca, Romania

This paper presents techniques for building text-to-speech front-ends in a way that avoids the need for language-specific expert knowledge, but instead relies on universal resources (such as the Unicode character database) and unsupervised learning from unannotated data to ease system development. The acquisition of expert language-specific knowledge and expert annotated data is a major bottleneck in the development of corpus-based TTS systems in new languages. The methods presented here side-step the need for such resources as pronunciation lexicons, phonetic feature sets, part of speech tagged data, etc. The paper explains how the techniques introduced are applied to the 14 languages of a corpus of 'found' audiobook data. Results of an evaluation of the intelligibility of the systems resulting from applying these novel techniques to this data are presented.

Keynote Session 1

Saturday, August 31
Chair: Keiichi Tokuda
Nagoya Institute of Technology, Japan

Saturday 16:30 – 17:20

Deep Learning in Speech Synthesis

Heiga Zen
Google, UK

Deep learning has been a hot research topic in various machine learning related areas including general object recognition and automatic speech recognition. This talk will present recent applications of deep learning to statistical parametric speech synthesis and contrast the deep learning-based approaches to the existing hidden Markov model-based one.

Oral Session 3: Robustness in synthetic speech.

Sunday, September 1

Chair: Frank Soong
Microsoft Research Asia, China

OS3-1 **Sunday 9:00 – 9:25**

A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices

Mauro Nicolao[1], *Fabio Tesser*[2], *Roger K. Moore*[1]

[1]University of Sheffield, United Kingdom; [2]ISTC-CNR, Italy

The effectiveness of phonetic-contrast motivated adaptation on HMM-based synthetic voices was previously tested on English successfully. The aim of this paper is to prove that such adaptation can be exported with minor changes to languages having different intrinsic characteristics. The Italian language was chosen because it has no obvious phonemic configuration towards which human speech tend when hypo-articulated such as the mid-central vowel (schwa) for English. Nonetheless, low-contrastive attractors were identified and a linear transformation was trained by contrasting each phone pronunciation with its nearest acoustic neighbour. Different degree of hyper and hypo articulated synthetic speech was then achieved by scaling such adaptation along the dimension identified by each contrastive pair. The Italian synthesiser outcome adapted with both the maximum and the minimum transformation strength was evaluated with two objective assessments: the analysis of some common acoustic correlates and the measurement of a intelligibility-in-noise index. For the latter, signals were mixed with different disturbances at various energy ratios and intelligibility was compared to the standard-TTS generated speech. The experimental results proved such transformation on the Italian voices to be as effective as those on the English one.

OS3-2 **Sunday 9:25 – 9:50**

Using neighbourhood density and selective SNR boosting to increase the intelligibility of synthetic speech in noise

Cassia Valentini-Botinhao, Mirjam Wester, Junichi Yamagishi, Simon King
University of Edinburgh, United Kingdom

Motivated by the fact that words are not equally confusable, we explore the idea of using word-level intelligibility predictions to selectively boost the harder-to-understand words in a sentence, aiming to improve overall intelligibility in the presence of noise. First, the intelligibility of a set of words from dense and sparse phonetic neighbourhoods was evaluated in isolation. The resulting intelligibility scores were used to inform two sentence-level experiments. In the first experiment the signal-to-noise ratio of one word was boosted to the detriment of another word. Sentence intelligibility did not generally improve. The intelligibility of words in isolation and in a sentence were found to be significantly different, both in clean and in noisy conditions. For the second experiment, one word was selectively boosted while slightly attenuating all other words in the sentence. This strategy was successful for words that were poorly recognised in that particular context. However, a reliable predictor of word-in-context intelligibility remains elusive, since this involves - as our results indicate - semantic, syntactic and acoustic information about the word and the sentence.

OS3-3 **Sunday 9:50 – 10:15**

Noise Robustness in HMM-TTS Speaker Adaptation

Kayoko Yanagisawa, Javier Latorre, Vincent Wan, Mark J. F. Gales, Simon King
Toshiba Research Europe Ltd., United Kingdom

Speaker adaptation for TTS applications has been receiving more attention in recent years for applications such as voice customisation or voice banking. If these applications are offered as an internet service, there is no control on the quality of the data that can be collected. It can be noisy with people talking in the background or recorded in a reverberant environment. This makes the adaptation more difficult. This paper explores the effect of different levels of additive and convolutional noise on speaker adaptation techniques based on cluster adaptive training (CAT) and average voice model (AVM). The results indicate that although both techniques suffer degradation to some extent, CAT is in general

more robust than AVM.

Oral Session 4: Issues in HMM-based speech synthesis.

Sunday, September 1
Chair: Tomoki Toda
Nara Institute of Science and Technology, Japan

OS4-1 **Sunday 10:45 – 11:10**

New Method for Rapid Vocal Tract Length Adaptation in HMM-based Speech Synthesis

Daniel Erro[1], *Agustin Alonso*[2], *Luis Serrano*[2], *Eva Navas*[2], *Inma Hernaez*[2]
[1]Ikerbasque - UPV/EHU, Spain; [2]University of the Basque Country (UPV/EHU), Spain

We present a new method to rapidly adapt the models of a statistical synthesizer to the voice of a new speaker. We apply a relatively simple linear transform that consists of a vocal tract length normalization (VTLN) part and a long-term average cepstral correction part. Despite the logical limitations of this approach, we will show that it effectively reduces the gap between source and target voices with only one reference utterance and without any phonetic segmentation. In addition, by using a minimum generation error criterion we avoid some of the problems that have been reported to arise when using a maximum likelihood criterion in VTLN.

OS4-2 **Sunday 11:10 – 11:35**

Text-to-speech synthesizer based on combination of composite wavelet and hidden Markov models

Nobukatsu Hojo[1], *Kota Yoshizato*[1], *Hirokazu Kameoka*[1,2], *Daisuke Saito*[1], *Shigeki Sagayama*[1]
[1]Graduate School of Information Science and Technology, The University

of Tokyo, Japan; [2]Communication Science Laboratories, NTT Corporation, Japan

This paper proposes a text-to-speech synthesis (TTS) system based on a combined model of the Composite Wavelet Model (CWM) and Hidden Markov Model (HMM). Conventional HMM-based TTS systems using cepstral features tend to produce over-smoothed spectra, which often result in muffled and buzzy synthesized speech. This is simply caused by the averaging of spectra associated with each phoneme during the learning process. To avoid the over-smoothing of generated spectra, we consider it important to focus on a different representation of the generative process of speech spectra. In particular, we choose to characterize speech spectra by the CWM, whose parameters correspond to the frequency, gain and peakiness of each underlying formant. This idea is motivated by our expectation that averaging of these parameters would not directly cause the over-smoothing of spectra, as opposed to the cepstral representations. To describe the entire generative process of a sequence of speech spectra, we combine the generative process of a formant trajectory using an HMM and the generative process of a speech spectrum using the CWM. A parameter learning algorithm for this combined model is derived based on an auxiliary function approach. We confirmed through experiments that our speech synthesis system was able to generate speech spectra with clear peaks and dips, which resulted in natural-sounding synthetic speech.

OS4-3 **Sunday 11:35 – 12:00**

An experimental comparison of multiple vocoder types

Qiong Hu[1]*, Korin Richmond*[1]*, Junichi Yamagishi*[1]*, Javier Latorre*[2]
[1]University of Edinburgh, United Kingdom; [2]Toshiba Research Europe, United Kingdom

This paper presents an experimental comparison of a broad range of the leading vocoder types which have been previously described. We use a reference implementation of each of these to create stimuli for a listening test using copy synthesis. The listening test is performed using both Lombard and normal read speech stimuli, and with two types of question for comparison. Multi-dimensional Scaling (MDS) is conducted on the listener responses to analyse similarities in terms of quality between the vocoders. Our MDS and clustering results show that the vocoders which use a sinusoidal synthesis approach are perceptually distinguishable from the source-filter vocoders. To help further interpret the axes of the resulting MDS space, we test for correlations with standard acoustic quality met-

rics and find one axis is strongly correlated with PESQ scores. We also find both speech style and the format of the listening test question may influence test results. Finally, we also present preference test results which compare each vocoder with the natural speech.

OS4-4 **Sunday 12:00 – 12:25**

Statistical Model Training Technique for Speech Synthesis Based on Speaker Class

Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno
NTT Corporation, Japan

To allow the average-voice-based speech synthesis technique to generate synthetic speech that is more similar to that of the target speaker, we propose a model training technique that introduces the label of speaker class. Speaker class represents the voice characteristics of speakers. In the proposed technique, first, all training data are clustered to determine classes of speaker type. The average voice model is trained using the labels of conventional context and speaker class. In the speaker adaptation process, the target speaker's class is estimated and is used to transform the average voice model into the target speaker's model. As a result, the speech of the target speaker is synthesized from the target speaker's model and the estimated target speaker's speaker class. The results of an objective experiment show that the proposed technique significantly reduces the RMS errors of log F0. Moreover, the results of a subjective experiment indicate that the proposal yields synthesized speech with better similarity than the conventional method.

Keynote Session 2

Sunday, September 1
Chair: Nick Campbell
University of Dublin, Ireland

Sunday 14:00 – 14:50

Prosodic Patterns in Dialog

Nigel Ward
Computer Science, University of Texas at El Paso, United States

In human-human dialog, over 80% of the variance in prosody can be explained by just 20 prosodic patterns, most of which involve actions of both speakers and most of which last several seconds. In dialog these patterns frequently occur simultaneously, at varying offsets, and they are additive at the signal level and apparently compositional at the semantic/pragmatic level. These patterns provide a simple, non-structural way to model the prosodic implications of various functions important in dialog, including managing turn-taking, framing topic structure, grounding, expressing attitude, and conveying instantaneous cognitive state, among others. These patterns have been used for language modeling, for detecting important moments in the speech stream, and for information retrieval from audio archives, and may be useful for speech synthesis for dialog applications.

Poster Session 2

Sunday, September 1

Chair: Junichi Yamagishi
University of Edinburgh, UK / National Institute of Informatics, Tokyo, Japan

PS2-1 **Sunday 14:50 – 16:30**

Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech

Florian Hinterleitner[1], *Christoph Norrenbrock*[2], *Sebastian Möller*[1]

[1]TU Berlin, Germany; [2]CAU Kiel, Germany

In this paper, we present a comparative overview of 9 studies on perceptual quality dimensions of synthetic speech. Differ-ent subjective assessment techniques have been used to evalu-ate the text-to-speech (TTS) stimuli in each of these tests: in a semantic differential, the test participants rate every stimulus on a given set of rating scales, while in a paired comparison test, the subjects rate the similarity of pairs of stimuli. Percep-tual quality dimensions can be derived from the results of both test methods, either by performing a factor analysis or via mul-tidimensional scaling. We show that even though the 9 tests differ in terms of used synthesizer types, stimulus duration, lan-guage, and quality assessment methods, the resulting perceptual quality dimensions can be linked to 5 universal quality dimen-sions of synthetic speech: (i) naturalness of voice, (ii) prosodic quality, (iii) fluency and intelligibility, (iv) disturbances, and (v) calmness.

PS2-2 **Sunday 14:50 – 16:30**

Evaluation of contextual descriptors for HMM-based speech syn-

thesis in French

Sébastien Le Maguer, Nelly Barbot, Olivier Boeffard
IRISA/University of Rennes 1, France

In HTS, a HMM-based speech synthesis system, about fifty contextual factors are introduced to label a segment to synthesize English utterances. Published studies indicate that most of them are used for clustering the prosodic component of speech. Nevertheless, the influence of all these factors on modeling is still unclear for French. The work presented in this paper deals with the analysis of contextual factors on acoustic parameters modeling in the context of a French synthesis purpose. Two objective and one subjective methodologies of evaluation are carried out to conduct this study. The first one relies on a GMM-approach to achieve a global evaluation of the synthetic acoustic space. The second one is based on a pairwise distance determined according to the acoustic parameter evaluated. Finally, a subjective evaluation is conducted to complete this study. Experimental results show that using phonetic context improves the overall spectrum and duration modeling and using syllable informations improves the F0 modeling. However other contextual factors do not significantly improve the quality of the HTS models.

PS2-3 **Sunday 14:50 – 16:30**

Towards Speaking Style Transplantation in Speech Synthesis

Jaime Lorenzo-Trueba[1]*, Roberto Barra-Chicote*[1]*, Junichi Yamagishi*[2]*, Oliver Watts*[2]*, Juan M. Montero*[1]
[1]Speech Technology Group, ETSI Telecomunicación, UPM, Spain; [2]University of Edinburgh, United Kingdom

One of the biggest challenges in speech synthesis is the production of naturally sounding synthetic voices. This means that the resulting voice must be not only of high enough quality but also that it must be able to capture the natural expressiveness imbued in human speech. This paper focus on solving the expressiveness problem by proposing a set of different techniques that could be used for extrapolating the expressiveness of proven high quality expressive models into neutral speakers in HMM-based synthesis. As an additional advantage, the proposed techniques are based on adaptation approaches, which means that they can be used with little training data (around 15 minutes of training data are used in each style for this paper). For the final implementation, a set of 4 speaking styles were considered: news broadcasts, live sports commentary, interviews and political speech. Finally, the implementation of the 5 techniques were tested through a perceptual

evaluation that proves that the deviations between neutral and expressive average models can be learned and used to imbue expressiveness into target neutral speakers as intended.

PS2-4 **Sunday 14:50 – 16:30**

Investigating the shortcomings of HMM synthesis

Thomas Merritt, Simon King
University of Edinburgh, United Kingdom

This paper presents the beginnings of a framework for formal testing of the causes of the current limited quality of HMM (Hidden Markov Model) speech synthesis. This framework separates each of the effects of modelling to observe their independent effects on vocoded speech parameters in order to address the issues that are restricting the progression to highly intelligible and natural-sounding speech synthesis. The simulated HMM synthesis conditions are performed on spectral speech parameters and tested via a pairwise listening test, asking listeners to perform a "same or different" judgement on the quality of the synthesised speech produced between these conditions. These responses are then processed using multidimensional scaling to identify the qualities in modelled speech that listeners are attending to and thus forms the basis of why they are distinguishable from natural speech. The future improvements to be made to the framework will finally be discussed which include the extension to more of the parameters modelled during speech synthesis.

PS2-5 **Sunday 14:50 – 16:30**

Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis

Raúl Montaño, Francesc Alías, Josep Ferrer
La Salle (Universitat Ramon Llull), Spain

The generation of synthetic speech with a certain degree of expressiveness has been successful for some particular applications or speaking styles (e.g. emotions). In this context, there is a particular speaking style with subtle speech nuances that may be of great interest for delivering expressive speech: the storytelling style. The purpose of this paper is to define a first step towards developing a storytelling Text-to-Speech (TTS) synthesis system by means of modelling the specific prosodic patterns (pitch, intensity and tempo) of this speaking style. We base our

analysis of a tale in Spanish on discourse modes present in storytelling: narrative, descriptive and dialogue. Moreover, we introduce narrative situations (neutral narrative, post-character, decreasing suspense and affective situations) within the narrative mode, which are analysed at the sentence level. After grouping the sentences into modes and narrative situations, we analyse their corresponding prosodic patterns both objectively (via statistical tests) and subjectively (via perceptual test considering resynthesized sentences). The results show that the statistically validated prosodic rules perform equally (or even better) than the original prosody in most sentences.

PS2-6 **Sunday 14:50 – 16:30**

Objective evaluation measures for speaker-adaptive HMM-TTS systems

Ulpu Remes, Reima Karhila, Mikko Kurimo

Aalto University School of Electrical Engineering, Finland

This paper investigates using objective quality measures to evaluate speaker adaptation performance in HMM-based speech synthesis. We compare severel objective measures to subjective evalution results from our earlier work about 1) comparison of speaker adaptation methods for child voices and 2) effects of noise in speaker adaptation. The results analysed in this work indicate a reasonable correlation between several objective and subjective quality measures.

PS2-7 **Sunday 14:50 – 16:30**

Experiments with Signal-Driven Symbolic Prosody for Statistical Parametric Speech Synthesis

Fabio Tesser[1]*, Giacomo Sommavilla*[1]*, Giulio Paci*[1]*, Piero Cosi*[2]

[1]ISTC-CNR, Italy; [2]ISTC-SPFD CNR, Italy

This paper presents a preliminary study on the use of symbolic prosody extracted from the speech signal to improve parameters prediction on HMM-based speech synthesis. The relationship between the prosodic labelling and the actual prosody of the training data is usually ignored in the building phase of corpus based TTS voices. In this work, different systems have been trained using prosodic labels predicted from speech and compared with the conventional system that predicts those labels solely from text. Experiments have been done using data from two speakers (one male and one female). Objective evaluation performed on a test set

of the corpora shows that the proposed systems improve the prediction accuracy of phonemes duration and F0 trajectories. Advantages on the use of signal-driven symbolic prosody in place of the conventional text-driven symbolic prosody, and future works about the effective use of these information in the synthesis stage of a Text To Speech systems are also described.

PS2-8 **Sunday 14:50 – 16:30**

Significance of word-terminal syllables for prediction of phrase breaks in Text-to-Speech systems for Indian languages

Anandaswarup Vadapalli, Peri Bhaskararao, Kishore Prahallad

International Institute of Information Technology Hyderabad, India

Phrase break prediction is very important for speech synthesis. Traditional methods of phrase break prediction have used linguistic resources like part-of-speech (POS) sequence information for modeling these breaks. In the context of Indian languages, we propose to look at syllable level features and explore the use of word-terminal syllables to model phrase breaks. We hypothesize that these terminal syllables serve to discriminate words based syntactic meaning, and can therefore be used to model phrase breaks. We utilize these terminal syllables in building models for automatic phrase break prediction from text and demonstrate by means of objective and subjective measures that these models perform as well as traditional models using POS sequence information. Thus the proposed method avoids the need for POS taggers for prosodic phrasing in Indian languages.

PS2-9 **Sunday 14:50 – 16:30**

The Effect of Age and Native Speaker Status on Synthetic Speech Intelligibility

Catherine Watson, Wei Liu, Bruce MacDonald

University of Auckland, New Zealand

We investigate whether listener age or native speaker status has the biggest impact on the intelligibility of a synthetic New Zealand English voice. The paper presents findings from a speech intelligibility experiment based on a reminding task involving 67 participants. There were no significant differences in the results due to age (young and old adults), however there was for native speaker status. The non- native listeners performed significantly worse than the native listeners in the synthetic speech condition although no differences were found in the natural

speech condition. We argue that despite the fact that aging impacts on speech perception, the older native listeners were able to draw on their in depth language model to help them parse the synthetic speech. The non-native speakers do not have such an in depth model to assist them.

PS2-10 **Sunday 14:50 – 16:30**

Exemplar-Based Voice Conversion using Non-Negative Spectrogram Deconvolution

Zhizheng Wu[1], *Tuomas Virtanen*[2], *Tomi Kinnunen*[3], *Eng Siong Chng*[1], *Haizhou Li*[4]

[1]Nanyang Technological University, Singapore, Singapore; [2]Tampere University of Technology, Finland, Finland; [3]University of Eastern Finland, Finland, Finland; [4]Institute for Infocomm Research, Singapore, Singapore

In the traditional voice conversion, converted speech is generated using statistical parametric models (for example Gaussian mixture model) whose parameters are estimated from parallel training utterances. A well-known problem of the statistical parametric methods is that statistical average in parameter estimation results in the over-smoothing of the speech parameter trajectories, and thus leads to low conversion quality. Inspired by recent success of so-called exemplar-based methods in robust speech recognition, we propose a voice conversion system based on non-negative spectrogram deconvolution with similar ideas. Exemplars, which are able to capture temporal context, are employed to generate converted speech spectrogram convolutely. The exemplar-based approach is seen as a data-driven, non-parametric approach as an alternative to the traditional parametric approaches to voice conversion. Experiments on VOICES database indicate that the proposed method outperforms the conventional joint density Gaussian mixture model by a wide margin in terms of both objective and subjective evaluations.

Keynote Session 3

Monday, September 2
Chair: Asunción Moreno
Universitat Politècnica de Catalunya, Barcelona, Spain

Monday 9:00 – 9:50

Singing voice synthesis in the context of music technology research

Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

The synthesis of the singing voice has always been very much tied to speech synthesis. Since the initial work of Max Mathews with Kelly and Lochbaum at Bell Labs in the 1950s many engineers and musicians have explored the potential of speech processing techniques in music applications. After reviewing some of this history I will present the work done in my research group to develop synthesis engines that could sound as natural and expressive as a real singer, or choir, and whose inputs could be just the score and the lyrics of the song. Some of this research is being done in collaboration with Yamaha and has resulted in the Vocaloid software synthesizer. In the talk I want to make special emphasis on the specificities of the music context and thus on the technical requirements needed for the use of a synthesis technology in music applications.

Oral Session 5: Synthetic singing voices.

Monday, September 2

Chair: Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona Spain

OS5-1 **Monday 9:50 – 10:15**

Mage - Reactive articulatory feature control of HMM-based parametric speech synthesis

Maria Astrinaki[1]*, Alexis Moinet*[1]*, Junichi Yamagishi*[2,3]*, Korin Richmond*[3]*, Zhen-Hua Ling*[4]*, Simon King*[3]*, Thierry Dutoit*[1]

[1]TCTS Lab., Numediart Institute, University of Mons, Belgium; [2]National Institute of Informatics, Tokyo, Japan; [3]University of Edinburgh, United Kingdom; [4]University of Science and Technology of China (USTC), China

In this paper, we present the integration of articulatory control into MAGE, a framework for realtime and interactive (reactive) parametric speech synthesis using hidden Markov models (HMMs). MAGE is based on the speech synthesis engine from HTS and uses acoustic features (spectrum and f0) to model and synthesize speech. In this work, we replace the standard acoustic models with models combining acoustic and articulatory features, such as tongue, lips and jaw positions. We then use feature-space-switched articulatory-to-acoustic regression matrices to enable us to control the spectral acoustic features by manipulating the articulatory features. Combining this synthesis model with MAGE allows us to interactively and intuitively modify phones synthesized in real time, for example transforming one phone into another, by controlling the configuration of the articulators in a visual display.

OS5-2 **Monday 10:15 – 10:40**

Systematic database creation for expressive singing voice synthesis control

Marti Umbert, Jordi Bonada, Merlijn Blaauw

Music Technology Group, Universitat Pompeu Fabra, Spain

In the context of singing voice synthesis, the generation of the synthesizer controls is a key aspect to obtain expressive performances. In our case, we use a system that selects, transforms and concatenates units of short melodic contours from a recorded database. This paper proposes a systematic procedure for the creation of such database. The aim is to cover relevant style-dependent combinations of features such as note duration, pitch interval and note strength. The higher the percentage of covered combinations is, the less transformed the units will be in order to match a target score. At the same time, it is also important that units are musically meaningful according to the target style. In order to create a style-dependent database, the melodic combinations of features to cover are identified, statistically modeled and grouped by similarity. Then, short melodic exercises of four measures are created following a dynamic programming algorithm. The Viterbi cost functions deal with the statistically observed context transitions, harmony, position within the measure and readability. The final systematic score database is formed by the sequence of the obtained melodic exercises.

Oral Session 6: Expressive speech synthesis.

Monday, September 2

Chair: Paul Taylor
Google, UK

OS6-1 **Monday 11:10 – 11:35**

Expressive Speech Synthesis: Synthesising Ambiguity

Matthew Aylett[1,2]*, Blaise Potard*[2]*, Christopher Pidcock*[2]

[1]University of Edinburgh, United Kingdom; [2]CereProc Ltd, United Kingdom

Previous work in HCI has shown that ambiguity, normally avoided in interaction design, can contribute to a user's engagement by increasing interest and uncertainty. In this work, we create and evaluate synthetic utterances where there is a conflict between text content, and the emotion in the voice. We show that: 1) text content measurably alters the negative/positive perception of a spoken utterance, 2) changes in voice quality also produce this effect, 3) when the voice quality and text content are conflicting the result is a synthesised ambiguous utterance. Results were analysed using an evaluation/activation space. Whereas the effect of text content was restricted to the negative/positive dimension (valence), voice quality also had a significant effect on how active or passive the utterance was perceived (activation).

OS6-2 **Monday 11:35 – 12:00**

Interactional Adequacy as a Factor in the Perception of Synthe-

sized Speech

Timo Baumann[1]*, David Schlangen*[2]

[1]Universität Hamburg, Germany; [2]Bielefeld University, Germany

Speaking as part of a conversation is different from reading out aloud. Speech synthesis systems, however, are typically developed using assumptions (at least implicitly) that are more true of the latter than the former situation. We address one particular aspect, which is the assumption that a fully formulated sentence is available for synthesis. We have built a system that does not make this assumption but rather can synthesize speech given incrementally extended input. In an evaluation experiment, we found that in a dynamic domain where what is talked about changes quickly, subjects rated the output of this system as more naturally pronounced than that of a baseline system that employed standard synthesis, despite the quality objectively being degraded. Our results highlight the importance of considering a synthesizer's ability to support interactive use-cases when determining the adequacy of synthesized speech.

OS6-3 **Monday 12:00 – 12:25**

A novel irregular voice model for HMM-based speech synthesis

Tamás Gábor Csapó, Géza Németh

Department of Telecommunications & Media Informatics, Budapest University of Technology & Economics, Hungary

State-of-the-art text-to-speech (TTS) synthesis is often based on statistical parametric methods. Particular attention is paid to hidden Markov model (HMM) based text-to-speech synthesis. HMM-TTS is optimized for ideal voices and may not produce high quality synthesized speech with voices having frequent non-ideal phonation. Such a voice quality is irregular phonation (also called as glottalization), which occurs frequently among healthy speakers. There are existing methods for transforming regular (also called as modal) to irregular voice, but only initial experiments have been conducted for statistical parametric speech synthesis with a glottalization model. In this paper we extend our previous residual codebook based excitation model with irregular voice modeling. The proposed model applies three heuristics, which were proven to be useful: 1) pitch halving, 2) pitch-synchronous residual modulation with periods multiplied by random scaling factors and 3) spectral distortion. In a perception test the extended HMM-TTS produced speech that is more similar to the original speaker than the baseline system. An acoustic experiment found the output of the model to be similar to

original irregular speech in terms of several parameters. Applications of the model may include expressive statistical parametric speech synthesis and the creation of personalized voices.

OS6-4 **Monday 12:25 – 12:50**

Expression of Speaker's Intentions through Sentence-Final Particle/Intonation Combinations in Japanese Conversational Speech Synthesis

Kazuhiko Iwata, Tetsunori Kobayashi
Waseda University, Japan

Aiming to provide the synthetic speech with the ability to express speaker's intentions and subtle nuances, we investigated the relationship between the speaker's intentions that the listener perceived and sentence-final particle/intonation combinations in Japanese conversational speech. First, we classified F0 contours of sentence-final syllables in actual speech and found various distinctive contours, namely, not only simple rising and falling ones but also rise-and-fall and fall-and-rise ones. Next, we conducted subjective evaluations to clarify what kind of intentions the listeners perceived depending on the sentence-final particle/intonation combinations. Results showed that adequate sentence-final particle/intonation combinations should be used to convey the intention to the listeners precisely. Whether the sentence was positive or negative also affected the listeners' perception. For example, a sentence-final particle 'yo' with a falling intonation conveyed the intention of an "order" in a positive sentence but "blame" in a negative sentence. Furthermore, it was found that some specific nuances could be added to some major intentions by subtle differences in intonation. The different intentions and nuances could be conveyed just by controlling the sentence-final intonation in synthetic speech.

Demo Session

Monday, September 2
Chair: Javier Latorre
Toshiba Research Europe Ltd.;United Kingdom

DS-1 **Monday 14:20 – 15:20**

Unified numerical simulation of the physics of voice. The EUNISON project.

Oriol Guasch[1], *Sten Ternström*[2], *Marc Arnela*[1], *Francesc Alías*[1]
[1]La Salle, Universitat Ramon Llull, Barcelona, Catalonia, Spain; [2]Department of Speech, Music and Hearing, Kungliga Tekniska Hgskolan, Stockholm, Sweden

In this demo we will briefly outline the scope of the european EUNISON project, which aims at a unified numerical simulation of the physics of voice by resorting to supercomputer facilities, and present some of its preliminary results obtained to date.

DS-2 **Monday 14:20 – 15:20**

Mage - HMM-based speech synthesis reactively controlled by the articulators

Maria Astrinaki[1], *Alexis Moinet*[1], *Junichi Yamagishi*[2,3], *Korin Richmond*[2], *Zhen-Hua Ling*[4], *Simon King*[2], *Thierry Dutoit*[1]
[1]TCTS Lab., Numediart Institute, University of Mons, Belgium; [2]University of Edinburgh, United Kingdom; [3]National Institute of Informatics, Tokyo, Japan; [4]University of Science and Technology of China (USTC), China

In this paper, we present the recent progress in the MAGE project. MAGE is a library for realtime and interactive (reactive) parametric speech synthesis using

hidden Markov models (HMMs). Here, it is broadened in order to support not only the standard acoustic features (spectrum and f0) to model and synthesize speech but also to combine acoustic and articulatory features, such as tongue, lips and jaw positions. Such an integration enables the user to have a straight forward and meaningful control space to intuitively modify the synthesized phones in real time only by configuring the position of the articulators.

DS-3 **Monday 14:20 – 15:20**

Reactive accent interpolation through an interactive map application

Maria Astrinaki[1], *Junichi Yamagishi*[2,3], *Simon King*[2], *Nicolas d'Alessandro*[1], *Thierry Dutoit*[1]

[1]TCTS Lab., Numediart Institute, University of Mons, Belgium; [2]University of Edinburgh, United Kingdom; [3]National Institute of Informatics, Tokyo, Japan

MAGE enables the reactive and continuous models modification in the HMM-based speech synthesis framework. Here, we present our first prototype system for extended interpolation applied for interactive accent control. Available accent models for American, Canadian and British English are manipulated in realtime by means of a gesturally controlled interactive geographical map. The accent interpolation is applied to one gender at a time, but the user is able to reactive alter between genders, while controlling the speakers to be interpolated at a time.

DS-4 **Monday 14:20 – 15:20**

Real-Time Control of Expressive Speech Synthesis Using Kinect Body Tracking

Christophe Veaux[1], *Maria Astrinaki*[2], *Keiichiro Oura*[3], *Robert A. J. Clark*[1], *Junichi Yamagishi*[1]

[1]University of Edinburgh, United Kingdom; [2]TCTS Lab., Numediart Institute, University of Mons, Belgium; [3]Departament of Computer Science, Nagoya Institute of Technology, Japan

The flexibility of statistical parametric speech synthesis has recently led to the development of interactive speech synthesis systems where different aspects of the voice output can be continuously controlled. The demonstration presented in this paper is based on MAGE/pHTS, a real-time synthesis system developed at Mons

University. This system enhances the controllability and the reactivity of HTS by enabling the generation of the speech parameters on the fly. This demonstration gives an illustration of the new possibilities offered by this approach in terms of interaction. A kinect sensor is used to follow the gestures and body posture of the user and these physical parameters are mapped to the prosodic parameters of an HMM-based singing voice model. In this way, the user can directly control various aspect of the singing voice such as the vibrato, the fundamental frequency or the duration. An avatar is used to encourage and facilitate the user interaction.

Poster Session 3

Monday, September 2
Chair: Keikichi Hirose
University of Tokyo, Japan

PS3-1 **Monday 15:20 – 17:00**

SASSC: A Standard Arabic Single Speaker Corpus

Ibrahim Almosallam, Atheer Alkhalifa, Mansour Alghamdi, Mohamed Alkanhal, Ashraf Alkhairy
KACST, Saudi Arabia

This paper describes the process of collecting and recording a large scale Arabic single speaker speech corpus. The collection and recording of the corpus was supervised by professional linguists and was recorded by a professional speaker in a soundproof studio using specialized equipments and stored in high quality formats. The pitch of the speaker (EGG) was also recorded and synchronized with the speech signal. Careful attempts were taken to insure the quality and diversity of the read text to insure maximum presence and combinations of words and phonemes. The corpus consists of 51 thousand words that required 7 hours of recording, and it is freely available for academic and research purposes.

PS3-2 **Monday 15:20 – 17:00**

Prosodically Modifying Speech for Unit Selection Speech Synthesis Databases

Ladan Golipour, Alistair Conkie, Ann Syrdal
AT&T, United States

This paper investigates the practical limits of artificially increasing the prosodic richness of a unit selection database by transforming the prosodic realization of

constituent sentences. The resulting high-quality transformed sentences are added to the database as new material. We examine in detail one of the most challenging prosodic transformations, namely converting statements into yes/no questions. Such transformations can require very large prosodic modifications while at the same time there is a need to retain as much naturalness of the signal as possible. Our data-driven approach relies on learning templates of pitch contours for different stress patterns of interrogative sentences from training data and later on applying these template pitch contours on unseen statements to generate the corresponding questions. We examine experimentally how the modified signals contribute to the perceived synthesis quality of the resulting database when compared with baseline unmodified databases.

PS3-3 **Monday 15:20 – 17:00**

Combining a Vector Space Representation of Linguistic Context with a Deep Neural Network for Text-To-Speech Synthesis

Heng Lu, Simon King, Oliver Watts

University of Edinburgh, United Kingdom

Conventional statistical parametric speech synthesis relies on decision trees to cluster together similar contexts, resulting in tied-parameter context-dependent hidden Markov models (HMMs). However, decision tree clustering has a major weakness: it use hard division and subdivides the model space based on one feature at a time, fragmenting the data and failing to exploit interactions between linguistic context features. These linguistic features themselves are also problematic, being noisy and of varied relevance to the acoustics. We propose to combine our previous work on vector-space representations of linguistic context, which have the added advantage of working directly from textual input, and Deep Neural Networks (DNNs), which can directly accept such continuous representations as input. The outputs of the network are probability distributions over speech features. Maximum Likelihood Parameter Generation is then used to create parameter trajectories, which in turn drive a vocoder to generate the waveform. Various configurations of the system are compared, using both conventional and vector space context representations and with the DNN making speech parameter predictions at two different temporal resolutions: frames, or states. Both objective and subjective results are presented.

PS3-4 **Monday 15:20 – 17:00**

Is Unit Selection Aware of Audible Artifacts?

Jindřich Matoušek, Daniel Tihelka, Milan Legát

University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics, Czech Republic

This paper presents a new analytic method that can be used for analysing perceptual relevance of unit selection costs and/or their sub-components as well as for tuning of the unit selection weights. The proposed method is leveraged to investigate the behaviour of a unit selection based system. The outcome is applied in a simple experiment with the aim to improve speech output quality of the system by setting limits on the costs and their sub-components during the search for optimal sequences of units. The experiments reveal that a large number (36.17 %) of artifacts annotated by listeners are not reflected by the values of the costs and their sub-componets as currently implemented and tuned in the evaluated system.

PS3-5 **Monday 15:20 – 17:00**

Development of Electrolarynx with Hands-Free Prosody Control

Kenji Matsui[1]*, Kenta Kimura*[1]*, Yoshihisa Nakatoh*[2]*, Yumiko O. Kato*[3]

[1]Osaka Institute of Technology, Japan; [2]Kyushu Institute of Technology, Japan; [3]St. Marianna University School of Medicine, Japan

The feasibility of using a motion sensor to replace a conventional electrolarynx(EL) user interface was explored. Forearm motion signals from MEMS accelerometer was used to provide on/off and pitch frequency control. The vibration device was placed against the throat using support bandage. Very small battery operated ARM-based control unit was developed and placed on the wrist. The control unit has a function to convert the tilt angle into the pitch frequency, as well as the device enable/disable function and pitch range adjustment function. As for the forearm tilt angle to pitch frequency conversion, two different conversion methods, linear mapping method and F0 model-based method, were investigated. A perceptual evaluation, with two well-trained normal speakers and ten subjects, was performed. Results of the evaluation study showed that both methods were able to produce better speech quality in terms of the naturalness.

PS3-6 **Monday 15:20 – 17:00**

A Hybrid TTS between Unit Selection and HMM-based TTS un-

der limited data conditions

Trung-Nghia Phung[1,2], *Chi Mai Luong*[2], *Masato Akagi*[1]
[1]JAIST, Japan; [2]IoIT, Vietnam

The intelligibility of HMM-based TTS can reach that of the original speech. However, HMM-based TTS is far from natural. On the contrary, unit selection TTS is the most-natural sounding TTS currently. However, its intelligibility and naturalness on segmental duration and timing are not stable. Additionally, unit selection needs to store a huge amount of data for concatenation. Recently, hybrid approaches between these two TTS, i.e. the HMM trajectory tiling (HTT) TTS, have been studied to take advantages of both unit selection and HMM-based TTS. However, such methods still require a huge amount of data for rendering. In this paper, a hybrid TTS among unit selection, HMM-based TTS, and Temporal Decomposition (TD) is pro- posed motivating to take advantages of both unit selection and HMM-based TTS under limited data conditions. Here, TD is a sparse representation of speech that decomposes a spectral or prosodic sequence into two mutual independent components: static event targets and correspondent dynamic event functions. Previous studies show that the dynamic event functions are related to the perception of speech intelligibility, one core linguistic or content information, while the static event targets convey non-linguistic or style information. Therefore, by borrowing the concepts of unit selection to render the event targets of the spectral sequence, and directly borrowing the prosodic sequences and the dynamic event functions of the spectral sequence generated by HMM-based TTS, the naturalness and the intelligibility of the proposed hybrid TTS can reach the naturalness of unit selection, and the intelligibility of HMM-based TTS, respec- tively. Due to the sparse representation of TD, the proposed hybrid TTS can also ensure a small amount of data for rendering, which suitable for limited data conditions. The experimental results with a small Vietnamese dataset, simulated to be a "limited data condition", show that the proposed hybrid TTS outperformed all HMM-based TTS, unit selection, HTT TTS under a limited data conditions.

PS3-7 **Monday 15:20 – 17:00**

Wavelets for intonation modeling in HMM speech synthesis

Antti Suni[1], *Daniel Aalto*[1], *Tuomo Raitio*[2], *Paavo Alku*[2], *Martti Vainio*[1]
[1]University of Helsinki, Finland; [2]Aalto University School of Electrical Engineering, Finland

The pitch contour in speech contains information about different linguistic units

at several distinct temporal scales. At the finest level, the microprosodic cues are purely segmental in nature, whereas in the coarser time scales, lexical tones, word accents, and phrase accents appear with both linguistic and paralinguistic functions. Consequently, the pitch movements happen on different temporal scales: the segmental perturbations are faster than typical pitch accents and so forth. In HMM-based speech synthesis paradigm, slower intonation patterns are not easy to model. The statistical procedure of decision tree clustering highlights instances that are more common, resulting in good reproduction of microprosody and declination, but with less variation on word and phrase level compared to human speech. Here we present a system that uses wavelets to decompose the pitch contour into five temporal scales ranging from microprosody to the utterance level. Each component is then individually trained within HMM framework and used in a superpositional manner at the synthesis stage. The resulting system is compared to a baseline where only one decision tree is trained to generate the pitch contour.

PS3-8 **Monday 15:20 – 17:00**

A Common Attribute based Unified HTS framework for Speech Synthesis in Indian Languages

Ramani B[1], *S Lilly Christina*[1], *G Anushiya Rachel*[1], *Sherlin Solomi V*[1], *Mahesh Kumar Nandwana*[2], *Anusha Prakash*[2], *Aswin Shanmugam S*[2], *Raghava Krishnan*[2], *S Kishore Prahalad*[3], *K Samudravijaya*[4], *P Vijayalakshmi*[1], *T Nagarajan*[1], *Hema Murthy*[2]

[1]SSN College of Engineering, India; [2]IIT Madras, India; [3]IIIT Hyderabad, India; [4]TIFR Bombay, India

State-of-the art approaches to speech synthesis are unit selection based concatenative speech synthesis (USS) and hidden Markov model based Text to speech synthesis (HTS). The former is based on waveform concatenation of subword units, while the latter is based on generation of an optimal parameter sequence from subword HMMs. The quality of an HMM based synthesiser in the HTS framework, crucially depends on an accurate description of the phoneset, and accurate description of the question set for clustering of the phones. Given the number of Indian languages, building a HTS system for every language is time consuming. Exploiting the properties of Indian languages, a uniform HMM framework for building speech synthesisers is proposed. Apart from the speech and text data used, the tasks involved in building a synthesis system can be made language-independent. A language-independent common phone set is first derived. Similar articulatory

descriptions also hold for sounds that are similar. The common phoneset and common question set are used to build HTS based systems for six Indian languages, namely, Hindi, Marathi, Bengali, Tamil, Telugu and Malayalam. Mean opinion score (MOS) is used to evaluate the system. An average MOS of 3.0 for naturalness and 3.4 for intelligibility is obtained for all language

PS3-9 **Monday 15:20 – 17:00**

Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for HMM-based speech synthesis

Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda

Nagoya Institute of Technology, Japan

This paper proposes a cross-lingual speaker adaptation (CLSA) method based on factor analysis using bilingual speech data. A state-mapping-based method has recently been proposed for CLSA. However, the method cannot transform only speaker-dependent characteristics. Furthermore, there is no theoretical framework for adapting prosody. To solve these problems, this paper presents a CLSA framework based on factor analysis using bilingual speech data. In this proposed method, model parameters representing language-dependent acoustic features and factors representing speaker characteristics are simultaneously optimized within a unified (maximum likelihood) framework based on a single statistical model by using bilingual speech data. This simultaneous optimization is expected to deliver a better quality of synthesized speech for the desired speaker characteristics. Experimental results show that the proposed method can synthesize better speech than the state-mapping-based method.

PS3-10 **Monday 15:20 – 17:00**

Residual Compensation based on Articulatory Feature-based Phone Clustering for Hybrid Mandarin Speech Synthesis

Yi-Chin Huang, Chung-Hsien Wu, Shih-Lun Lin

National Cheng-Kung University, Taiwan

While speech synthesis based on Hidden Markov Models (HMMs) has been developed to successfully synthesize stable and intelligible speech with flexibility and small footprints in recent years, HMM-based method is still incapable to generate the speech with good quality and high naturalness. In this study, a hybrid method

combining the unit-selection and HMM-based methods is proposed to compensate the residuals between the feature vectors of the natural phone units and the HMM-synthesized phone units to select better units and improve the naturalness of the synthesized speech. Articulatory features are adopted to cluster the phone units with similar articulation to construct the residual models of phone clusters. One residual model is characterized for each phone cluster using state-level linear regression. The candidate phone units of the natural corpus are selected by considering the compensated synthesized phone units of the same phone cluster, and then an optimal phone sequence is decided by the spectral features, contextual articulatory features, and pitch values to generate the synthesized speech with better naturalness. Objective and subjective evaluations were conducted and the comparison results to the HMM-based method and the conventional hybrid-based method confirm the improved performance of the proposed method.

Index

Notes