SSW10

The 10th ISCA Speech Synthesis Workshop

The Austrian museum of folk life and folk art, Laudongasse 15-19, A-1080 Vienna

SSW10 Sponsors

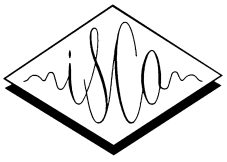
We would like to thank our workshop sponsors for their support.

Acoustics Research Institute

Amazon Alexa

Austrian Academy of Sciences

The International Speech Communication Association (ISCA)

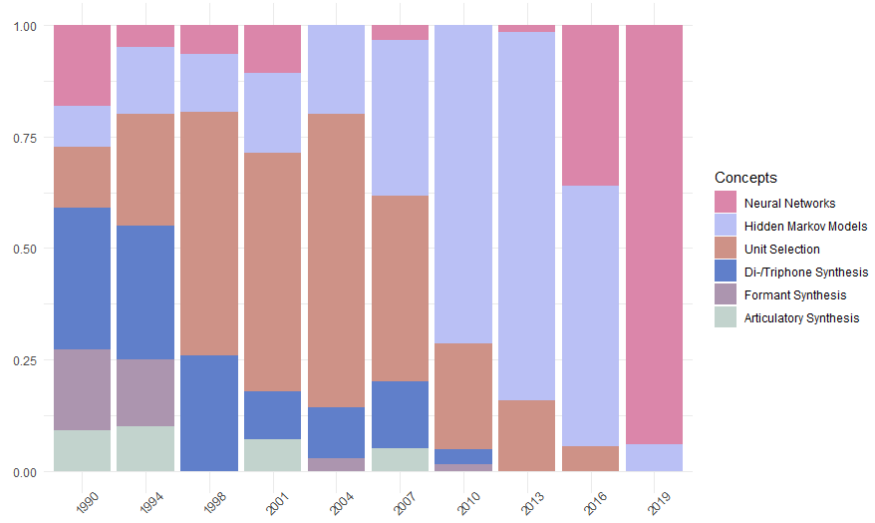


Message from the chair

Dear SSW10 participants,

I am happy to welcome you in Vienna for the 10th ISCA Speech Synthesis Workshop (SSW10), a satellite event of Interspeech 2019 that is organised in Graz, Austria.

The technical program contains 49 papers: 6 oral sessions with 21 presentations and 3 poster sessions with 28 posters. Each paper has been reviewed by at least two reviewers. The papers cover a diverse range of topics from neural vocoders and speech science to prosody. It has always been a unique feature of SSW to cover the many aspects of speech synthesis. However, at this SSW a paradigm shift towards Deep Neural Network (DNN) based speech synthesis is taking place. In a survey of topics of this and previous SSW workshops that was published recently [1], you can clearly see this paradigm shift as shown in the figure below where DNNs have taken over.



This paradigm shift is responsible for recent progress in the field, which also results in strong interest from the research community and industry. At SSW10, as the scientific forum for exchanging ideas on speech synthesis, this interest in the topic led to 200 registered participants. Apart from the speech synthesis research community speech synthesis has also become a benchmark for machine learning. The goal for this SSW was to put SSW10 at the center of this recent paradigm shift in speech synthesis, but also show the diversity of speech synthesis in speech science, evaluation, language varieties, and prosody.

Each day of the workshop begins with a keynote talk from a leading researcher, starting on Friday with Aäron van den Oord, who will talk about his recent work on „Deep learning for speech synthesis“. On Saturday Tecumseh Fitch and Bart de Boer will introduce us into the field of animal communication with a talk on

„Synthesizing animal vocalizations and modelling animal speech“. On Sunday Claire Gardent presents her work on „Natural Language Generation: Creating Text“, a possible new field for cooperation with speech synthesis research.

This is now the 10th workshop in this successful series of workshops that were held in Autrans 1990, Mohonk 1994, Jenolan Caves 1998, Pitlochry 2001, Pittsburgh 2004, Bonn 2007, Kyoto 2010, Barcelona 2013, Sunnyvale 2016, and Vienna 2019. To celebrate this 10th anniversary of SSW, an advisory board composed by the chairmans of the previous workshops, voted for the most important paper of each of the past SSW. The best papers will be awarded at SSW10 and will also be listed on the workshops website. This list will give participants of past SSW workshops the opportunity to revisit landmark papers, and for those colleagues that are new to speech synthesis it can show the different paradigms, problems, and steady progress of the field.

I would like to thank the organising committee, especially my colleagues Junichi Yamagishi and Sébastien Le Maguer for the work they have put into the preparation of the workshop. I would also like to thank the advisory committee for selection of the best papers of past SSW workshops and the scientific committee for thorough review of all submitted papers. Special thanks go to Konstantin Ulitsch from the Acoustics Research Institute (ARI) for his help on social events, venue, and more and to the team from ARI that helped during the workshop.

Thanks to the sponsoring from Amazon the regular registration fees could be kept at a level of previous SSW workshops, and the student registration fees could be significantly reduced. Two social events accompany the technical program. A welcome reception on the first day at the conference venue, and a dinner at a so-called „Buschenschank“, a restaurant in the vineyards that surround Vienna. In this restaurant you can taste local Viennese wine and food.

Finally I wish you a productive and interesting time at the workshop and nice experiences in Vienna. Should this be your first time in Vienna there are plenty of things to see and do whatever your interests are. A good way to start is always to walk around the Ringstrasse and in the first district. You can also go to the Prater, a big park or visit the Zentralfriedhof (central cemetery), if you are after a very Viennese experience. Definitely you should visit a Viennese coffee house if you want to be alone but need others for that (A. Polgar).

Yours sincerley

Michael Pucher (SSW chair)

Vienna, 10. September 2019

[1] C. Lozo, J. Luttenberger, M. Pucher, The thought collective behind thirty years of progress in speech synthesis. HSCR19 - The 3rd International Workshop on the History of Speech Communication Research, Vienna, Austria, 2019, 49-58.

Organizing committee

Michael Pucher, Acoustics Research Institute, Austria
Junichi Yamagishi, National Institute of Informatics, Japan
Sébastien Le Maguer, ADAPT Centre/Trinity College Dublin, Ireland
Christian Kaseß, Acoustics Research Institute, Austria
Friedrich Neubarth, Austrian Research Center for Artificial Intelligence

Advisory Committee

Gérard Bailly, GIPSA-lab, France
Antonio Bonafonte, Amazon, UK
Nick Campbell, Trinity College Dublin, Ireland
Julia Hirschberg, Columbia University, USA
Simon King, University of Edinburgh, UK
Bernd Möbius, Saarland University, Germany
Kishore Prahallad, Apple, USA
Keiichi Tokuda, Nagoya Institute of Technology, Japan
David Winarsky, Apple, USA

Logo design

Hieu-Thi Luong, National Institute of Informatics, Japan

Abstract book design

Sébastien Le Maguer, ADAPT Centre/Trinity College Dublin, Ireland
Ingmar Steiner, audEERING, Germany

Scientific Committee

Nagaraj Adiga, University of Crete, Greece
Francesc Alías, Ramon Llull University, Spain
Sercan Arik, Google, USA
Gérard Bailly, GIPSA-lab, France
Pallavi Baljekar, Google Brain, UK
Roberto Barra-Chicote, Amazon, UK
Timo Baumann, Universität Hamburg, Germany
Antonio Bonafonte, Amazon, UK
Joao Cabral, ADAPT Centre/Trinity College Dublin, Ireland
Nick Campbell, Trinity College Dublin, Ireland
Rob Clark, Google AI, UK
Alistair Conkie, Apple, USA
Erica Cooper, National Institute of Informatics, Japan
Daniel Erro, Cirrus Logic, Spain
Raul Fernandez, IBM, USA
Tamás Gábor-Csapó, Budapest University, Hungary
Phil Garner, Idiap, Switzerland
Gustav Eje-Henter, KTH, Sweden
Inma Hernáez, University of the Basque Country, Spain
Pierre-Edouard Honnet, Idiap Research Institute, Switzerland
Christian Kaseß, Acoustics Research Institute, Austria
Simon King, University of Edinburgh, UK
Esther Klabbers, ReadSpeaker, Netherlands
Javier Latorre, Amazon, USA
Jaime Lorenzo-Trueba, Amazon, UK
Gwénolé Lecorvé, University of Rennes 1/IRISA, France
Sébastien Le Maguer, ADAPT Centre/Trinity College Dublin, Ireland
Zhenhua Ling, University of Science and Technology of China
Damien Lolive, University of Rennes 1/IRISA, France
Jindrich Matousek, University of West Bohemia, Czech Republic
Thomas Merritt, Amazon, UK
Bernd Möbius, Saarland University, Germany
Eva Navas, University of the Basque Country, Spain
Yamato Ohtani, AI Inc, Japan
Michael Pucher, Acoustics Research Institute, Austria
Tuomo Raitio, Apple, USA
Sam Ribeiro, University of Edinburgh, UK

Srikanth Ronanki, Amazon, UK
Andrew Rosenberg, Google, USA
Ingmar Steiner, audEERING, Germany
Eva Szekely, KTH, Sweden
Shinji Takaki, Nagoya Institute of Technology, Japan
Tomoki Toda, Nagoya University, Japan
Markus Toman, VocalID, USA
Cassia Valentini-Botinhao, University of Edinburgh, UK
Xin Wang, National Institute of Informatics, Japan
Oliver Watts, University of Edinburgh, UK
David Winarsky, Apple, USA
Zhizheng Wu, JD.com, China
Junichi Yamagishi, National Institute of Informatics, Japan
Yi Zhao, National Institute of Informatics, Japan
Heiga Zen, Google, Japan

PROGRAM

Friday, September 20, 2019

SSW Opening

08:00	Registration opens	
09:15	Welcome	

Keynote 1 - Deep learning for speech synthesis [Chair: Xin Wang]

9:30	Aäron van den Oord: *Deep learning for speech synthesis*	1
10:30	Coffee break	

Oral Session 1 - Neural vocoder [Chair: Rob Clark]

10:50	Xin Wang & Junichi Yamagishi: *Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis*	2
	Prachi Govalkar, Johannes Fischer, Frank Zalkow & Christian Dittmar: *A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction*	2
	Keiichiro Oura, Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku & Keiichi Tokuda: *Deep neural network based real-time speech vocoder with periodic and aperiodic inputs*	2
	Qiao Tian, Xucheng Wan & Shan Liu: *Generative Adversarial Network based Speaker Adaptation for High Fidelity WaveNet Vocoder*	3
12:10	Lunch break	

Oral Session 2 - Adaptation [Chair: Oliver Watts]

14:00	Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik & Sachin Kajarekar: *Neural Text-to-Speech Adaptation from Low Quality Public Recordings*	4
	Bastian Schnell & Philip N. Garner: *Neural VTLN for Speaker Adaptation in TTS*	4
	David Álvarez de la Torre, Santiago Pascual de la Puente & Antonio Bonafonte Cávez: *Problem-Agnostic Speech Embeddings for Multi-Speaker Text-to-Speech with SampleRNN*	4
15:00	SynSIG meeting	
15:15	Coffee break	

Poster Session 1 - Voice conversion and multi-speaker TTS [Chair: Qiong Hu]

15:30	Hiroki Kanagawa & Yusuke Ijima: *Multi-Speaker Modeling for DNN-based Speech Synthesis Incorporating Generative Adversarial Networks*	6
	Ivan Himawan, Sandesh Aryal, Iris Ouyang, Shukhan Ng & Pierre Lanchantin: *Speaker Adaptation of Acoustic Model using a Few Utterances in DNN-based Speech Synthesis Systems*	6
	Yuki Saito, Shinnosuke Takamichi & Hiroshi Saruwatari: *DNN-based Speaker Embedding Using Subjective Inter-speaker Similarity for Multi-speaker Modeling in Speech Synthesis*	6
	Wen-Chin Huang, Yi-Chiao Wu, Kazuhiro Kobayashi, Yu-Huai Peng, Hsin-Te Hwang, Patrick Lumban Tobing, Yu Tsao, Hsin-Min Wang & Tomoki Toda: *Generalization of Spectrum Differential based Direct Waveform Modification for Voice Conversion*	7
	Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi & Tomoki Toda: *Statistical Voice Conversion with Quasi-periodic WaveNet Vocoder*	8

Evening program

17:30	Welcome Reception - Gartenpalais Schönborn, Austrian museum of folk life and folk art

Saturday, September 21, 2019

Keynote 2 - Synthesizing animal vocalizations and modelling animal speech
[Chair: Michael Pucher]

Oral 3 - Evaluation and performance [Chair: Cassia Valentini-Botinhao]

Oral 4 - Speech science [Chair: Rasmus Dall]

Evening program

18:00 Dinner - Buschenschank Fuhrgassl-Huber, Stadl, Buschenschank Fuhrgassl-Huber, Neustift am Walde 68, A-1190 Wien

Sunday, September 22, 2019

Keynote 3 - Natural Language Generation: Creating Text [Chair: Erica Cooper]

Oral Session 5 - Language varieties [Chair: Hermant Patil]

Oral Session 6 - Sequence to sequence model [Chair: Heiga Zen]

Poster Session 3 - Prosody [Chair: Petra Wagner]

Slava Shechtman & Alex Sorin:
Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities 33

Tomoya Yanagita, Sakriani Sakti & Satoshi Nakamura:
Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework 33

Closing ceremony

17:00 Closing ceremony

Keynote 1 - Deep learning for speech synthesis

Deep learning for speech synthesis

Aäron van den Oord

With the advent of Deep Learning, Generative Modeling has dramatically improved, almost reaching the point that generated samples cannot be distinguished from real data. WaveNet has shown that it is possible to model high-dimensional audio so well that it can be used for speech synthesis, outperforming the best known methods such as concatenative and vocoder based systems. The main advantage of generative TTS, however, may be the flexibility of these learning-based approaches. The same system that learns to speak English fluently can also be trained for other languages, such as Mandarin, or even synthesize non-voice audio such as music. A single model can learn different speaker voices at once and can switch between them by conditioning on the speaker identity. It can also learn to adapt more quickly to new unseen data, learning new speakers from as little as a few sentences. Finally, generative TTS systems open the door to a wide variety of new applications, such as unsupervised phonetic unit discovery and speech compression.

Oral Session 1 - Neural vocoder

Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis
Xin Wang & Junichi Yamagishi

Neural source-filter (NSF) models are deep neural networks that produce waveforms given input acoustic features. They use dilated-convolution-based neural filter modules to filter sinebased excitation for waveform generation, which is different from WaveNet and flow-based models. One of the NSF models, called harmonic-plus-noise NSF (h-NSF) model, uses separate pairs of source and neural filters to generate harmonic and noise waveform components. It is close to WaveNet in terms of speech quality while being superior in generation speed. The h-NSF model can be improved even further. While h-NSF merges the harmonic and noise components using predefined digital low- and high-pass filters, it is well known that the maximum voice frequency (MVF) that separates the periodic and aperiodic spectral bands are time-variant. Therefore, we propose a new h-NSF model with time-variant and trainable MVF. We parameterize the digital low- and highpass filters as windowed-sinc filters and predict their cut-off frequency (i.e., MVF) from the input acoustic features. Our experiments demonstrated that the new model can predict a good trajectory of the MVF and produce high-quality speech for a text-to-speech synthesis system.

A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction
Prachi Govalkar, Johannes Fischer, Frank Zalkow & Christian Dittmar

In recent years, text-to-speech (TTS) synthesis has benefited from advanced machine learning approaches. Most prominently, since the introduction of the WaveNet architecture, neural vocoders have exhibited superior performance in terms of the naturalness of synthesized speech signals in comparison to traditional vocoders. In this paper, a fair comparison of recent neural vocoders is presented in a signal reconstruction scenario. That means we use such techniques to resynthesize speech waveforms from mel-scaled spectrograms, a compact and generally non-invertible representation of the underlying audio signal. In that context, we conduct listening tests according to the well established MUSHRA standard and compare the attained results to similar studies. Weighing off the perceptual quality to the computational requirements, our findings shall serve as a guideline to both practitioners and researchers in speech synthesis.

Deep neural network based real-time speech vocoder with periodic and aperiodic inputs

Keiichiro Oura, Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku & Keiichi Tokuda

In this paper, we propose a framework for speech synthesis taking both periodic and aperiodic inputs. Recently, a method of modeling speech waveforms directly, called WaveNet [1], was proposed. WaveNet is able to model speech waveforms accurately and is able to generate natural speech directly, so it is being used, particularly as a speech vocoder [2], in various research [3, 4, 5]. However, it has an autoregressive structure that generates speech sample from the sequence of past speech samples, so parallel computation cannot be used for synthesis, and consequently real-time synthesis is not possible. It also uses pitch information as an auxiliary feature, so it is unable to generate waveforms with a pitch outside of the range in the training data [6], and even if a pitch within the range of the training data is specified, a waveform with a different pitch could be generated. To address these issues, we propose a method that uses periodic and aperiodic input signals to generate the speech sample sequence at once. With the proposed method, speech can be generated faster than real-time, and speech waveforms with pitch outside the range of the training data can be generated. We also conducted a subjective evaluation of the naturalness of the speech, which indicated better synthesized speech quality than WaveNet.

Generative Adversarial Network based Speaker Adaptation for High Fidelity WaveNet Vocoder

Qiao Tian, Xucheng Wan & Shan Liu

Although state-of-the-art parallel WaveNet has addressed the issue of real-time waveform generation, there remains problems. Firstly, due to the noisy input signal of the model, there is still a gap between the quality of generated and natural waveforms. Secondly, a parallel WaveNet is trained under a distillation framework, which makes it tedious to adapt a well trained model to a new speaker. To address these two problems, in this paper we propose an end-to-end adaptation method based on the generative adversarial network (GAN), which can reduce the computational cost for the training of new speaker adaptation. Our subjective experiments shows that the proposed training method can further reduce the quality gap between generated and natural waveforms.

Oral Session 2 - Adaptation

Neural Text-to-Speech Adaptation from Low Quality Public Recordings
Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik & Sachin Kajarekar

Neural Text-to-Speech (TTS) synthesis is able to generate highquality speech with natural prosody. However, these systems typically require a large amount of data, preferably recorded in a clean and noise-free environment. We focus on creating target voices from low quality public recordings and our findings show that even with a large amount of data from a specific speaker, it is challenging to train a speaker-dependent neural TTS model. In order to improve the voice quality, while simultaneously reducing the amount of data required, we introduce meta-learning to adapt the neural TTS front-end. We propose three approaches for multi-speaker systems: (1) a lookup-table-based system, (2) a speaker representation derived from the Personalized Hey Siri (PHS) system, and (3) a system with no speaker encoder. Results show that: i) By using a significantly smaller number of target voice recordings, the proposed system based on embeddings trained from the PHS system can generate comparable quality and speaker similarity to the speaker-dependent model trained solely on the target voice. ii) Applying meta-learning to Tacotron can effectively learn a representation of an unseen speaker. iii) For low quality public recordings, the adaptation based on the multi-speaker corpus can generate a cleaner target voice in comparison with the speaker-dependent model.

Neural VTLN for Speaker Adaptation in TTS
Bastian Schnell & Philip N. Garner

Vocal tract length normalisation (VTLN) is well established as a speaker adaptation technique that can work with very little adaptation data. It is also well known that VTLN can be cast as a linear transform in the cepstral domain. Building on this latter property, we show that it can be cast as a (linear) layer in a deep neural network (DNN) for speech synthesis. We show that VTLN parameters can then be trained in the same framework as the rest of the DNN using automatic gradients. Experimental results show that the DNN is capable of predicting phonedependent warpings on artificial data, and that such warpings improve the quality of an acoustic model on real data in subjective listening tests.

Problem-Agnostic Speech Embeddings for Multi-Speaker Text-to-Speech with SampleRNN

David Álvarez de la Torre, Santiago Pascual de la Puente & Antonio Bonafonte Cávez

Text-to-speech (TTS) acoustic models map linguistic features into an acoustic representation out of which an audible waveform is generated. The latest and most natural TTS systems build a direct mapping between linguistic and waveform domains, like SampleRNN. This way, possible signal naturalness losses are avoided as intermediate acoustic representations are discarded. Another important dimension of study apart from naturalness is their adaptability to generate voice from new speakers that were unseen during training. In this paper we first propose the use of problem-agnostic speech embeddings in a multi-speaker acoustic model for TTS based on SampleRNN. This way, we feed the acoustic model with speaker acousticallydependent representations that enrich the waveform generation more than embeddings unrelated to these factors. Our first results suggest that the proposed embeddings lead to better quality voices than those obtained with one-hot embeddings. Furthermore, as we can use any speech segment as an encoded representation during inference, the model is capable to generalize to new speaker identities without retraining the network. We finally show that, with a small increase of speech duration in the embedding extractor, we dramatically reduce the spectral distortion to close the gap towards the target identities.

Poster Session 1 - Voice conversion and multi-speaker TTS

Multi-Speaker Modeling for DNN-based Speech Synthesis Incorporating Generative Adversarial Networks
Hiroki Kanagawa & Yusuke Ijima

This paper presents a novel DNN-based speech synthesis method we derived from multi-speaker training data. In general, speaker-dependent modeling techniques based on generative adversarial networks (GANs) improve synthesized speech quality. However, they are inadequate for multi-speaker training because conventional discriminators cannot take into account speaker identity, which degrades anti-spoofing performance in GAN discriminators. We introduce two approaches as means to learn GAN speaker characteristics, i.e., auxiliary features and tasks. The first uses speaker codes as additional discriminator input. The second uses speaker identification as a means to verify that anti-spoofing verification methods are effective. Experimental results showed that our proposed techniques outperformed conventional and GAN-based methods.

Speaker Adaptation of Acoustic Model using a Few Utterances in DNN-based Speech Synthesis Systems
Ivan Himawan, Sandesh Aryal, Iris Ouyang, Shukhan Ng & Pierre Lanchantin

Synthesizing a person's voice from only a few utterances is a highly desirable feature for personalized text-to-speech systems. This can be achieved by adapting an existing speakerindependent model to a target speaker such that the speaker variabilities due to a mismatch between training and testing conditions are minimized. In deep neural network (DNN) based speech synthesis, directly fine-tuning a large number of parameters is susceptible to over-fitting problem, especially when the adaptation set is small. In this paper, we present a novel technique to estimate a speaker-specific model using a partial copy of the speaker-independent model by creating a separate parallel branch stemmed from the intermediate hidden layer of the base network. This allows the fine-tuning of a speaker-specific model to take into account the difference between the target speaker and a speaker-independent model output. Experimental results show that the proposed adaptation method achieves improved audio quality and higher speaker similarity compared to another DNN speaker adaptation technique.

DNN-based Speaker Embedding Using Subjective Inter-speaker Similarity for Multi-speaker Modeling in Speech Synthesis

Yuki Saito, Shinnosuke Takamichi & Hiroshi Saruwatari

This paper proposes novel algorithms for speaker embedding using subjective inter-speaker similarity based on deep neural networks (DNNs). Although conventional DNN-based speaker embedding such as a d-vector can be applied to multi-speaker modeling in speech synthesis, it does not correlate with the subjective inter-speaker similarity and is not necessarily appropriate speaker representation for open speakers whose speech utterances are not included in the training data. We propose two training algorithms for DNN-based speaker embedding model using an inter-speaker similarity matrix obtained by large-scale subjective scoring. One is based on similarity vector embedding and trains the model to predict a vector of the similarity matrix as speaker representation. The other is based on similarity matrix embedding and trains the model to minimize the squared Frobenius norm between the similarity matrix and the Gram matrix of d-vectors, i.e., the inter-speaker similarity derived from the d-vectors. We crowdsourced the inter-speaker similarity scores of 153 Japanese female speakers, and the experimental results demonstrate that our algorithms learn speaker embedding that is highly correlated with the subjective similarity. We also apply the proposed speaker embedding to multispeaker modeling in DNN-based speech synthesis and reveal that the proposed similarity vector embedding improves synthetic speech quality for open speakers whose speech utterances are unseen during the training.

Generalization of Spectrum Differential based Direct Waveform Modification for Voice Conversion

Wen-Chin Huang, Yi-Chiao Wu, Kazuhiro Kobayashi, Yu-Huai Peng, Hsin-Te Hwang, Patrick Lumban Tobing, Yu Tsao, Hsin-Min Wang & Tomoki Toda

We present a modification to the spectrum differential based direct waveform modification for voice conversion (DIFFVC) so that it can be directly applied as a waveform generation module to voice conversion models. The recently proposed DIFFVC avoids the use of a vocoder, meanwhile preserves rich spectral details hence capable of generating high quality converted voice. To apply the DIFFVC framework, a model that can estimate the spectral differential from the F0 transformed input speech needs to be trained beforehand. This requirement imposes several constraints, including a limitation on the estimation model to parallel training and the need of extra training on each conversion pair, which make DIFFVC inflexible. Based on the above motivations, we propose a new DIFFVC framework based on an F0 transformation in the residual domain. By performing inverse filtering on the input signal followed by synthesis filtering on the F0 transformed residual signal using the converted spectral features directly, the spectral conversion model does not need to be retrained or capable of predicting the spectral differential. We describe several details that need to be taken care of under this modification, and by applying our proposed method to a non-parallel, variational autoencoder (VAE)-based spectral conversion model, we demonstrate that this framework can be generalized to any spectral conversion model, and experimental evaluations show that it can outperform a baseline framework whose waveform generation process is carried out by a vocoder.

Statistical Voice Conversion with Quasi-periodic WaveNet Vocoder
Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi & Tomoki Toda

In this paper, we investigate the effectiveness of a quasi-periodic WaveNet (QPNet) vocoder combined with a statistical spectral conversion technique for a voice conversion task. The WaveNet (WN) vocoder has been applied as the waveform generation module in many different voice conversion frameworks and achieves significant improvement over conventional vocoders. However, because of the fixed dilated convolution and generic network architecture, the WN vocoder lacks robustness against unseen input features and often requires a huge network size to achieve acceptable speech quality. Such limitations usually lead to performance degradation in the voice conversion task. To overcome this problem, the QPNet vocoder is applied, which includes a pitch-dependent dilated convolution component to enhance the pitch controllability and attain a more compact network than the WN vocoder. In the proposed method, input spectral features are first converted using a framewise deep neural network, and then the QPNet vocoder generates converted speech conditioned on the linearly converted prosodic and transformed spectral features. The experimental results confirm that the QPNet vocoder achieves significantly better performance than the same-size WN vocoder while maintaining comparable speech quality to the double-size WN vocoder.

Voice Conversion without Explicit Separation of Source and Filter Components Based on Non-negative Matrix Factorization
Hitoshi Suda, Daisuke Saito & Nobuaki Minematsu

This paper introduces a new voice conversion (VC) technique which performs spectrogram-to-spectrogram conversion. Conventional studies on VC focus on spectral envelopes, which represent vocal tract information. While vocoders have enabled light-weight and high-quality synthesis from the features, flexibility and quality is still limited by parameterization. To overcome the limitation, this paper aims to model and convert spectrograms themselves. In general, spectrograms are too complicated to be modeled because they contain not only spectral envelopes but also source structures. This paper adopts source-filter non-negative matrix factorization (SF-NMF) as a conversion model of spectrograms. SF-NMF is an extended model of non-negative matrix factorization (NMF), and models source and filter components jointly without explicit separation. The proposed method generates waveforms by reconstructing phase information from amplitude spectrograms. Since SFNMF requests log-frequency spectrograms, the method utilizes scalograms, which are obtained by continuous wavelet transform (CWT). Experimental results showed the proposed method achieved spectrogram-to-spectrogram speaker conversion.

Voice conversion based on full-covariance mixture density networks for time-variant linear transformations
Gaku Kotani & Daisuke Saito

This paper integrates a density estimation scheme based on neural networks with voice conversion (VC) under constraints of time-variant linear transformation. In VC, deep neural networks (DNNs) are used as conversion models that represent mapping from source to target features, in which a stack of multiple nonlinear transformations is applied to source ones. In automatic speech recognition and text-to-speech synthesis, direct mapping between source and target features by DNNs works effectively and flexibly since DNNs are suitable for such tasks in which input and output feature domains are heterogeneous, i.e. speech-to-text or text-to-speech. On the other hand, the case of VC is different from them, i.e. input and output features usually exist on the same domain, such as cepstral space. This condition may help more effective and flexible DNN-based VC. From this point of view, VC based on DNNs for time-variant linear transformations has been suggested. The method can utilize the condition, in which a trained model outputs parameters of linear transformations for each time index t: a linear transformation matrix At and a bias vector bt . It was observed that the method improved the performance of VC. However, the detailed properties of At and bt have still been obscure. In this paper, in order to reveal it, full-covariance mixture density networks are introduced to the VC framework. In the proposed method, joint density of source and target features is directly estimated from the source features by mixture density networks. From the help of tight relationship between Gaussian and linear transformation, the correspondence between the parameters At and bt , and density of the feature space become clear. The proposed scheme was investigated by experiments of VC, and the results showed that naturalness improvement was observed compared with naive DNN-based VC and the decided correspondence between At and bt was observed.

Unsupervised Learning of a Disentangled Speech Representation for Voice Conversion
Tobias Gburrek, Thomas Glarner, Janek Ebbers, Reinhold Haeb-Umbach & Petra Wagner

This paper presents an approach to voice conversion, which does neither require parallel data nor speaker or phone labels for training. It can convert between speakers which are not in the training set by employing the previously proposed concept of a factorized hierarchical variational autoencoder. Here, linguistic and speaker induced variations are separated upon the notion that content induced variations change at a much shorter time scale, i.e., at the segment level, than speaker induced variations, which vary at the longer utterance level. In this contribution we propose to employ convolutional instead of recurrent network layers in the encoder and decoder blocks, which is shown to achieve better phone recognition accuracy on the latent segment variables at frame-level due to their better temporal resolution. For voice conversion the mean of the utterance variables is replaced with the respective estimated mean of the target speaker. The resulting log-mel spectra of the decoder output are used as local conditions of a WaveNet which is utilized for synthesis of the speech waveforms. Experiments show both good disentanglement properties of the latent space variables, and good voice conversion performance.

Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device
Riku Arakawa, Shinnosuke Takamichi & Hiroshi Saruwatari

Voice conversion (VC) enables us to change speech while preserving the linguistic information and is expected to play a significant role in augmented human communication. Recently, deep neural network (DNN)-based VC has been attracting attention because it can synthesize high-quality speech. However, existing methods typically assume offline processes (i.e., analysis, conversion, and synthesis) and cannot be directly applied to real-time VC. Therefore, we propose an implementation method of DNN-based VC that works online with low latency. We also propose audio data augmentation to improve the speech quality of real-time VC. Finally, we develop a maskbased real-time VC device to improve robustness against background noise. Experimental results demonstrate that 1) the proposed real-time VC works with 0.50 of the real-time factor, 2) the proposed data augmentation improves speech quality, and 3) the proposed mask-based VC device is more robust to noise than a standard microphone-based VC device.

Keynote 2 - Synthesizing animal vocalizations and modelling animal speech

---

Synthesizing animal vocalizations and modelling animal speech
Tecumseh Fitch & Bart de Boer

In the last two decades, theory from speech science and methods from digital signal processing have been productively used to study animal communication in many different ways. This has led to fundamental advances in our understanding of how animals produce and perceive their vocalizations, and use them to communicate with one another. A central insight was that the source-filter theory of vocal production, initially developed in speech science, applies to most vertebrate vocal systems as well. This opened the door to using methods like linear prediction to analyze source and filter characteristics, and to re-synthesize realistic vocalizations with precise changes to fundamental frequency, formants and other characteristics. We give an overview of this progress, with several specific examples from our own work covered in more detail.

Oral 3 - Evaluation and performance

Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs
Rob Clark, Hanna Silen, Tom Kenter & Ralph Leith

Text-to-speech systems are typically evaluated on single sentences. When long-form content, such as data consisting of full paragraphs or dialogues is considered, evaluating sentences in isolation is not always appropriate as the context in which the sentences are synthesized is missing. In this paper, we investigate three different ways of evaluating the naturalness of long-form text-to-speech synthesis. We compare the results obtained from evaluating sentences in isolation, evaluating whole paragraphs of speech, and presenting a selection of speech or text as context and evaluating the subsequent speech. We find that, even though these three evaluations are based upon the same material, the outcomes differ per setting, and moreover that these outcomes do not necessarily correlate with each other. We show that our findings are consistent between a single speaker setting of read paragraphs and a two-speaker dialogue scenario. We conclude that to evaluate the quality of long-form speech, the traditional way of evaluating sentences in isolation does not suffice, and that multiple evaluations are required.

Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program
Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander & Jana Voße

Speech synthesis applications have become an ubiquity, in navigation systems, digital assistants or as screen or audio book readers. Despite their impact on the acceptability of the systems in which they are embedded, and despite the fact that different applications probably need different types of TTS voices, TTS evaluation is still largely treated as an isolated problem. Even though there is strong agreement among researchers that the mainstream approaches to Text-to-Speech (TTS) evaluation are often insufficient and may even be misleading, there exist few clear-cut suggestions as to (1) how TTS evaluations may be realistically improved on a large scale, and (2) how such improvements may lead to an informed feedback for system developers and, ultimately, better systems relying on TTS. This paper reviews the current state-of-the-art in TTS evaluation, and suggests a novel user-centered research program for this area.

Rakugo speech synthesis using segment-to-segment neural transduction and style tokens — toward speech synthesis for entertaining audiences
Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki & Junichi Yamagishi

We have been working on constructing rakugo speech synthesis as a challenging example of speech synthesis that entertains audiences. Rakugo is a traditional Japanese form of verbal entertainment that is similar to one-person stand-up comedy. In rakugo, a performer himself/herself plays multiple characters, and conversations by them make the story progress. We tried to build a rakugo synthesizer with state-of-the-art encoder-decoder models with attention such as Tacotron 2. However, it did not work well because the expressions of rakugo speech are far more diverse than those of read speech. We therefore use segment-to-segment neural transduction (SSNT) in place of a combination of attention and decoder. Furthermore, we experimented with global style tokens (GST) and manually-labeled context features to enrich the speaking styles of synthesized rakugo speech. The results show that SSNT greatly helps align the encoder and decoder time steps and that GST help reproduce characteristics better.

Voice Puppetry: Exploring Dramatic Performance to Develop Speech Synthesis
Matthew Aylett, David Braude, Christopher Pidcock & Blaise Potard

Technology and innovation is often inspired by nature. However, when technology enters the social domain, such as creating human-like voices or having human-like conversations, mimicry can become an objective rather than an inspiration. In this paper we argue that performance and acting can offer a radically different design agenda to the mimicry objective. We compare a human mimic's vocal performance (Alec Baldwin) of a target voice (Donald Trump) with the synthesis and copy resynthesis of a cloned synthetic voice. We show the conversational speaking style of natural performance is still a challenge to recreate with modern synthesis methods, and that resynthesis is hampered by current limitations in speech alignment approaches. We conclude by discussing how voice puppetry where a human voice is used to drive a synthesis engine - could be used to advance the state-of-the-art and the challenges involved in developing a voice puppetry system.

Oral 4 - Speech science

Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis
Avashna Govender, Cassia Valentini-Botinhao & Simon King

Listening to even high quality text-to-speech - such as that generated by a Deep Neural Network (DNN) driving a vocoder - still requires greater cognitive effort than natural speech, under noisy conditions. Vocoding itself, plus errors in predictions of the vocoder speech parameters by the DNN model are assumed to be responsible. To better understand the contribution of each parameter, we construct a range of systems that vary from copysynthesis (i.e., vocoding) to full text-to-speech generated using a Deep Neural Network system. Each system combines some speech parameters (e.g., spectral envelope) from copy-synthesis with other speech parameters (e.g., F0) predicted from text. Cognitive load was measured using a pupillometry paradigm described in our previous work. Our results reveal the differing contributions that each predicted speech parameter makes to increasing cognitive load.

Statistical parametric synthesis of budgerigar songs
Lorenz Gutscher, Michael Pucher, Carina Lozo, Marisa Hoeschele & Daniel Mann

In this paper we present the synthesis of budgerigar songs with Hidden Markov Models (HMMs) and the HMM-based Speech Synthesis System (HTS). Budgerigars can produce complex and diverse sounds that are difficult to categorize. We adapted techniques that are commonly used in the area of speech synthesis so that we can use them for the synthesis of budgerigar songs. To segment the recordings, the songs are broken down into phrases, which are sounds separated by silence. Complex phrases furthermore can be subdivided into smaller units and then be clustered to identify recurring elements. These element categories along with additional contextual information are used together to enhance the training and synthesis. Overall, the aim of the process is to offer an interface that generates new sequences and compositions of bird songs based on user input, consisting of the desired song structure and contextual information. Finally, an objective evaluation comparing the synthesized output to the natural recording is performed, and a subjective evaluation with human listeners shows that they prefer resynthesized over natural recordings and that they perceive no significant differences in terms of naturalness between natural, resynthesized, and synthesized versions1 .

GlottDNN-based spectral tilt analysis of tense voice emotional styles for the expressive 3D numerical synthesis of vowel [a]

Marc Freixes, Marc Arnela, Francesc Alías & Joan Claudi Socoró

Three-dimensional (3D) acoustic models allow for an accurate modelling of acoustic wave propagation in 3D realistic vocal tracts. However, voice generated by these approaches is still limited in terms of expressiveness, which could be improved through proper modifications of the glottal source excitation. This work aims at adding some expressiveness to a 3D numerical synthesis approach based on the Finite Element Method (FEM) that uses as input an LF (Liljencrants-Fant) model controlled by the glottal shape parameter Rd . To that effect, a parallel Spanish speech corpus containing neutral and tense voice emotional styles is analysed with the GlottDNN vocoder, obtaining F0 and spectral tilt parameters associated with the glottal excitation. The variations of these two parameters are computed for happy and aggressive styles with reference to neutral speech, differentiating between stressed and unstressed vowels [a]. From this analysis, F0 and Rd values are then derived and used in the LF-FEM based synthesis of vowels [a] to resemble the aforementioned expressive styles. Results show that it is necessary to increase F0 and decrease Rd with respect to neutral speech, with larger deviations for happy than aggressive style, especially for the stressed vowels.

Poster Session - Applications

Preliminary guidelines for the efficient management of OOV words for spoken text

Christina Tånnander & Jens Edlund

We investigate the practical short-term and long-term effects of five different frequency ranks used for selecting which out-ofvocabulary (OOV) words to add to a pronunciation lexicon for text-to-speech (TTS) of university textbooks. The work is an empirical study on a corpus of 200 university text books selected for talking book production and it takes the extensive pronunciation lexicon of a commercial text-to-speech system as its baseline. The main take-home message is a short but succinct set of guidelines that promise to increase the efficiency of OOV management, at least for text-to-speech production of university text books.

Loss Function Considering Temporal Sequence for Feed-Forward Neural Network–Fundamental Frequency Case

Noriyuki Matsunaga, Yamato Ohtani & Tatsuya Hirahara

This paper describes a novel loss function for training feedforward neural networks (FFNNs), which can generate smooth speech parameter sequences without post-processing. In statistical parametric speech synthesis based on deep neural networks (DNNs), maximum likelihood parameter generation (MLPG) or recurrent neural networks (RNNs) are generally used to generate smooth speech parameter sequences. However, because the MLPG process requires utterance-level processing, it is not suitable for speech synthesis requiring low latency. Furthermore, networks such as long short-term memory RNNs (LSTM-RNNs) have high computational costs. As RNNs are not recommended in limited computational resource situations, we look at employing FFNNs as an alternative. One limitation of FFNNs is that they train to ignore relationships between speech parameters in adjacent frames. To overcome this limitation and generate smooth speech parameter sequences from FFNNs alone, we propose a novel loss function that uses long- and short-term features from speech parameters. We evaluated the proposed loss function with a focus on the fundamental frequency (F0) at found that, using the proposed loss function, an FFNN-only approach can generate F0 contours that are perceptually equal to or better in terms of naturalness than those generated by MLPG or LSTM-RNNs.

Sparse Approximation of Gram Matrices for GMMN-based Speech Synthesis
Tomoki Koriyama, Shinnosuke Takamichi & Takao Kobayashi

This paper discusses a training method of speech synthesis framework using generative moment matching network (GMMN). GMMN is a deep generative model optimized by minimizing conditional maximum mean discrepancy (CMMD), and the GMMN-based speech synthesis system models the distribution of acoustic features. Although CMMD is computationally infeasible for a large amount of data, the reduction methods of computation complexity were not examined in the previous study. In this paper, we propose an approximation method based on random Fourier features (RFFs) and minibatch selection technique using K-means clustering. Experimental evaluations show that the proposed method outperformed the conventional one in the perception of inter-utterance variation.

Speaker Anonymization Using X-vector and Neural Waveform Models
Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans & Jean-Francois Bonastre

The social media revolution has produced a plethora of web services to which users can easily upload and share multimedia documents. Despite the popularity and convenience of such services, the sharing of such inherently personal data, including speech data, raises obvious security and privacy concerns. In particular, a user's speech data may be acquired and used with speech synthesis systems to produce high-quality speech utterances which reflect the same user's speaker identity. These utterances may then be used to attack speaker verification systems. One solution to mitigate these concerns involves the concealing of speaker identities before the sharing of speech data. For this purpose, we present a new approach to speaker anonymization. The idea is to extract linguistic and speaker identity features from an utterance and then to use these with neural acoustic and waveform models to synthesize anonymized speech. The original speaker identity, in the form of timbre, is suppressed and replaced with that of an anonymous pseudo identity. The approach exploits state-of-the-art x-vector speaker representations. These are used to derive anonymized pseudo speaker identities through the combination of multiple, random speaker x-vectors. Experimental results show that the proposed approach is effective in concealing speaker identities. It increases the equal error rate of a speaker verification system while maintaining high quality, anonymized speech.

V2S attack: building DNN-based voice conversion from automatic speaker verification

Taiki Nakamura, Yuki Saito, Shinnosuke Takamichi, Yusuke Ijima & Hiroshi Saruwatari

This paper presents a new voice impersonation attack using voice conversion (VC). Enrolling personal voices for automatic speaker verification (ASV) offers natural and flexible biometric authentication systems. Basically, the ASV systems do not include the users' voice data. However, if the ASV system is unexpectedly exposed and hacked by a malicious attacker, there is a risk that the attacker will use VC techniques to reproduce the enrolled user's voices. We name this the "verification-to-synthesis (V2S) attack" and propose VC training with the ASV and pre-trained automatic speech recognition (ASR) models and without the targeted speaker's voice data. The VC model reproduces the targeted speaker's individuality by deceiving the ASV model and restores phonetic property of an input voice by matching phonetic posteriorgrams predicted by the ASR model. The experimental evaluation compares converted voices between the proposed method that does not use the targeted speaker's voice data and the standard VC that uses the data. The experimental results demonstrate that the proposed method performs comparably to the existing VC methods that trained using a very small amount of parallel voice data.

Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis

Takato Fujimoto, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku & Keiichi Tokuda

We investigate the impact of input linguistic feature representation on Japanese end-to-end speech synthesis. An end-toend speech synthesis system, which directly generates natural speech from text, has recently been proposed. The English endto-end system Tacotron 2 achieves sound quality close to that of natural speech. However, unlike alphabetic language that use stress accent, such as English and Spanish, it is difficult to achieve end-to-end speech synthesis with other non-alphabetic languages (e.g., Japanese and Chinese, which use pitch accent and tone, respectively, and use ideograms). We investigated the units of an input sequence, contexts, pause insertion, vowel devoicing, and pronunciation of particles for Japanese end-to-end speech synthesis. Experimental results indicate improvement in the naturalness of the synthesized speech using high or low accents. The results also indicate that the accent-phrase information can help to predict pause insertion, and an end-to-end text-to-speech model may be able to change the pronunciation for devoiced vowels and particles.

Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis

Takato Fujimoto, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku & Keiichi Tokuda

We investigate the impact of input linguistic feature representation on Japanese end-to-end speech synthesis. An end-toend speech synthesis system, which directly generates natural speech from text, has recently been proposed. The English endto-end system Tacotron 2 achieves sound quality close to that of natural speech. However, unlike alphabetic language that use stress accent, such as English and Spanish, it is difficult to achieve end-to-end speech synthesis with other non-alphabetic languages (e.g., Japanese and Chinese, which use pitch accent and tone, respectively, and use ideograms). We investigated the units of an input sequence, contexts, pause insertion, vowel devoicing, and pronunciation of particles for Japanese end-to-end speech synthesis. Experimental results indicate improvement in the naturalness of the synthesized speech using high or low accents. The results also indicate that the accent-phrase information can help to predict pause insertion, and an end-to-end text-to-speech model may be able to change the pronunciation for devoiced vowels and particles.

Low computational cost speech synthesis based on deep neural networks using hidden semi-Markov model structures
Motoki Shimada, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku & Keiichi Tokuda

We propose a method of changing the units of input features from states used conventionally to phonemes and moras to reduce the computational cost of deep neural networks (DNNs) with a hidden semi-Markov model structure for speech synthesis, which can model acoustic features and a temporal structure in a unified framework. Neural networks with very deep and wide structures have recently been applied successfully in the field of speech synthesis. However, such models have very high computational cost, so they are not being applied on platforms with limited resources. To solve this problem, we increased the length of time of DNN input units. We used phoneme or mora units, which are longer than the state units used conventionally. Increasing the length in time of units of input features reduces the number of DNN forward propagations required for speech synthesis, reducing the computational cost. Since a mora in Japanese exhibits isochronism, the duration can be represented more appropriately than the phoneme units expressing consonants and vowels of different lengths with one neural network. Experimental results indicate that compared with speech synthesis based on a DNN with frame inputs, computational cost can be reduced by 97% without degrading the naturalness of the synthesized speech with the proposed method.

Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework
Tomoya Yanagita, Sakriani Sakti & Satoshi Nakamura

Real-time machine speech interpreters aim to mimic human interpreters that able to produce high-quality speech translations on the fly. It requires all system components, including speech recognition, machine translation, and text-to-speech (TTS), to perform incrementally before the speaker has spoken an entire sentence. For TTS, this poses problems as a standard framework commonly requires language-dependent contextual linguistics of a full sentence to produce a natural-sounding speech waveform. Existing studies of incremental TTS (iTTS) have mainly been conducted on a model based on hidden Markov model (HMM). Recently, end-to-end TTS based on a neural net has synthesized more natural speech than HMM-based systems. In this paper, we take an initial step to construct iTTS based on end-to-end neural framework (Neural iTTS) and investigate the effects of various incremental units on the quality of end-to-end neural speech synthesis in both English and Japanese.

Keynote 3 - Natural Language Generation: Creating Text

Natural Language Generation: Creating Text

Claire Gardent

Natural Language Generation (NLG) aims at creating text based on some input (data, text, meaning representation) and some communicative goal (summarising, verbalising, comparing etc.). In the pre-neural era, differing input types and communicative goals led to distinct computational models. In contrast, deep learning encoder-decoder models introduced a shift of paradigm in that they provide a unifying framework for all NLG tasks. In my talk, I will start by briefly introducing the three main types of input considered in NLG. I will then give an overview of how neural models handle these and present some of the work we did on generating text from meaning representations, from data and from text.

Oral Session 5 - Language varieties

Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN
Aye Mya Hlaing, Win Pa Pa & Ye Kyaw Thu

Recently, Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) has become an attractive architecture in speech synthesis for its ability to learn long time-dependencies. Contextual linguistic information is an important feature for naturalness in speech synthesis and using that feature in various speech synthesis models improves the quality of the synthesized speeches for languages. In this paper, LSTM-RNN was applied in Myanmar speech synthesis, and the importance of contextual linguistic features and the effect of applying explicit tone information in different architectures of LSTM-RNN was examined using our proposed Myanmar question set. Experiments of LSTM-RNN, and a hybrid system of DNN and LSTM-RNN, i.e., four feedforward hidden layers followed by two LSTMRNN layers, were done on Myanmar speech synthesis and compared with the baseline DNN. Both objective and subjective evaluations show that the hybrid of DNN and LSTM-RNN system gives more satisfiable synthesized speeches for Myanmar language than the LSTM-RNN and baseline DNN systems.

Building Multilingual End-to-End Speech Synthesisers for Indian Languages
Anusha Prakash, Anju Leela Thomas, Umesh S & Hema A Murthy

Building text-to-speech (TTS) synthesisers is a difficult task, especially for low resource languages. Language-specific modules need to be developed for system building. End-to-end speech synthesis has become a popular paradigm as a TTS can be trained using only <text, audio> pairs. However, end-to-end speech synthesis is not scalable in a multilanguage scenario, as the vocabulary increases with the number of different scripts. In this paper, TTSes are trained for Indian languages using two text representations– character-based and phone-based. For the character-based approach, a multi-language character map (MLCM) is proposed to easily train Indic speech synthesisers. The phone-based approach uses the common label set (CLS) representation for Indian languages. Both approaches leverage the similarities that exist among the languages. The advantage is a compact representation across multiple languages. Experiments are conducted by building TTSes using monolingual data and by pooling data across two languages. The ability to synthesise code-mixed text using the phone-based approach is also assessed. Subjective evaluations indicate that reasonably good quality Indic TTSes can be developed using both approaches. This emphasises the need to incorporate multilingual text processing in the end-to-end framework.

Diphthong interpolation, phone mapping, and prosody transfer for speech synthesis of similar dialect pairs
Michael Pucher, Carina Lozo, Philip Vergeiner & Dominik Wallner

Dialect synthesis is a challenging area of research and contrasts the synthesis of standard varieties not only as to the non standard nature of dialects but also in collecting proper data. In this paper we describe a diphthong interpolation and phone mapping based method that can be used to synthesize a new dialect with an existing dialect model of a similar dialect. The method only uses transcriptions of original dialect data, which are then mapped onto the phones in the model. We improve the basic mapping model further by transferring prosodic features such as original duration and F0. In addition to prosody transfer we want to investigate, if interpolation between two diphthong parts can substitute satisfactorily a missing phone in the target dialect. The methods are applied to two South-Bavarian dialects from Tyrol in Austria.

Subset Selection, Adaptation, Gemination and Prosody Prediction for Amharic Text-to-Speech Synthesis
Elshadai Tesfaye Biru, Yishak Tofik Mohammed, David Tofu, Erica Cooper & Julia Hirschberg

While large TTS corpora exist for commercial systems created for high-resource languages such as Mandarin, English, and Spanish, for many languages such as Amharic, which are spoken by millions of people, this is not the case. We are working with "found" data collected for other purposes (e.g. training ASR systems) or available on the web (e.g. news broadcasts, audiobooks) to produce TTS systems for low-resource languages which do not currently have expensive, commercial systems. This study describes TTS systems built for Amharic from "found" data and includes systems built from different acoustic-prosodic subsets of the data, systems built from combined high and lower quality data using adaptation, and systems which use prediction of Amharic gemination to improve naturalness as perceived by evaluators.

Oral Session 6 - Sequence to sequence model

Initial investigation of encoder-decoder end-to-end TTS using marginalization of monotonic hard alignments

Yusuke Yasuda, Xin Wang & Junichi Yamagishi

End-to-end text-to-speech (TTS) synthesis is a method that directly converts input text to output acoustic features using a single network. A recent advance of end-to-end TTS is due to a key technique called attention mechanisms, and all successful methods proposed so far have been based on soft attention mechanisms. However, although network structures are becoming increasingly complex, end-to-end TTS systems with soft attention mechanisms may still fail to learn and to predict accurate alignment between the input and output. This may be because the soft attention mechanisms are too flexible. Therefore, we propose an approach that has more explicit but natural constraints suitable for speech signals to make alignment learning and prediction of end-to-end TTS systems more robust. The proposed system, with the constrained alignment scheme borrowed from segment-to-segment neural transduction (SSNT), directly calculates the joint probability of acoustic features and alignment given an input text. The alignment is designed to be hard and monotonically increase by considering the speech nature, and it is treated as a latent variable and marginalized during training. During prediction, both the alignment and acoustic features can be generated from the probabilistic distributions. The advantages of our approach are that we can simplify many modules for the soft attention and that we can train the end-to-end TTS model using a single likelihood function. As far as we know, our approach is the first end-to-end TTS without a soft attention mechanism.

Where do the improvements come from in sequence-to-sequence neural TTS?
Oliver Watts, Gustav Eje Henter, Jason Fong & Cassia Valentini-Botinhao

Sequence-to-sequence neural networks with attention mechanisms have recently been widely adopted for text-to-speech. Compared with older, more modular statistical parametric synthesis systems, sequence-to-sequence systems feature three prominent innovations: 1) They replace substantial parts of traditional fixed front-end processing pipelines (like Festival's) with learned text analysis; 2) They jointly learn to align text and speech and to synthesise speech audio from text; 3) They operate autoregressively on previously-generated acoustics. Naturalness improvements have been reported relative to earlier systems which do not contain these innovations. It would be useful to know how much each of the various innovations contribute to the improved performance. We here propose one way of associating the separately-learned components of a representative older modular system, specifically Merlin, with the different sub-networks within recent neural sequence-to-sequence architectures, specifically Tacotron 2 and DCTTS. This allows us to swap in and out various components and subnets to produce intermediate systems that step between the two paradigms; subjective evaluation of these systems then allows us to isolate the perceptual effects of the various innovations. We report on the design, evaluation, and findings of such an experiment.

A Comparison of Letters and Phones as Input to Sequence-to-Sequence Models for Speech Synthesis

Jason Fong, Jason Taylor, Korin Richmond & Simon King

Neural sequence-to-sequence (S2S) models for text-tospeech synthesis (TTS) may take letter or phone input sequences. Since for many languages phones have a more direct relationship to the acoustic signal, they lead to improved quality. But generating phone transcriptions from text requires an expensive dictionary and an error-prone grapheme-to-phoneme (G2P) model, and the relative improvement over using letters has yet to be quantified. In approaching this question, we presume that letter-input S2S models must implicitly learn an internal counterpart to G2P conversion and therefore inevitably make errors. Such a model may thus be viewed as phone-input S2S with inaccurate phone input. To quantify this inaccuracy, we compare in this paper a letter-input S2S system to several phone-input systems trained on data with a varying level of error in the phonetic transcription. Our findings show our letterinput system is equivalent in quality to the phone-input system in which 25% of word tokens in the training data have incorrect phonetic transcriptions. Furthermore, we find that for phoneinput systems up to 15% of word tokens in the training data can have incorrect phonetic transcriptions without any significant difference in performance to a 0% error rate system. This suggests it is acceptable to use G2P to predict pronunciations for out-of-vocabulary words (OOVs) provided they are less than around 15% of the training data, removing the need to manually add OOVs to the dictionary for every new training set.

Poster Session 3 - Prosody

Generative Modeling of F0 Contours Leveraged by Phrase Structure and Its Application to Statistical Focus Control
Yuma Shirahata, Daisuke Saito & Nobuaki Minematsu

In this paper, we propose a statistical generative model of fundamental frequency (F0) contours that incorporates a phrase structure of Japanese ("bunsetsu"), and apply this model to control of the focus point in a sentence. Fujisaki model is a mathematical model that formulates F0 contours as the superposition of phrase and accent components, considering the control mechanism of vocal fold vibration. In the Fujisaki model, model parameters are closely related to linguistic information. Thus, flexible and interpretable conversion of F0 contours corresponding to linguistic information is achieved by changing the model parameters. Recently, a method of treating the Fujisaki model as a stochastic model has been proposed. In this method, the model parameters are inferred from observed F0 contours by a maximum likelihood manner. However, since there are no constraints of linguistic information in inference, unnatural parameters are occasionally estimated. In the proposed method, occurrence of phrase commands is linked to the boundaries of bunsetsu, and then the Fujisaki model parameters and phrase structure correspond to each other. It enables simultaneous modeling of two different F0 contours in every bunsetsu unit. The proposed modeling can be applied to pairs of neutral and focused utterances, and it enables bunsetsu-by-bunsetsu focus control . Experimental results show that the proposed method achieved reasonable control of focus in 74% accuracy rate compared with natural speech. Though there is room for improvement in naturalness, the proposed scheme achieves interpretable conversion of prosody.

Subword tokenization based on DNN-based acoustic model for end-to-end prosody generation
Masashi Aso, Shinnosuke Takamichi, Norihiro Takamune & Hiroshi Saruwatari

This paper presents a method for determining subword units for end-to-end prosody generation. End-to-end prosody generation using deep neural networks (DNNs) is expected to directly generate a prosody sequence from text without any professional knowledge in the target language. In natural language processing, language model-based language-independent subword tokenization was previously proposed for determining subwords suitable for end-to-end language processing. However, the subwords determined by the language models are not appropriate for end-to-end speech processing. In this paper, we propose a language-independent algorithm for determining subwords that maximize acoustic model likelihoods. The proposed algorithm iterates expectation-maximization (EM)-based training of DNN acoustic models and likelihood-based construction of the subword vocabulary. In the experimental evaluation, we discuss the stability of the EM-based training and analyze subword vocabularies determined by the conventional language model-based and proposed acoustic model-based methods.

Using generative modelling to produce varied intonation for speech synthesis
Zack Hodari, Oliver Watts & Simon King

Unlike human speakers, typical text-to-speech (TTS) systems are unable to produce multiple distinct renditions of a given sentence. This has previously been addressed by adding explicit external control. In contrast, generative models are able to capture a distribution over multiple renditions and thus produce varied renditions using sampling. Typical neural TTS models learn the average of the data because they minimise mean squared error. In the context of prosody, taking the average produces flatter, more boring speech: an "average prosody". A generative model that can synthesise multiple prosodies will, by design, not model average prosody. We use variational autoencoders (VAE) which explicitly place the most "average" data close to the mean of the Gaussian prior. We propose that by moving towards the tails of the prior distribution, the model will transition towards generating more idiosyncratic, varied renditions. Focusing here on intonation, we investigate the trade-off between naturalness and intonation variation and find that typical acoustic models can either be natural, or varied, but not both. However, sampling from the tails of the VAE prior produces much more varied intonation than the traditional approaches, whilst maintaining the same level of naturalness.

How to train your fillers: uh and um in spontaneous speech synthesis
Éva Székely, Gustav Eje Henter, Jonas Beskow & Joakim Gustafson

Using spontaneous conversational speech for TTS raises questions on how disfluencies such as filled pauses (FPs) should be approached. Detailed annotation of FPs in training data enables precise control at synthesis time; coarse or nonexistent FP annotation, when combined with stochastic attention-based neural TTS, leads to synthesisers that insert these phenomena into fluent prompts on their own accord. In this study we investigate, objectively and subjectively, the effects of FP annotation and the impact of relinquishing control over FPs in a Tacotron TTS system. The training corpus comprised 9 hours of singlespeaker breath groups extracted from a conversational podcast. Systems trained with no or location-only FP annotation were found to reproduce FP locations and types (uh/um) in a pattern broadly similar to that of the corpus. We also studied the effect of FPs on natural and synthetic speech rate and the interchangeability of FP types. Interestingly, subjective tests indicate that synthesiser-predicted FP types from location-only annotation often were preferred over specifying the ground-truth type. In contrast, a more precise annotation, allowing us to focus training on the most fluent parts of the corpus, improved rated naturalness when synthesising fluent speech.

An Investigation of Features for Fundamental Frequency Pattern Prediction in Electrolaryngeal Speech Enhancement

Mohammad Eshghi, Kou Tanaka, Kazuhiro Kobayashi, Hirokazu Kameoka & Tomoki Toda

Despite abundance of research, natural voice restoration after total laryngectomy (i. e., removal of the vocal folds of the larynx), has remained a challenge. A typical way of producing a relatively intelligible speech for patients suffering from this inability is to use an electrolarynx. However, the outcome voice sounds artificial and has "robotic" quality owing to constant fundamental frequency (F0) patterns generated by the electrolarynx. In existing frameworks on natural F0 patterns prediction, a model is trained on a massive amount of parallel training data to find a mapping that maps spectral features of the source speech into F0 contours of the target speech. However, creating big datasets for electrolaryngeal (EL) speech is considered as a cumbersome and expensive task. Moreover, EL speech spectral features are significantly different from spectral features of the normal speech, and therefore, it is not straightforward to effectively use easily available normal speech datasets in training of the model for EL speech. Consequently, the quality of the models could be still low due to the lack of sufficient training data. To address this problem, we investigate F0 pattern prediction based on other features that could be shared between normal speech and EL speech. By using shared input features, we would be to train the prediction model using a large amount of training data. As such features, in this work, we examine F0 prediction accuracy based on phoneme-related features. The findings show that by considering phoneme labels for both vowels and consonants and one-hot encoding of these labels, we are able to predict F0 contours with high correlation coefficients.

PROMIS: a statistical-parametric speech synthesis system with prominence control via a prominence network
Zofia Malisz, Harald Berthelsen, Jonas Beskow & Joakim Gustafson

We implement an architecture with explicit prominence learning via a prominence network in Merlin, a statistical-parametric DNN-based text-to-speech system. We build on our previous results that successfully evaluated the inclusion of an automatically extracted, speech-based prominence feature into the training and its control at synthesis time. In this work, we expand the PROMIS system by implementing the prominence network that predicts prominence values from text. We test the network predictions as well as the effects of a prominence control module based on SSML-like tags. Listening tests for the complete PROMIS system, combining a prominence feature, a prominence network and prominence control, show that it effectively controls prominence in a diagnostic set of target words. It also does not negatively impact the perceived naturalness relative to the baseline when one of the tested tagging methods is used.

Deep Mixture-of-Experts Models for Synthetic Prosodic-Contour Generation
Raul Fernandez

Deep recurrent neural networks have been shown to provide state-of-art performance when generating prosodic contours in a speech-synthesis system. These models benefit from the representational capacity obtained by increased compositionality across many layers. As larger amounts of data become available, larger and deeper architectures can be trained at the expense of obtaining models that are expensive both in terms of computation and latency. In this work we take an alternative approach and divide the learning among an ensemble of experts, each of which is a smaller and/or shallower learner whose predictions are then arbitrated by a switching module that assigns sequences of linguistic features to global, sequence-level posteriors, and uses this information to weigh the members of the ensemble. Compared with a single deep cascaded model, this approach is more parallelizable, and can be exploited to obtain a more efficient model in terms of computation (as measured by overall model-size reduction) and latency (as measured by reduction of parameters by branching). We present an architecture where the cluster assignment and prediction models can be trained simultaneously, and demonstrate such gains in efficiency without sacrificing the perceptual quality of the predictions in a subjective listening test.

Prosody Prediction from Syntactic, Lexical, and Word Embedding Features
Rose Sloan, Syed Sarfaraz Akhtar, Bryan Li, Ritvik Shrivastava, Agustin Gravano & Julia Hirschberg

Accurate prosody prediction from text leads to more natural-sounding TTS. In this work, we employ a new set of features to predict ToBI pitch accent and phrase boundaries from text. We investigate a wide variety of text-based features, including many new syntactic features, several types of word embeddings, co-reference features, LIWC features, and specificity information. We focus our work on the Boston Radio News Corpus, a ToBI-labeled corpus of relatively clean news broadcasts, but also test our classifiers on Audix, a smaller corpus of read news, and on the Columbia Games Corpus, a corpus of conversational speech, in order to test the applicability of our model in cross-corpus settings. Our results show strong performance on both tasks, as well as some promising results for cross-corpus applications of our models.

Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities
Slava Shechtman & Alex Sorin

Modern sequence to sequence neural TTS systems provide close to natural speech quality. Such systems usually comprise a network converting linguistic/phonetic features sequence to an acoustic features sequence, cascaded with a neural vocoder. The generated speech prosody (i.e. phoneme durations, pitch and loudness) is implicitly present in the acoustic features, being mixed with spectral information. Although the speech sounds natural, its prosody realization is randomly chosen and cannot be easily altered. The prosody control becomes an even more difficult task if no prosodic labeling is present in the training data. Recently, much progress has been achieved in unsupervised speaking style learning and generation, however human inspection is still required after the training for discovery and interpretation of the speaking styles learned by the system. In this work we introduce a fully automatic method that makes the system aware of the prosody and enables sentencewise speaking pace and expressiveness control on a continuous scale. While being useful by itself in many applications, the proposed prosody control can also improve the overall quality and expressiveness of the synthesized speech, as demonstrated by subjective listening evaluations. We also propose a novel augmented attention mechanism, that facilitates better pace control sensitivity and faster attention convergence.

Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework

Tomoya Yanagita, Sakriani Sakti & Satoshi Nakamura

Real-time machine speech interpreters aim to mimic human interpreters that able to produce high-quality speech translations on the fly. It requires all system components, including speech recognition, machine translation, and text-to-speech (TTS), to perform incrementally before the speaker has spoken an entire sentence. For TTS, this poses problems as a standard framework commonly requires language-dependent contextual linguistics of a full sentence to produce a natural-sounding speech waveform. Existing studies of incremental TTS (iTTS) have mainly been conducted on a model based on hidden Markov model (HMM). Recently, end-to-end TTS based on a neural net has synthesized more natural speech than HMM-based systems. In this paper, we take an initial step to construct iTTS based on end-to-end neural framework (Neural iTTS) and investigate the effects of various incremental units on the quality of end-to-end neural speech synthesis in both English and Japanese.

INDEX