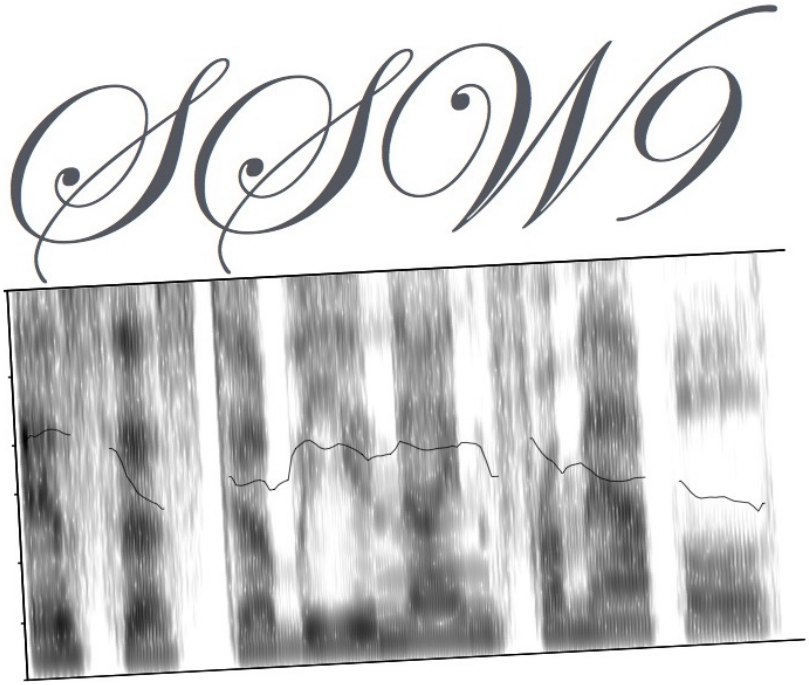

9th ISCA Workshop on Speech Synthesis

Proceedings

Sunnyvale (CA, USA) • Septembre 13 – 15, 2016

Edited by Antonio Bonafonte and Kishore Prahallad

At the time of release, the proceedings can be downloaded from the website of the SSW9: ssw9.net or ssw9.talp.cat

Contents

Program Committee

Chair:
Alan W Black, Carnegie Mellon University, USA

Members:
Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain.
Peter Cahill, Voysis, Ireland.
Alistair Conkie, Apple, USA.
Thierry Dutoit, Faculté Polytechnique de Mons, Belgium.
Raul Fernandez, IBM, USA.
Simon King, University of Edinburgh, UK.
Gopala Krishna Anumachipalli, University of California, San Francisco, USA.
Javier Latorre, Amazon, UK.
Kevin Lenzo, Apple, USA.
Zhenhua Ling, University of Science and Technology of China.
Hema A. Murthy, Indian Institute of Technology Madras, India.
Kishore Prahallad, Apple, USA.
Yao Qian, Educational Testing Service (ETS), USA.
Tuomo Raitio, Apple, USA.
Yannis Stylianou, University of Crete/Toshiba Cambridge Research Labs, UK.
David Suendermann-Oeft, Educational Testing Service (ETS), USA.
Tomoki Toda, Nagoya University, Japan.
Keiichi Tokuda, Nagoya Institute of Technology, Japan.
Junichi Yamagishi, National Institute of Informatics, Japan / Univ. of Edinburg, UK.
Heiga Zen, Google, UK.

Contributing Reviewers:
David Escudero, Universidad de Valladolid, Spain
Igor Jauk, Universitat Politècnica de Catalunya, Spain
Ladan Grolipour, Apple, USA
Asunción Moreno, Universitat Politècnica de Catalunya, Spain
Santiago Pascual, Universitat Politècnica de Catalunya, Spain
Zhizheng Wu, Apple, USA

Organizing Committee

Members:

Kishore Prahallad, Apple
Antonio Bonafonte, Universitat Politècnica de Catalunya
David Winarsky, Apple
Gopala Krishna Anumachipalli, University of California
Peter Cahill, Voysis

Technical Program

Tuesday, September 13

Keynote Session 1

Tuesday, September 13, 09:30 – 10:30

Chair: Simon King

Oral Session 1: Prosody.

Tuesday, September 13, 11:00 – 13:00

Chair: Ingmar Steiner

Poster Session 1

Tuesday, September 13, 15:00 – 17:00

Chair: Sébastien Le Maguer

John Andersson, Sebastian Berlin, André Costa, Harald Berthelsen, Hanna Lindgren, Nikolaj Lindberg, Jonas Beskow, Jens Edlund, Joakim Gustafson

Wednesday, September 14

Keynote Session 2

Wednesday, September 14, 09:30 – 10:30
Chair: Alan W. Black

Oral Session 2: Deep Learning in Speech Synthesis

Wednesday, September 14, 11:00 – 13:00
Chair: Tomoki Toda

Demo Session

Wednesday, September 14, 15:00 – 16:00
Chair: Keiichi Tokuda

Poster Session 2

Wednesday, September 14, 16:00 – 18:00
Chair: Sunayana Sitaram and Zhizheng WU

Thursday, September 15

Keynote Session 3

Thursday, September 15, 09:30 – 10:30

Chair: Keiichi Tokuda

KN3 09:30 – 10:30 — End-to-end Learning for Text and Speech *40*
Quoc V. Le

Oral Session 3: Analysis and Modeling for Speech Synthesis

Thursday, September 15, 11:00 – 13:00

Chair: Oliver Watts

OS3-1 11:00 – 11:30 — A hybrid harmonics-and-bursts modelling approach to speech synthesis *41*
Jonas Beskow, Harald Berthelsen

OS3-2 11:30 – 12:00 — A Pulse Model in Log-domain for a Uniform Synthesizer *42*
Gilles Degottex, Pierre Lanchantin, Mark Gales

OS3-3 12:00 – 12:30 — Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis *42*
Hideki Kawahara, Yannis Agiomyrgiannakis, Heiga Zen

OS3-4 12:30 – 13:00 — Wideband Harmonic Model: Alignment and Noise Modeling for High Quality Speech Synthesis *43*
Slava Shechtman, Alex Sorin

Abstracts

Keynote Session 1

Tuesday, September 13
Chair: Simon King
University of Edinburgh, UK

Tuesday 09:30 – 10:30

Large-scale finite element simulations of the physics of voice

Oriol Guasch
La Salle - Universitat Ramon Llull, Barcelona, Spain

The physics of voice is very complex and encompasses turbulent airflows interacting with vibrating, colliding and deforming bodies, like the vocal folds or the lips, and with acoustic waves propagating in a dynamic contorted vocal tract. Numerical approaches, and in particular the finite element method (FEM), have revealed as the most suitable option to solve many of those physical phenomena, and why not, attempting at a unified simulation, from muscle articulation and phonation to the emitted sound, in the mid-term. In this talk we will review some of the state of the art and current challenges in numerical voice production; from static and dynamic vowel sounds to sibilants and the self oscillations of the vocal folds. Numerical methods can be very appealing because they allow one not only to listen to a simulated sound but also to visualize the sound sources and the propagation of acoustic waves through the vocal tract. However, care should be taken not to use FEM as a black box. Even if a fully unified simulation of the whole process of voice generation was possible in an ideal supercomputer, would this reveal all the physics beneath voice production.

Oral Session 1: Prosody.

Tuesday, September 13
Chair: Ingmar Steiner
DFKI, Germany

OS1-1 **Tuesday 11:00 – 11:30**

Automatic, model-based detection of pause-less phrase boundaries from fundamental frequency and duration features

Mahsa Sadat Elyasi Langarani, Jan van Santen
Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA

Prosodic phrase boundaries (PBs) are a key aspect of spoken communication. In automatic PB detection, it is common to use local acoustic features, textual features, or a combination of both. Most approaches – regardless of features used – succeed in detecting major PBs (break score "4" in ToBI annotation, typically involving a pause) while detection of intermediate PBs (break score "3" in ToBI annotation) is still challenging. In this study we investigate the detection of intermediate, "pauseless" PBs using prosodic models, using a new corpus characterized by strong prosodic dynamics and an existing (CMU) corpus. We show how using duration and fundamental frequency modeling can improve detection of these PBs, as measured by the F1 score, compared to Festival, which only uses textual features to detect PBs. We believe that this study contributes to our understanding of the prosody of phrase breaks.

OS1-2 **Tuesday 11:30 – 12:00**

Synthesising Filled Pauses: Representation and Datamixing

Rasmus Dall[1], *Marcus Tomalin*[2], *Mirjam Wester*[1]
[1]The Centre for Speech Technology Research, The University of Edinburgh, UK; [2]Cambridge University Engineering Department, University of Cambridge, UK

Filled pauses occur frequently in spontaneous human speech, yet modern text-to-speech synthesis systems rarely model these disfluencies overtly, and consequently they do not output convincing synthetic filled pauses. This paper presents a text-to-speech system that is specifically designed to model these particular disfluencies more efffectively. A preparatory investigation shows that a synthetic voice trained exclusively on spontaneous speech is perceived to be inferior in quality to a voice trained entirely on read speech, even though the latter does not handle filled pauses well. This motivates an investigation into the phonetic representation of filled pauses which show that, in a preference test, the use of a distinct phone for filled pauses is preferred over the standard /V/ phone and the alternative /@/ phone. In addition, we present a variety of data-mixing techniques to combine the strengths of standard synthesis systems trained on read speech corpora with the supplementary advantages offered by systems trained on spontaneous speech. In a MUSHRA-style test, it is found that the best overall quality is obtained by combining the two types of corpora using a source marking technique. Specifically, general speech is synthesised with a standard mark, while filled pauses are synthesised with a spontaneous mark, which has the added benefit of also producing filled pauses that are comparatively well synthesised.

OS1-3 **Tuesday 12:00 – 12:30**

Emphasis recreation for TTS using intonation atoms

Pierre-Edouard Honnet[1,2], *Philip N. Garner*[1]
[1]Idiap Research Institute, Martigny, Switzerland; [2]Ecole Polytechnique Fédérale de Lausanne, Switzerland

We are interested in emphasis for text to speech synthesis. In speech to speech translation, emphasising the correct words is important to convey the underlying meaning of a message. In this paper, we propose to use a generalised command-response (CR) model of intonation to generate emphasis in synthetic speech. We first analyse the differences in the model parameters between emphasised words in an acted emphasis scenario and their neutral counterpart. We investigate word level intonation modelling using simple random forest as a basis framework, to predict the parameters of the model in the specific case of emphasised word.

Based on the linguistic context of the words we want to emphasise, we attempt at recovering emphasis pattern in the intonation in originally neutral synthetic speech by generating word-level model parameters with similar context. The method is presented and initial results are given, on synthetic speech.

OS1-4 **Tuesday 12:30 – 13:00**

Prediction of Emotions from Text using Sentiment Analysis for Expressive Speech Synthesis

Eva Vanmassenhove[1], *João P. Cabral*[2], *Fasih Haider*[2]
[1]Dublin City University, Ireland; [2]Trinity College Dublin, Ireland

The generation of expressive speech is a great challenge for text-to-speech synthesis in audiobooks. One of the most important factors is the variation in speech emotion or voice style. In this work, we developed a method to predict the emotion from a sentence so that we can convey it through the synthetic voice. It consists of combining a standard emotion-lexicon based technique with the polarity-scores (positive/negative polarity) provided by a less fine-grained sentiment analysis tool, in order to get more accurate emotion-labels. The primary goal of this emotion prediction tool was to select the type of voice (one of the emotions or neutral) given the input sentence to a stateof-the-art HMM-based Text-to-Speech (TTS) system. In addition, we also combined the emotion prediction from text with a speech clustering method to select the utterances with emotion during the process of building the emotional corpus for the speech synthesizer. Speech clustering is a popular approach to divide the speech data into subsets associated with different voice styles. The challenge here is to determine the clusters that map out the basic emotions from an audiobook corpus that contains high variety of speaking styles, in a way that minimizes the need for human annotation. The evaluation of emotion classification from text showed that, in general, our system can obtain accuracy results close to that of human annotators. Results also indicate that this technique is useful in the selection of utterances with emotion for building expressive synthetic voices.

Poster Session 1

Tuesday, September 13

Chair: Sébastien Le Maguer
Saarland University, Germany

PS1-1 **Tuesday 15:00 – 17:00**

Non-filter waveform generation from cepstrum using spectral phase reconstruction

Yasuhiro Hamada[1], *Nobutaka Ono*[2], *Shigeki Sagayama*[1]

[1]Meiji University, Nakano, Tokyo, Japan; [2]National Institute of Informatics / The Graduate University for Advanced Studies, Tokyo, Japan

This paper discusses non-filter waveform generation from cepstral features using spectral phase reconstruction as an alternative method to replace the conventional source-filter model in text-to-speech (TTS) systems. As the primary purpose of the use of filters is considered as producing a waveform from the desired spectrum shape, one possible alternative of the sourcefilter framework is to directly convert the designed spectrum into a waveform by utilizing a recently developed " phase reconstruction "from the power spectrogram. Given cepstral features and fundamental frequency (F0) as desired spectrum from a TTS system, the spectrum to be heard by the listener is calculated by converting the cepstral features into a linear-scale power spectrum and multiplying with the pitch structure of F0 . The signal waveform is generated from the power spectrogram by spectral phase reconstruction. An advantageous property of the proposed method is that it is free from undesired amplitude and long time decay often caused by sharp resonances in recursive filters. In preliminary experiments, we compared temporal and gain characteristics of the synthesized speech using the proposed method and mel-log spectrum approximation (MLSA) filter. Results show the proposed

method performed better than the MLSA filter in the both characteristics of the synthesized speech, and imply a desirable properties of the proposed method for speech synthesis.

PS1-2 **Tuesday 15:00 – 17:00**

Investigating Spectral Amplitude Modulation Phase Hierarchy Features in Speech Synthesis

Alexandros Lazaridis, Milos Cernak, Pierre-Edouard Honnet, Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

In our recent work, a novel speech synthesis with enhanced prosody (SSEP) system using probabilistic amplitude demodulation (PAD) features was introduced. These features were used to improve prosody in speech synthesis. The PAD was applied iteratively for generating syllable and stress amplitude modulations in a cascade manner. The PAD features were used as a secondary input scheme along with the standard text-based input features in deep neural network (DNN) speech synthesis. Objective and subjective evaluation validated the improvement of the quality of the synthesized speech. In this paper, a spectral amplitude modulation phase hierarchy (S-AMPH) technique is used in a similar to the PAD speech synthesis scheme, way. Instead of the two modulations used in PAD case, three modulations, i.e., stress-, syllable- and phoneme-level ones (2, 5 and 20 Hz respectively) are implemented with the S-AMPH model. The objective evaluation has shown that the proposed system using the S-AMPH features improved synthetic speech quality in respect to the system using the PAD features; in terms of relative reduction in mel-cepstral distortion (MCD) by approximately 9% and in terms of relative reduction in root mean square error (RMSE) of the fundamental frequency (F0) by approximately 25%. Multi-task training is also investigated in this work, giving no statistically significant improvements.

PS1-3 **Tuesday 15:00 – 17:00**

Multidimensional scaling of systems in the Voice Conversion Challenge 2016

Mirjam Wester[1]*, Zhizheng Wu*[1]*, Junichi Yamagishi*[1,2]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK; [2]National Institute of Informatics, Japan

This study investigates how listeners judge the similarity of voice converted voices using a talker discrimination task. The data used is from the Voice Conversion Challenge 2016. 17 participants from around the world took part in building voice converted voices from a shared data set of source and target speakers. This paper describes the evaluation of similarity for four of the source-target pairs (two intra-gender and two cross-gender) in more detail. Multidimensional scaling was performed to illustrate where each system was perceived to be in an acoustic space compared to the source and target speakers and to each other.

PS1-4 **Tuesday 15:00 – 17:00**

An Automatic Voice Conversion Evaluation Strategy Based on Perceptual Background Noise Distortion and Speaker Similarity

Dong-Yan Huang[1], *Lei Xie*[2], *Yvonne Siu Wa Lee*[1], *Jie Wu*[2], *Huaiping Ming*[1], *Xiaohai Tian*[3], *Shaofei Zhang*[2], *Chuang Ding*[1], *Mei Li*[1], *Quy Hy Nguyen*[3], *Minghui Dong*[1], *Haizhou LI*[1]

[1]Institute for Infocomm Research, A* STAR, Singapore; [2]School of Computer Science, Northwestern Polytechnical University, Xi'an, China School of Computer Engineering, Nanyang Technological University (NTU);Singapore

Voice conversion aims to modify the characteristics of one speaker to make it sound like spoken by another speaker without changing the language content. This task has attracted considerable attention and various approaches have been proposed since two decades ago. The evaluation of voice conversion approaches, usually through time-intensive subject listening tests, requires a huge amount of human labor. This paper proposes an automatic voice conversion evaluation strategy based on perceptual background noise distortion and speaker similarity. Experimental results show that our automatic evaluation results match the subjective listening results quite well. We further use our strategy to select best converted samples from multiple voice conversion systems and our submission achieves promising results in the voice conversion challenge (VCC2016).

PS1-5 **Tuesday 15:00 – 17:00**

Nonaudible murmur enhancement based on statistical voice conversion and noise suppression with external noise monitoring

Yusuke Tajiri, Tomoki Toda

Graduate School of Information Science, Nagoya University, Japan

This paper presents a method for making nonaudible murmur (NAM) enhancement based on statistical voice conversion (VC) robust against external noise. NAM, which is an extremely soft whispered voice, is a promising medium for silent speech communication thanks to its faint volume. Although such a soft voice can still be detected with a special body-conductive microphone, its quality significantly degrades compared to that of air-conductive voices. It has been shown that the statistical VC technique is capable of significantly improving quality of NAM by converting it into the air-conductive voices. However, this technique is not helpful under noisy conditions because a detected NAM signal easily suffers from external noise, and acoustic mismatches are caused between such a noisy NAM signal and a previously trained conversion model. To address this issue, in this paper we apply our proposed noise suppression method based on external noise monitoring to the statistical NAM enhancement. Moreover, a known noise superimposition method is further applied in order to alleviate the effects of residual noise components on the conversion accuracy. The experimental results demonstrate that the proposed method yields significant improvements in the conversion accuracy compared to the conventional method.

PS1-6 **Tuesday 15:00 – 17:00**

Prosodic and Spectral iVectors for Expressive Speech Synthesis

Igor Jauk, Antonio Bonafonte

Universitat Politècnica de Catalunya, Barcelona, Spain

This work presents a study on the suitability of prosodic and acoustic features, with a special focus on i-vectors, in expressive speech analysis and synthesis. For each utterance of two different databases, a laboratory recorded emotional acted speech, and an audiobook, several prosodic and acoustic features are extracted. Among them, i-vectors are built not only on the MFCC base, but also on F0, power and syllable durations. Then, unsupervised clustering is performed using different feature combinations. The resulting clusters are evaluated calculating cluster entropy for labeled portions of the databases. Additionally, synthetic voices are trained, applying speaker adaptive training, from the clusters built from the audiobook. The voices are evaluated in a perceptual test where the participants have to edit an audiobook paragraph using the synthetic voices. The objective results suggest that i-vectors are very useful for the audiobook, where different speakers (book characters) are imitated. On the other hand, for the laboratory recordings,

traditional prosodic features outperform i-vectors. Also, a closer analysis of the created clusters suggest that different speakers use different prosodic and acoustic means to convey emotions. The perceptual results suggest that the proposed ivector based feature combinations can be used for audiobook clustering and voice training.

PS1-7 **Tuesday 15:00 – 17:00**

Development of a statistical parametric synthesis system for operatic singing in German

Michael Pucher[1], *Fernando Villavicencio*[2], *Junichi Yamagishi*[2,3]

[1]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria; [2]National Institute of Informatics, Japan; [3]Centre for Speech Technology Research, University of Edinburgh,UK

In this paper we describe the development of a Hidden Markov Model (HMM) based synthesis system for operatic singing in German, which is an extension of the HMM-based synthesis system for popular songs in Japanese and English called "Sinsy". The implementation of this system consists of German text analysis, lexicon and Letter-To-Sound (LTS) conversion, and syllable duplication, which enables us to convert a German MusicXML input into context-dependent labels for acoustic modelling. Using the front-end, we develop two operatic singing voices, female mezzo-soprano and male bass voices, based on our new database, which consists of singing data of professional opera singers based in Vienna. We describe the details of the database and the recording procedure that is used to acquire singing data of four opera singers in German. For HMM training, we adopt a singer (speaker)-dependent training procedure. For duration modelling we propose a simple method that hierarchically constrains note durations by the overall utterance duration and then constrains phone durations by the synthesised note duration. We evaluate the performance of the voices with two vibrato modelling methods that have been proposed in the literature and show that HMM-based vibrato modelling can improve the overall quality.

PS1-8 **Tuesday 15:00 – 17:00**

DNN-based Speech Synthesis for Indian Languages from ASCII text

Srikanth Ronanki[1], *Siva Reddy*[2], *Bajibabu Bollepalli*[3], *Simon King*[1]
[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom; [2]ILCC, School of Informatics, University of Edinburgh, United Kingdom; [3]Department of Signal Processing and Acoustics, Aalto University, Finland

Text-to-Speech synthesis in Indian languages has seen a lot of progress over the decade partly due to the annual Blizzard challenges. These systems assume the text to be written in Devanagari or Dravidian scripts which are nearly phonemic orthography scripts. However, the most common form of computer interaction among Indians is ASCII written transliterated text. Such text is generally noisy with many variations in spelling for the same word. In this paper we evaluate three approaches to synthesize speech from such noisy ASCII text: a naive UniGrapheme approach, a Multi-Grapheme approach, and a supervised Grapheme-to-Phoneme (G2P) approach. These methods first convert the ASCII text to a phonetic script, and then learn a Deep Neural Network to synthesize speech from that. We train and test our models on Blizzard Challenge datasets that were transliterated to ASCII using crowdsourcing. Our experiments on Hindi, Tamil and Telugu demonstrate that our models generate speech of competetive quality from ASCII text compared to the speech synthesized from the native scripts. All the accompanying transliterated datasets are released for public access.

PS1-9 **Tuesday 15:00 – 17:00**

Experiments with Cross-lingual Systems for Synthesis of Code-Mixed Text

Sunayana Sitaram[1], *Sai Krishna Rallabandi*[1], *Shruti Rijhwani*[1], *Alan W. Black*[2]
[1]Microsoft Research, India; [2]Carnegie Mellon University, USA

Most Text to Speech (TTS) systems today assume that the input is in a single language written in its native script, which is the language that the TTS database is recorded in. However, due to the rise in conversational data available from social media, phenomena such as code-mixing, in which multiple languages are used together in the same conversation or sentence are now seen in text. TTS systems capable of synthesizing such text need to be able to handle multiple languages at the same time, and may also need to deal with noisy input. Previously, we proposed a framework to synthesize code-mixed text by using a TTS database in a single language, identifying the language that each word was from, normalizing

spellings of a language written in a non-standardized script and mapping the phonetic space of mixed language to the language that the TTS database was recorded in. We extend this cross-lingual approach to more language pairs, and improve upon our language identification technique. We conduct listening tests to determine which of the two languages being mixed should be used as the target language. We perform experiments for code-mixed Hindi-English and German-English and conduct listening tests with bilingual speakers of these languages. From our subjective experiments we find that listeners have a strong preference for cross-lingual systems with Hindi as the target language for code-mixed Hindi and English text. We also find that listeners prefer cross-lingual systems in English that can synthesize German text for codemixed German and English text.

PS1-10 **Tuesday 15:00 – 17:00**

Jerk Minimization for Acoustic-To-Articulatory Inversion

Avni Rajpal, Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, India

The effortless speech production in humans requires coordinated movements of the articulators such as lips, tongue, jaw, velum, etc. Therefore, measured trajectories obtained are smooth and slowly varying. However, the trajectories estimated from acoustic-to-articulatory inversion (AAI) are found to be jagged. Thus, energy minimization is used as smoothness constraint for improving performance of the AAI. Besides energy minimization, jerk (i.e., rate of change of acceleration) is known for quantification of smoothness in case of human motor movements. Human motors are organized to achieve intended goal with smoothest possible movements, under the constraint of minimum accelerative transients. In this paper, we propose jerk minimization as an alternative smoothness criterion for frame-based acoustic-to-articulatory inversion. The resultant trajectories obtained are smooth in the sense that for articulatorspecific window size, they will have minimum jerk. The results using this criterion were found to be comparable with inversion schemes based on existing energy minimization criteria for achieving smoothness.

PS1-11 **Tuesday 15:00 – 17:00**

How to select a good voice for TTS

Sunhee Kim
Naver Labs, Naver Corporation, Korea

Even though the quality of synthesized speech is not necessarily guaranteed by the perceived quality of the speaker's natural voice, it is required to select a certain number of candidates based on their natural voice before moving to the evaluation stage of synthesized sentences. This paper describes a male speaker selection procedure for unit selection synthesis systems in English and Japanese based on perceptive evaluation and acoustic measurements of the speakers' natural voice. A perceptive evaluation is performed on eight professional voice talents of each language. A total of twenty native-speaker listeners are recruited in both languages and each listener is asked to rate on eight analytical factors by using a five-scale score and rank three best speakers. Acoustic measurement focuses on the voice quality by extracting two measures, Long Term Average Spectrum (LTAS), the so-called Speakers Formant (SPF), which corresponds to the peak intensity between 3 kHz and 4 kHz, and the Alpha ratio, lower level difference between 0 and 1 kHz and 1 and 4 kHz ranges. The perceptive evaluation results show a very strong correlation between the total score and the preference in both languages, 0.9183 in English and 0.8589 in Japanese. The correlations between the perceptive evaluation and acoustic measurements are moderate with respect to SPF and AR, 0.473 and -0.494 in English, and 0.288 and -0.263 in Japanese.

PS1-12 **Tuesday 15:00 – 17:00**

WikiSpeech – enabling open source text-to-speech for Wikipedia

John Andersson[1], *Sebastian Berlin*[1], *André Costa*[1], *Harald Berthelsen*[2], *Hanna Lindgren*[2], *Nikolaj Lindberg*[2], *Jonas Beskow*[3], *Jens Edlund*[3], *Joakim Gustafson*[3]
[1]Wikimedia Sverige, Sweeden; [2]STTS, Sweeden; [3]KTH, Sweeden

We present WikiSpeech, an ambitious joint project aiming to (1) make open source text-to-speech available through Wikimedia Foundation's server architecture; (2) utilize the large and active Wikipedia user base to achieve continuously improving text-to-speech; (3) improve existing and develop new crowdsourcing methods for text-to-speech; and (4) develop new and adapt current evaluation methods so that they are well suited for the particular use case of reading Wikipedia articles out loud while at the same time capable of harnessing the huge user base made available by Wikipedia. At its inauguration, the project is backed by The Swedish Post and Telecom Authority and headed by Wikimedia Sverige, STTS and KTH,

but in the long run, the project aims at broad multinational involvement. The vision of the project is freely available text-to-speech for all Wikipedia languages (currently 293). In this paper, we present the project itself and its first steps: requirements, initial architecture, and initial steps to include crowdsourcing and evaluation.

Keynote Session 2

Wednesday, September 14
Chair: Alan W. Black
Carnegie Mellon University, USA

Wednesday 09:30 – 10:30

Siri's voice gets deep learning

Alex Acero
Apple, USA

In iOS 10, the new Siri voices are built on a hybrid speech synthesizer leveraging deep learning. The goodness of a concatenation between two units is modeled by a Gaussian distribution on the acoustic vectors (MFCC, F0, and their deltas) with the means and variances being a function of the linguistic features. The goodness of a target is modeled similarly with the addition of duration to the acoustic vector. The means and variances of these Gaussians are obtained through a Mixture Density Network. The new Siri voices are more natural, smoother, and allow Siri's personality to shine through.

Oral Session 2: Deep Learning in Speech Synthesis

Wednesday, September 14
Chair: Tomoki Toda
Graduate School of Information Science, Nagoya University, Japan

OS2-1 **Wednesday 11:00 – 11:30**

Parallel and cascaded deep neural networks for text-to-speech synthesis

Manuel Sam Ribeiro[1], *Oliver Watts*[1], *Junichi Yamagishi*[1,2]
[1]Centre for Speech Technology Research, University of Edinburgh, UK; [2]National Institute of Informatics, Tokyo, Japan

An investigation of cascaded and parallel deep neural networks for speech synthesis is conducted. In these systems, suprasegmental linguistic features (syllable-level and above) are processed separately from segmental features (phone-level and below). The suprasegmental component of the networks learns compact distributed representations of high-level linguistic units without any segmental influence. These representations are then integrated into a frame-level system using a cascaded or a parallel approach. In the cascaded network, suprasegmental representations are used as input to the framelevel network. In the parallel network, segmental and suprasegmental features are processed separately and concatenated at a later stage. These experiments are conducted with a standard set of high-dimensional linguistic features as well as a hand-pruned one. It is observed that hierarchical systems are consistently preferred over the baseline feedforward systems. Similarly, parallel networks are preferred over cascaded networks.

OS2-2 **Wednesday 11:30 – 12:00**

Temporal modeling in neural network based statistical parametric speech synthesis

Keiichi Tokuda, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku
Nagoya Institute of Technology, Japan

This paper proposes a novel neural network structure for speech synthesis, in which spectrum, F0 and duration parameters are simultaneously modeled in a unified framework. In the conventional neural network approaches, spectrum and F0 parameters are predicted by neural networks while phone and/or state durations are given from other external duration predictors. In order to consistently model not only spectrum and F0 parameters but also durations, we adopt a special type of mixture density network (MDN) structure, which models utterance level probability density functions conditioned on the corresponding input feature sequence. This is achieved by modeling the conditional probability distribution of utterance level output features, given input features, with a hidden semi-Markov model, where its parameters are generated using a neural network trained with a log likelihood-based loss function. Variations of the proposed neural network structure are also discussed. Subjective listening test results show that the proposed approach improves the naturalness of synthesized speech.

OS2-3 **Wednesday 12:00 – 12:30**

Multi-output RNN-LSTM for multiple speaker speech synthesis with α-interpolation model

Santiago Pascual, Antonio Bonafonte
Universitat Politècnica de Catalunya, Barcelona, Spain

Deep Learning has been applied successfully to speech processing. In this paper we propose an architecture for speech synthesis using multiple speakers. Some hidden layers are shared by all the speakers, while there is a specific output layer for each speaker. Objective and perceptual experiments prove that this scheme produces much better results in comparison with single speaker model. Moreover, we also tackle the problem of speaker interpolation by adding a new output layer (α-layer) on top of the multi-output branches. An identifying code is injected into the layer together with acoustic features of many speakers. Experiments show that the α-layer can effectively learn to interpolate the acoustic features between speakers.

OS2-4 **Wednesday 12:30 – 13:00**

A Comparative Study of the Performance of HMM, DNN, and RNN based Speech Synthesis Systems Trained on Very Large Speaker-Dependent Corpora

Xin Wang[1,2], *Shinji Takaki*[1], *Junichi Yamagishi*[1,2,3]

[1]National Institute of Informatics, Japan; [2]SOKENDAI University, Japan; [3]University of Edinburgh, UK

This study investigates the impact of the amount of training data on the performance of parametric speech synthesis systems. A Japanese corpus with 100 hours' audio recordings of a male voice and another corpus with 50 hours' recordings of a female voice were utilized to train systems based on hidden Markov model (HMM), feed-forward neural network and recurrent neural network (RNN). The results show that the improvement on the accuracy of the predicted spectral features gradually diminishes as the amount of training data increases. However, different from the "diminishing returns" in the spectral stream, the accuracy of the predicted F0 trajectory by the HMM and RNN systems tends to consistently benefit from the increasing amount of training data.

Demo Session

Wednesday, September 14

Chair: Keiichi Tokuda
Nagoya Institute of Technology, Japan

DS-1 **Wednesday 15:00 – 16:00**

Prosodic Reading Tutor of Japanese, Suzuki-kun: The first and only educational tool to teach the formal Japanese

Nobuaki Minematsu[1], *Daisuke Saito*[1], *Nobuyuki Nishizawa*[2]
[1]The University of Tokyo; [2]KDDI R&D Lab.

A text typed to a speech synthesizer is generally converted into its corresponding phoneme sequence on which various kinds of prosodic symbols are attached by a prosody prediction module. By using this module effectively, we build a prosodic reading tutor of Japanese, called Suzuki-kun, and it is provided as one feature of OJAD (Online Japanese Accent Dictionary). In Suzuki-kun, any Japanese text is converted into its reading (Hiragana 1 sequence) on which the pitch pattern that sounds natural as Tokyo Japanese (the formal Japanese) is visualized as a smooth curve drawn by the F0 contour generation process model. Further, the positions of accent nuclei and unvoiced vowels are illustrated. Suzuki-kun also reads that text out following the prosodic features that are visualized. Suzuki-kun has become the most popular feature of OJAD and so far, we gave 90 tutorial workshops of OJAD in 27 countries. After INTERSPEECH, we'll give 6 workshops in the USA this year.

DS-2 **Wednesday 15:00 – 16:00**

Aliasing-free L-F model and its application to an interactive MATLAB tool and test signal generation for speech analysis procedures

Hideki Kawahara
Wakayama University

This demo introduces a closed form representation of the L-F model for excitation source. The representation provides flexible of source parameters in continuous time axis and aliasing-free excitation signal. MATLAB implementation of the model combined with an interactive parameter control and visual and sound feedback is a central component of educational/research tools for speech science. The model also provides flexible and accurate test signals applicable to test speech analysis procedures, such as F0 trackers and spectrum envelope estimator.

DS-3 **Wednesday 15:00 – 16:00**

A Demonstration of the Merlin Open Source Neural Network Speech Synthesis System

Srikanth Ronanki, Zhizheng Wu, Oliver Watts, Simon King
University of Edinburgh, United Kingdom

This demonstration showcases our new Open Source toolkit for neural network-based speech synthesis, Merlin. We wrote Merlin because we wanted free, simple, maintainable code that we understood. No existing toolkits met all of those requirements. Merlin is designed for speech synthesis, but can be put to other uses. It has already also been used for voice conversion, classification tasks, and for predicting head motion from speech.

DS-4 **Wednesday 15:00 – 16:00**

WaveNet: A Generative Model for Raw Audio

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu
Google

This demo presents WaveNet, a deep generative model of raw audio waveforms. We show that WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech (TTS) systems, reducing the gap in subjective quality relative to natural speech by over

50%. We also demonstrate that the same network can be used to synthesize other audio signals such as music, and present some striking samples of automatically generated piano pieces. WaveNets open up a lot of possibilities for text-to-speech, music generation and audio modelling in general.

DS-5 **Wednesday 15:00 – 16:00**

Demo of Idlak Tangle, An Open Source DNN-Based Parametric Speech Synthesiser

Blaise Potard, Matthew P. Aylett, David A. Baude

CereProc Ltd., United Kingdom; ; CSTR, University of Edinburgh, United Kingdom

We present a live demo of Idlak Tangle, a TTS extension to the ASR toolkit Kaldi [1]. Tangle combines the Idlak front-end and newly released MLSA vocoder, with two DNNs modelling respectively the units duration and acoustic parameters, providing a fully functional end-to-end TTS system. The system has none of the licensing restrictions of currently available HMM style systems, such as the HTS toolkit, and can be used free of charge for any type of applications. Experimental results using the freely available SLT speaker from CMU ARCTIC, reveal that the speech output is rated in a MUSHRA test as significantly more natural than the output of HTS-demo. The tools, audio database and recipe required to reproduce the results presented are fully available online at https://github.com/bpotard/idlak . The live demo will allow participants to measure the quality of TTS output on several ARCTIC voices, and on voices created from commercial-grade recordings.

Poster Session 2

Wednesday, September 14

Chair: Sunayana Sitaram; Microsoft
Zhizheng Wu; Apple

PS2-1 **Wednesday 16:00 – 18:00**

Non-intrusive Quality Assessment of Synthesized Speech using Spectral Features and Support Vector Regression

Meet H. Soni, Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology, India

In this paper, we propose a new quality assessment method for synthesized speech. Unlike previous approaches which uses Hidden Markov Model (HMM) trained on natural utterances as a reference model to predict the quality of synthesized speech, proposed approach uses knowledge about synthesized speech while training the model. The previous approach has been successfully applied in the quality assessment of synthesized speech for the German language. However, it gave poor results for English language databases such as Blizzard Challenge 2008 and 2009 databases. The problem of quality assessment of synthesized speech is posed as a regression problem. The mapping between statistical properties of spectral features extracted from the speech signal and corresponding speech quality score (MOS) was found using Support Vector Regression (SVR). All the experiments were done on Blizzard Challenge Databases of the year 2008, 2009, 2010 and 2012. The results of experiments show that by including knowledge about synthesized speech while training, the performance of quality assessment system can be improved. Moreover, the accuracy of quality assessment system heavily depends on the kind of synthesis system used for signal generation. On Blizzard 2008 and 2009

database, proposed approach gives correlation of 0.28 and 0.49, respectively, for about 17 % data used in training. Previous approach gives correlation of 0.3 and 0.09, respectively, using spectral features. For Blizzard 2012 database, proposed approach gives correlation of 0.8 by using 12 % of available data in training.

PS2-2 **Wednesday 16:00 – 18:00**

Novel Pre-processing using Outlier Removal in Voice Conversion

Sushant V. Rao, Nirmesh J Shah, Hemant A. Patil
Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar

Voice conversion (VC) technique modifies the speech utterance spoken by a source speaker to make it sound like a target speaker is speaking. Gaussian Mixture Model (GMM)-based VC is a state-of-the-art method. It finds the mapping function by modeling the joint density of source and target speakers using GMM to convert spectral features framewise. As with any real dataset, the spectral parameters contain a few points that are inconsistent with the rest of the data, called outliers. Until now, there has been very few literature regarding the effect of outliers in voice conversion. In this paper, we have explored the effect of outliers in voice conversion, as a pre-processing step. In order to remove these outliers, we have used the score distance, which uses the scores estimated using Robust Principal Component Analysis (ROBPCA). The outliers are determined by using a cut-off value based on the degrees of freedom in a chi-squared distribution. They are then removed from the training dataset and a GMM is trained based on the least outlying points. This pre-processing step can be applied to various methods. Experimental results indicate that there is a clear improvement in both, the objective (8 %) as well as the subjective (4 % for MOS and 5 % for XAB) results.

PS2-3 **Wednesday 16:00 – 18:00**

Emotional Voice Conversion Using Neural Networks with Different Temporal Scales of F0 based on Wavelet Transform

Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki
Graduate School of System Informatics, Kobe University, Japan

An artificial neural network is one of the most important models for training features of voice conversion (VC) tasks. Typically, neural networks (NNs) are

very effective in processing nonlinear features, such as mel cepstral coefficients (MCC) which represent the spectrum features. However, a simple representation for fundamental frequency (F0) is not enough for neural networks to deal with an emotional voice, because the time sequence of F0 for an emotional voice changes drastically. Therefore, in this paper, we propose an effective method that uses the continuous wavelet transform (CWT) to decompose F0 into different temporal scales that can be well trained by NNs for prosody modeling in emotional voice conversion. Meanwhile, the proposed method uses deep belief networks (DBNs) to pretrain the NNs that convert spectral features. By utilizing these approaches, the proposed method can change the spectrum and the prosody for an emotional voice at the same time, and was able to outperform other state-of-the-art methods for emotional voice conversion.

PS2-4 **Wednesday 16:00 – 18:00**

Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech

Cassia Valentini-Botinhao[1], *Xin Wang*[2,3], *Shinji Takaki*[2], *Junichi Yamagishi*[1,2,3]

[1]The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK; [2]National Institute of Informatics, Japan; [3]SOKENDAI University, Japan

The quality of text-to-speech (TTS) voices built from noisy speech is compromised. Enhancing the speech data before training has been shown to improve quality but voices built with clean speech are still preferred. In this paper we investigate two different approaches for speech enhancement to train TTS systems. In both approaches we train a recursive neural network (RNN) to map acoustic features extracted from noisy speech to features describing clean speech. The enhanced data is then used to train the TTS acoustic model. In one approach we use the features conventionally employed to train TTS acoustic models, i.e Mel cepstral (MCEP) coefficients, aperiodicity values and fundamental frequency (F0). In the other approach, following conventional speech enhancement methods, we train an RNN using only the MCEP coefficients extracted from the magnitude spectrum. The enhanced MCEP features and the phase extracted from noisy speech are combined to reconstruct the waveform which is then used to extract acoustic features to train the TTS system. We show that the second approach results in larger MCEP distortion but smaller F0 errors. Subjective evaluation shows that

synthetic voices trained with data enhanced with this method were rated higher and with similar to scores to voices trained with clean speech.

PS2-5 **Wednesday 16:00 – 18:00**

Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis

Shinji Takaki[1], *SangJin Kim*[2], *Junichi Yamagishi*[1,3]

[1]National Institute of Informatics, Tokyo, Japan; [2]Naver Labs, Naver Corporation, Seongnam, Korea; [3]University of Edinburgh, United Kingdom

In this paper, we investigate the effectiveness of speaker adaptation for various essential components in deep neural network based speech synthesis, including acoustic models, acoustic feature extraction, and post-filters. In general, a speaker adaptation technique, e.g., maximum likelihood linear regression (MLLR) for HMMs or learning hidden unit contributions (LHUC) for DNNs, is applied to an acoustic modeling part to change voice characteristics or speaking styles. However, since we have proposed a multiple DNN-based speech synthesis system, in which several components are represented based on feed-forward DNNs, a speaker adaptation technique can be applied not only to the acoustic modeling part but also to other components represented by DNNs. In experiments using a small amount of adaptation data, we performed adaptation based on LHUC and simple additional fine tuning for DNNbased acoustic models, deep auto-encoder based feature extraction, and DNN-based post-filter models and compared them with HMM-based speech synthesis systems using MLLR.

PS2-6 **Wednesday 16:00 – 18:00**

Mandarin Prosodic Phrase Prediction based on Syntactic Trees

Zhengchen Zhang[1], *Fuxiang Wu*[2], *Chenyu Yang*[1], *Minghui Dong*[1], *Fugen Zhou*[2]

[1]Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore; [2]BeiHang University, Beijing, China

Prosodic phrases (PPs) are important for Mandarin Text-To-Speech systems. Most of the existing PP detection methods need large manually annotated corpora to learn the models. In this paper, we propose a rule based method to predict the PP boundaries employing the syntactic information of a sentence. The method is

based on the observation that a prosodic phrase is a meaningful segment of a sentence with length restrictions. A syntactic structure allows to segment a sentence according to grammars. We add some length restrictions to the segmentations to predict the PP boundaries. An F-Score of 0.693 was obtained in the experiments, which is about 0.02 higher than the one got by a Conditional Random Field based method.

PS2-7 **Wednesday 16:00 – 18:00**

Investigating Very Deep Highway Networks for Parametric Speech Synthesis

Xin Wang[1,2], *Shinji Takaki*[1], *Junichi Yamagishi*[1,2,3]
[1]National Institute of Informatics, Japan; [2]SOKENDAI University, Japan; [3]University of Edinburgh, UK

The depth of the neural network is a vital factor that affects its performance. Recently a new architecture called highway network was proposed. This network facilitates the training process of a very deep neural network by using gate units to control a information highway over the conventional hidden layer. For the speech synthesis task, we investigate the performance of highway networks with up to 40 hidden layers. The results suggest that a highway network with 14 non-linear transformation layers is the best choice on our speech corpus and this highway network achieves better performance than a feed-forward network with 14 hidden layers. On the basis of these results, we further investigate a multi-stream highway network where separate highway networks are used to predict different kinds of acoustic features such as the spectral and F0 features. Results of the experiments suggest that the multi-stream highway network can achieve better objective results than the single network that predicts all the acoustic features. Analysis on the output of highway gate units also supports the assumption for the multi-stream network that different hidden representation may be necessary to predict spectral and F0 features.

PS2-8 **Wednesday 16:00 – 18:00**

Contextual Representation using Recurrent Neural Network Hidden State for Statistical Parametric Speech Synthesis

Sivanand Achanta, Rambabu Banoth, Ayushi Pandey, Anandaswarup Vadapalli, Suryakanth V Gangashetty

Speech and Vision Laboratory, IIIT, Hyderabad, India.

In this paper, we propose to use hidden state vector obtained from recurrent neural network (RNN) as a context vector representation for deep neural network (DNN) based statistical parametric speech synthesis. While in a typical DNN based system, there is a hierarchy of text features from phone level to utterance level, they are usually in 1-hot-k encoded representation. Our hypothesis is that, supplementing the conventional text features with a continuous frame-level acoustically guided representation would improve the acoustic modeling. The hidden state from an RNN trained to predict acoustic features is used as the additional contextual information. A dataset consisting of 2 Indian languages (Telugu and Hindi) from Blizzard challenge 2015 was used in our experiments. Both the subjective listening tests and the objective scores indicate that the proposed approach performs significantly better than the baseline DNN system.

PS2-9 **Wednesday 16:00 – 18:00**

Wide Passband Design for Cosine-Modulated Filter Banks in Sinusoidal Speech Synthesis

Nobuyuki Nishizawa, Tomonori Yazaki
KDDI R&D Laboratories Inc., Japan

A new filter design strategy to shorten the length of the filter is introduced for sinusoidal speech synthesis using cosinemodulated filter banks. Multiple sinusoidal waveforms for speech synthesis can be effectively synthesized by using pseudo-quadrature mirror filter (pseudo-QMF) banks, which are constructed by cosine modulation of the coefficients of a lowpass filter. This is because stable sinusoids are represented as sparse vectors on the subband domain of the pseudo-QMF banks and computation for the filter banks can be effectively performed with fast algorithms for discrete cosine transformation (DCT). However, the pseudo-QMF banks require relatively long filters to reduce noise caused by aliasing. In this study, a wider passband design with a perfect reconstruction (PR) QMF bank is introduced. The properties of experimentally designed filters indicated that the length of the filters can be reduced from 448 taps to 384 taps for 32-subband systems with less than -96dB errors where the computational cost for speech synthesis does not significantly increase.

PS2-10 **Wednesday 16:00 – 18:00**

Utterance Selection Techniques for TTS Systems Using Found Speech

Pallavi Baljekar, Alan W. Black

Language Technologies Institute, Carnegie Mellon University, USA

The goal in this paper is to investigate data selection techniques for found speech. Found speech unlike clean, phoneticallybalanced datasets recorded specifically for synthesis contain a lot of noise which might not get labeled well and it might contain utterances with varying channel conditions. These channel variations and other noise distortions might sometimes be useful in terms of adding diverse data to our training set, however in other cases it might be detrimental to the system. The approach outlined in this work investigates various metrics to detect noisy data which degrade the performance of the system on a held-out test set. We assume a seed set of 100 utterances to which we then incrementally add in a fixed set of utterances and find which metrics can capture the misaligned and noisy data. We report results on three datasets, an artificially degraded set of clean speech, a single speaker database of found speech and a multi - speaker database of found speech. All of our experiments are carried out on male speakers. We also show comparable results are obtained on a female multi-speaker corpus.

PS2-11 **Wednesday 16:00 – 18:00**

Open-Source Consumer-Grade Indic Text To Speech

Andrew Wilkinson[1]*, Alok Parlikar*[1]*, Sunayana Sitaram*[1]*, Tim White*[1,2]*, Alan W. Black*[1]*, Suresh Bazaj*[2]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA; [2]Hear2Read, Indians for Collective Action, Palo Alto, CA, USA

Open-source text-to-speech (TTS) software has enabled the development of voices in multiple languages, including many high-resource languages, such as English and European languages. However, building voices for low-resource languages is still challenging. We describe the development of TTS systems for 12 Indian languages using the Festvox framework, for which we developed a common frontend for Indian languages. Voices for eight of these 12 languages are available for use with Flite, a lightweight, fast run-time synthesizer, and the Android Flite app available in the Google Play store. Recently, the baseline Punjabi TTS voice was built end-to-end in a month by two undergraduate students (without any prior knowledge of TTS) with help from two of the authors of this paper. The frame-

work can be used to build a baseline Indic TTS voice in two weeks, once a text corpus is selected and a suitable native speaker is identified.

PS2-12 **Wednesday 16:00 – 18:00**

On the impact of phoneme alignment in DNN-based speech synthesis

Mei Li[1]*, Zhizheng Wu*[2]*, Lei Xie*[1]

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China; [2]The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK

Recently, deep neural networks (DNNs) have significantly improved the performance of acoustic modeling in statistical parametric speech synthesis (SPSS). However, in current implementations, when training a DNN-based speech synthesis system, phonetic transcripts are required to be aligned with the corresponding speech frames to obtain the phonetic segmentation, called phoneme alignment. Such an alignment is usually obtained by forced alignment based on hidden Markov models (HMMs) since manual alignment is labor-intensive and timeconsuming. In this work, we study the impact of phoneme alignment on the DNN-based speech synthesis system. Specifically, we compare the performances of different DNN-based speech synthesis systems, which use manual alignment and HMM-based forced alignment from three types of labels: HMM mono-phone, tri-phone and full-context. Objective and subjective evaluations are conducted in term of the naturalness of synthesized speech to compare the performances of different alignments.

PS2-13 **Wednesday 16:00 – 18:00**

Merlin: An Open Source Neural Network Speech Synthesis System

Zhizheng Wu, Oliver Watts, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

We introduce the Merlin speech synthesis toolkit for neural network-based speech synthesis. The system takes linguistic features as input, and employs neural net-

works to predict acoustic features, which are then passed to a vocoder to produce the speech waveform. Various neural network architectures are implemented, including a standard feedforward neural network, mixture density neural network, recurrent neural network (RNN), long short-term memory (LSTM) recurrent neural network, amongst others. The toolkit is Open Source, written in Python, and is extensible. This paper briefly describes the system, and provides some benchmarking results on a freely available corpus.

Keynote Session 3

Thursday, September 15
Chair: Keiichi Tokuda
Nagoya Institute of Technology, Japan

Thursday 09:30 – 10:30

End-to-end Learning for Text and Speech

Quoc V. Le
Google, USA

In this talk, I will discuss our recent work on using neural networks for NLP and speech recognition tasks. Our work started with the sequence-to-sequence learning framework that can read a variable-length input sequence and produce a variable-length output sequence. The framework allows neural networks to be applied to new tasks in text and speech domains. I will talk about the implementation details and results of our implementation on machine translation, dialogue modeling, and speech recognition. We also find that unsupervised learning in our framework is simple, and improves the performance of our networks significantly.

Oral Session 3: Analysis and Modeling for Speech Synthesis

Thursday, September 15
Chair: Oliver Watts
University of Edinburgh, UK

OS3-1 **Thursday 11:00 – 11:30**

A hybrid harmonics-and-bursts modelling approach to speech synthesis

Jonas Beskow[1]*, Harald Berthelsen*[2]
[1]KTH Speech, Music and Hearing, Stockholm, Sweden; [2]STTS Speech Technology Services, Stockholm, Sweden

Statistical speech synthesis systems rely on a parametric speech generation model, typically some sort of vocoder. Vocoders are great for voiced speech because they offer independent control over voice source (e.g. pitch) and vocal tract filter (e.g. vowel quality) through control parameters that typically vary smoothly in time and lend themselves well to statistical modelling. Voiceless sounds and transients such as plosives and fricatives on the other hand exhibit fundamentally different spectro-temporal behaviour. Here the benefits of the vocoder are not as clear. In this paper, we investigate a hybrid approach to modeling the speech signal, where speech is decomposed into an harmonic part and a noise burst part through spectrogram kernel filtering. The harmonic part is modeled using vocoder and statistical parameter generation, while the burst part is modeled by concatenation. The two channels are then mixed together to form the final synthesized waveform. The proposed method was compared against a state of the art statistical speech synthesis system (HTS 2.3) in a perceptual evaluation, which reveled that the harmonics plus bursts method was perceived as significantly more natural than the purely statistical variant.

OS3-2 **Thursday 11:30 – 12:00**

A Pulse Model in Log-domain for a Uniform Synthesizer

Gilles Degottex, Pierre Lanchantin, Mark Gales
Cambridge University, Engineering Department, Cambridge, UK

The quality of the vocoder plays a crucial role in the performance of parametric speech synthesis systems. In order to improve the vocoder quality, it is necessary to reconstruct as much of the perceived components of the speech signal as possible. In this paper, we first show that the noise component is currently not accurately modelled in the widely used STRAIGHT vocoder, thus, limiting the voice range that can be covered and also limiting the overall quality. In order to motivate a new, alternative, approach to this issue, we present a new synthesizer, which uses a uniform representation for voiced and unvoiced segments. This synthesizer has also the advantage of using a simple signal model compared to other approaches, thus offering a convenient and controlled alternative for future developments. Experiments analysing the synthesis quality of the noise component shows improved speech reconstruction using the suggested synthesizer compared to STRAIGHT. Additionally an experiment about analysis/resynthesis shows that the suggested synthesizer solves some of the issues of another uniform vocoder, Harmonic Model plus Phase Distortion (HMPD). In text-to-speech synthesis, it outperforms HMPD and exhibits a similar, or only slightly worse, quality to STRAIGHT's quality, which is encouraging for a new vocoding approach.

OS3-3 **Thursday 12:00 – 12:30**

Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis

Hideki Kawahara[1,2], *Yannis Agiomyrgiannakis*[1], *Heiga Zen*[1]
[1]Google, Wakayama University;Japan

This paper introduces a general and flexible framework for F0 and aperiodicity (additive non periodic component) analysis, specifically intended for high-quality speech synthesis and modification applications. The proposed framework consists of three subsystems: instantaneous frequency estimator and initial aperiodicity detector, F0 trajectory tracker, and F0 refinement and aperiodicity extractor. A preliminary implementation of the proposed framework substantially outperformed

(by a factor of 10 in terms of RMS F0 estimation error) existing F0 extractors in tracking ability of temporally varying F0 trajectories. The front end aperiodicity detector consists of a complex-valued wavelet analysis filter with a highly selective temporal and spectral envelope. This front end aperiodicity detector uses a new measure that quantifies the deviation from periodicity. The measure is less sensitive to slow FM and AM and closely correlates with the signal to noise ratio. The front end combines instantaneous frequency information over a set of filter outputs using the measure to yield an observation probability map. The second stage generates the initial F0 trajectory using this map and signal power information. The final stage uses the deviation measure of each harmonic component and F0 adaptive time warping to refine the F0 estimate and aperiodicity estimation. The proposed framework is flexible to integrate other sources of instantaneous frequency when they provide relevant information.

OS3-4 **Thursday 12:30 – 13:00**

Wideband Harmonic Model: Alignment and Noise Modeling for High Quality Speech Synthesis

Slava Shechtman, Alex Sorin

IBM Research, Haifa, Israel

Speech sinusoidal modeling has been successfully applied to a broad range of speech analysis, synthesis and modification tasks. However, developing a high fidelity full band sinusoidal model that preserves its high quality on speech transformation still remains an open research problem. Such a system can be extremely useful for high quality speech synthesis. In this paper we present an enhanced harmonic model representation for voiced/mixed wide band speech that is capable of high quality speech reconstruction and transformation in the parametric domain. Two key elements of the proposed model are a proper phase alignment and a decomposition of a speech frame to "deterministic" and dense "stochastic" harmonic model representations that can be separately manipulated. The coupling of stochastic harmonic representation with the deterministic one is performed by means of intra-frame periodic energy envelope, estimated at analysis time and preserved during original/transformed speech reconstruction. In addition, we present a compact representation of the stochastic harmonic component, so that the proposed model has less parameters than the regular full band harmonic model, with better Signal to Reconstruction Error performance. On top of that, the improved phase alignment of the proposed model provides better phase

coherency in transformed speech, resulting in better quality of speech transformations. We demonstrate the subjective and objective performance of the new model on speech reconstruction and pitch modification tasks. Performance of the proposed model within unit selection TTS is also presented.

Author's Index

Index

Notes