SSW11

The 11th ISCA Speech Synthesis Workshop

Budapest, Hungary

SSW11 Sponsors

Platinum Sponsor

Google

Gold Sponsors

SAMSUNG

The International Speech Communication Association (ISCA)

MTA

HTEnet Innovációs Nonprofit Kft.

MŰEGYETEM 1782

BME VIK

Dear Participants,

At the time of submitting our bid for SSW11, I hoped that we can re-create the atmosphere of SSW1 in Autrans that was the first international speech conference I could attend in 1990. I enjoyed the single-track approach that allowed everyone to listen to all the presentations. At the age of 31 years, coming from a country where an ordinary citizen until then only once in 3 years was allowed to travel to Western Europe, that conference was a sort of "open door" to the speech synthesis research community. We could have a chat anytime with world-famous researchers (e.g. Ken Stevens). That was when we could start a friendship with one of the organizers, Christian Benoit, whose name is still well-known although he passed away 23 years ago. Fortunately, the other organizer of SSW1 – Gérard Bailly- will be with us on-site. That is one of the reasons why I encourage everyone to come to Budapest in person.

Another important reason is that Hungary was among the first TTS developers of the world. The HungaroVox formant-based TTS system was demonstrated in 1981 by the Hungarian Academy of Sciences. One of the HungaroVox developers - Gábor Olaszy - will be also on-site with us.

Unfortunately, due to the COVID pandemic, our original plans could not be fulfilled. Initially, we wanted to have everyone in the same hotel (just as it was at SSW1 in Autrans) by Lake Velence. Due to travel restrictions and uncertainty, we moved the conference to a hybrid setting, and the location is a wellness hotel in Budapest which is more flexible in terms of lecture room size and reservations. Although Hungary and Budapest are some of the safest places worldwide, it seems that most participants could only afford remote registration. Even with these limitations, we try to provide as much interaction possibility during the conference as possible.

All submitted papers were reviewed by three members of the Scientific Committee. The accepted 40 papers will be presented in oral sessions, where 24 minutes are devoted to the introduction, the talk, and Q&A activity. On each day, we start with an oral session. One-hour keynotes are planned for the middle of the day in Europe so that the audience spanning from Vancouver to Tokyo had a chance to be fully aware. Discussions may be mixed with lunch after the keynotes. In selecting the keynotes, we tried to follow the special topic proposed for SSW11: "*Speech uniqueness and deep learning*". The first plenary speaker is Lior Wolf, who will present the crossroads of speech, singing, and music. Our second keynote speaker, István Winkler, will introduce us into his basic research on the development of the communication

of infants. On the last day, Thomas Drugman will present the latest results on expressive TTS, which is an essential requirement for more extended human-machine speech dialogues.

Authors are encouraged to submit an extended version of the papers to a special issue on speech synthesis of the Infocommunications Journal (www.infocommunications.hu).

Please use the breaks and the session discussions for both private and group forms of exchanging ideas and opinions. The platform of the presentations will be Zoom Webinar. To facilitate discussions, we shall use the Spatial Chat infrastructure.

Those who come in person have the advantage of personal participation at the welcome reception and the social event besides the regular program.

I would like to thank members of the Organizing Committee for all the effort that made SSW11 possible. Both the past and current Synsig Board members have helped a lot during the preparation. Péter Nagy and Mária Tézsla from the Scientific Association for Infocommunications have taken up the burden of organizational and financial administration. Scientific Committee members spent several hours on thoroughly evaluating the papers and give helpful feedback to the authors.

Our sponsors, Google, Apple, Samsung, iFlytek, ISCA, and the Hungarian Academy of Sciences, allowed us to keep registration fees low while providing high-quality services to the audience. I hope that against all the difficulties caused by the COVID pandemic, SSW11 will provide a remarkable contribution to the development of speech synthesis. Although smaller in numbers, but hopefully through intensive interaction with each other and the two keynote speakers who will be in Budapest, on-site participants will have a unique chance. On behalf of our Speech Communication and Smart Interaction Labs of the Budapest University of Technology and Economics, you are all welcome anytime when you visit Budapest.

I hope that one of the young researchers participating at SSW11 will chair SSW26 30 years from now. That would be a remarkable result.

Looking forward to meeting you at SSW11.

Budapest, August 10, 2021.

Géza Németh, Chairman

Organizing committee

Géza Németh	BME TMIT, Hungary
Junichi Yamagishi	National Institute of Informatics Japan, University of Edinburgh, UK
Sébastien Le Maguer	ADAPT Centre/TCD, Ireland
Esther Klabbers	Readspeaker, Netherlands
Mátyás Bartalis	BME TMIT, Hungary
Tamás Gábor Csapó	BME TMIT, Hungary
Bálint Gyires-Tóth	BME TMIT, Hungary
Gábor Olaszy	BME TMIT, Hungary
Csaba Zainkó	BME TMIT, Hungary

Abstract book design

Csaba Zainkó	Budapest University of Technology and Economic
Sébastien Le Maguer	Trinity College Dublin / Adapt Centre

Scientific Committee

Nagaraj Adiga	University of Crete
Gerard Bailly	GIPSA-Lab
Pallavi Baljekar	Google
Timo Baumann	University of Hamburg
Antonio Bonafonte	Universitat Politècnica de Catalunya
Joao Cabral	Trinity College Dublin
Robert Clark	Google
Erica Cooper	National Institute of Informatics
Tamás Gabor Csapo	Budapest University of Technology and Economic
Daniel Erro	Cirrus Logic
Raul Fernandez	IBM
Philip N. Garner	Idiap Research Institute
Balint Gyires-Toth	Budapest University of Technology and Economic
Qiong Hu	Google
Esther Klabbers	ReadSpeaker
Zhen-Hua Ling	University of Science and Technology of China
Sébastien Le Maguer	Trinity College Dublin / Adapt Centre
Jindrich Matousek	University of West Bohemia
Thomas Merritt	Amazon
Bernd Möbius	Saarland University
Eva Navas	University of the Basque Country
Géza Németh	Budapest University of Technology and Economic
Yamato Ohtani	AI Inc.,
Michael Pucher	Acoustics Research Institute
Francesc Alias Pujol	La Salle - Universitat Ramon Llull
Tuomo Raitio	Apple Inc
Manuel Sam Ribeiro	The University of Edinburgh
Andrew Rosenberg	Google
Adriana Stan	Technical University of Cluj-Napoca
Eva Szekely	KTH Royal Institute of Technology
Tomoki Toda	Nagoya University
Markus Toman	Neuratec
Jaime Lorenzo Trueba	Universidad Politecnica de Madrid
Pirros Tsiakoulis	Samsung Electronics
Junichi Yamagishi	The University of Edinburgh
Csaba Zainkó	Budapest University of Technology and Economic
Heiga Zen	Google

PROGRAM

Thursday, August 26, 2021

SSW Opening

08:30	Welcome	

Session 1: Special synthesis problems [Chair: Lior Wolf]

09:00	Sai Sirisha Rallabandi, Babak Naderi & Sebastian Möller: *Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech*	1
	Tamás Gábor Csapó: *Extending Text-to-Speech Synthesis with Articulatory Movement Prediction using Ultrasound Tongue Imaging*	1
	Martin Lenglet, Olivier Perrotin & Gérard Bailly: *Impact of Segmentation and Annotation in French end-to-end Synthesis*	1
	Marc Illa, Bence Mark Halpern, Rob van Son, Laureano Moro-Velazquez & Odette Scharenborg: *Pathological voice adaptation with autoencoder-based voice conversion*	2
	Elijah Gutierrez, Pilar Oplustil-Gallegos & Catherine Lai: *Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm*	3
11:00	Coffee break	

Keynote 1: Expressive Neural TTS [Chair: Erica Cooper]

11:10	Lior Wolf: *Deep Audio Conversion Technologies and Their Applications in Speech, Singing, and Music*	4

	Abdelhamid Ezzerg, Adam Gabrys, Bartosz Putrycz, Daniel Korzekwa, Daniel Saez-Trigueros, David McHardy, Kamil Pokora, Jakub Lachowicz, Jaime Lorenzo-Trueba & Viacheslav Klimkov: *Enhancing audio quality for expressive Neural Text-to-Speech*	9
	Lucas H. Ueda, Paula D. P. Costa, Flavio O. Simoes & Mário U. Neto: *Are we truly modeling expressiveness? A study on expressive TTS in Brazilian Portuguese for real-life application styles*	9
17:20	Afternoon Session discussion	

Friday, August 27, 2021

Session 4: Articulation and Naturalness [Chair: Tamás Gábor Csapó]

Keynote 2: Early Development of Infantile Communication by Sound [Chair: Cassia Valentini Botinhao]

Session 5: Emotion, singing and voice conversion [Chair: Simon King]

Saturday, August 28, 2021

Session 7: Modeling and evaluation [Chair: Gérard Bailly]

Keynote 3: Deep Audio Conversion Technologies and Their Applications in Speech, Singing, and Music [Chair: Gustav Eje Henter]

Session 8: Synthesis and Context [Chair: Thomas Drugman]

SSW closing

15:10 Closing & SynSIG announcement

Session 1: Special synthesis problems

Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech

Sai Sirisha Rallabandi, Babak Naderi & Sebastian Möller

The advent of neural Text-to-Speech (TTS) synthesizers has enhanced the expressivity of synthetic speech in the recent past. However, there is very little work on understanding the acoustic correlates of paralinguistic traits, emotions, speaker attributes and characteristics from synthetic speech. This paper investigates the acoustic correlates of the speaker attributes: likeability, friendliness, and skillfulness. Our study was carried out on the voices derived from the two commercial TTS systems, Amazon Polly (9 voices) and Google TTS engine (10 voices). In our previous study, we performed a crowd-sourcing-based evaluation to collect the subjective ratings for the desired speaker attributes. In this work, we perform the acoustic feature prediction using the backward elimination algorithm. We show that the level of loudness, spectral flux, fundamental frequency, its formant frequencies, and their combinations contribute to the desired speaker attributes. We further combine the ratings of friendliness and likeability to investigate the characteristic, warmth in synthetic speech and correspondingly, skilfullness represents the characteristic, competence.

Extending Text-to-Speech Synthesis with Articulatory Movement Prediction using Ultrasound Tongue Imaging

Tamás Gábor Csapó

In this paper, we present our first experiments in text-toarticulation prediction, using ultrasound tongue image targets. We extend a traditional (vocoder-based) DNN-TTS framework with predicting PCA-compressed ultrasound images, of which the continuous tongue motion can be reconstructed in synchrony with synthesized speech. We use the data of eight speakers, train fully connected and recurrent neural networks, and show that FC-DNNs are more suitable for the prediction of sequential data than LSTMs, in case of limited training data. Objective experiments and visualized predictions show that the proposed solution is feasible and the generated ultrasound videos are close to natural tongue movement. Articulatory movement prediction from text input can be useful for audiovisual speech synthesis or computer-assisted pronunciation training.

Impact of Segmentation and Annotation in French end-to-end Synthesis
Martin Lenglet, Olivier Perrotin & Gérard Bailly

Audio books are commonly used to train text-to-speech models (TTS), as they offer large phonetic content with rather expressive pronunciation, but number and sizes of publicly available audio books corpora differ between languages. Moreover, the quality and accuracy of the available utterance segmentations are debatable. Yet, the impact of segmentation on the output synthesis is not well established. Additionally, utterances are generally used individually, without taking advantage of text level structuring information, even though they influence speaker reading. In this paper, we conduct a multidimensional evaluation of Tacotron2 trained on different segmentations and text level annotations of the same French corpus. We show that both spectrum accuracy and expressiveness depend on the segmentation used. In particular, a shorter segmentation, in addition with the annotation of paragraphs, benefits to spectrum reconstruction at the detriment of phrasing. Multidimensional analysis of mean opinion scores obtained with a MUSHRAexperiment revealed that phrasing was relatively more important than spectrum accuracy in perceptual judgement. This work serves as evidence that particular attention must be given to models evaluation, as well as how to use the training corpus to maximize synthesis characteristics of interest.

Pathological voice adaptation with autoencoder-based voice conversion
Marc Illa, Bence Mark Halpern, Rob van Son, Laureano Moro-Velazquez & Odette Scharenborg

In this paper, we propose a new approach to pathological speech synthesis. Instead of using healthy speech as a source, we customise an existing pathological speech sample to a new speaker's voice characteristics. This approach alleviates the evaluation problem one normally has when converting typical speech to pathological speech, as in our approach, the voice conversion (VC) model does not need to be optimised for speech degradation but only for the speaker change. This change in the optimisation ensures that any degradation found in naturalness is due to the conversion process and not due to the model exaggerating characteristics of a speech pathology. To show a proof of concept of this method, we convert dysarthric speech using the UASpeech database and an autoencoder-based VC technique. Subjective evaluation results show reasonable naturalness for high intelligibility dysarthric speakers, though lower intelligibility seems to introduce a marginal degradation in naturalness scores for mid and low intelligibility speakers compared to ground truth. Conversion of speaker characteristics for low and high intelligibility speakers is successful, but not for mid. Whether the differences in the results for the different intelligibility levels is due to the intelligibility levels or due to the speakers needs to be further investigated.

Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm

Elijah Gutierrez, Pilar Oplustil-Gallegos & Catherine Lai

Text-to-Speech synthesis systems are generally evaluated using Mean Opinion Score (MOS) tests, where listeners score samples of synthetic speech on a Likert scale. A major drawback of MOS tests is that they only offer a general measure of overall quality—i.e., the naturalness of an utterance—and so cannot tell us where exactly synthesis errors occur. This can make evaluation of the appropriateness of prosodic variation within utterances inconclusive. To address this, we propose a novel evaluation method based on the Rapid Prosody Transcription paradigm. This allows listeners to mark the locations of errors in an utterance in real-time, providing a probabilistic representation of the perceptual errors that occur in the synthetic signal. We conduct experiments that confirm that the fine-grained evaluation can be mapped to system rankings of standard MOS tests, but the error marking gives a much more comprehensive assessment of synthesized prosody. In particular, for standard audiobook test set samples, we see that error marks consistently cluster around words at major prosodic boundaries indicated by punctuation. However, for question-answer based stimuli, where we control information structure, we see differences emerge in the ability of neural TTS systems to generate context-appropriate prosodic prominence.

Keynote 1: Expressive Neural TTS

Deep Audio Conversion Technologies and Their Applications in Speech, Singing, and Music
Lior Wolf

Thomas Drugman is a Science Manager in Amazon TTS Research team. He received his PhD in 2011 from the University of Mons, winning the IBM Belgium award for "Best Thesis in Computer Science". His PhD thesis studied the use of glottal source analysis in Speech Processing. He then made a 3-year post-doc on speech/audio analysis for two biomedical applications: trachea-esophageal speech reconstruction and cough detection in chronic respiratory diseases. In 2014, he joined Amazon as a Scientist in the Alexa ASR team. He then transferred to the TTS team in 2016, where he is Science Manager since 2017. He has contributed in making Amazon's Neural TTS more natural and expressive, notably by enriching Alexa's experience with different speaking styles: emotions, newscaster, whispering, etc. His current research interests lie in improving the naturalness and flow of longer synthetic speech interactions. He has about 125 publications in the field of Speech Processing. He got the Interspeech Best Student Paper awards in 2009 and 2014 (as supervisor). He is also member of the IEEE Speech and Language Technical Committee since 2019.

Session 2: Articulation and speech styles

Speech Synthesis from Text and Ultrasound Tongue Image-based Articulatory Input

Tamás Gábor Csapó, László Tóth, Gábor Gosztolya & Alexandra Markó

Articulatory information has been shown to be effective in improving the performance of HMM-based and DNN-based textto- speech synthesis. Speech synthesis research focuses traditionally on text-to-speech conversion, when the input is text or an estimated linguistic representation, and the target is synthesized speech. However, a research field that has risen in the last decade is articulation-to-speech synthesis (with a target application of a Silent Speech Interface, SSI), when the goal is to synthesize speech from some representation of the movement of the articulatory organs. In this paper, we extend traditional (vocoder-based) DNN-TTS with articulatory input, estimated from ultrasound tongue images. We compare text-only, ultrasound-only, and combined inputs. Using data from eight speakers, we show that that the combined text and articulatory input can have advantages in limited-data scenarios, namely, it may increase the naturalness of synthesized speech compared to single text input. Besides, we analyze the ultrasound tongue recordings of several speakers, and show that misalignments in the ultrasound transducer positioning can have a negative effect on the final synthesis performance.

Combining speakers of multiple languages to improve quality of neural voices

Javier Latorre, Charlotte Bailleul, Tuuli Morrill, Alistair Conkie & Yannis Stylianou

In this work, we explore multiple architectures and training procedures for developing a multi-speaker and multi-lingual neural TTS system with the goals of a) improving the quality when the available data in the target language is limited and b) enabling cross-lingual synthesis. We report results from a large experiment using 30 speakers in 8 different languages across 15 different locales. The system is trained on the same amount of data per speaker. Compared to a single-speaker model, when the suggested system is fine tuned to a speaker, it produces significantly better quality in most of the cases while it only uses less than 40

Methods of slowing down speech
Christina Tånnander & Jens Edlund

A slower speaking rate of human or synthetic speech is often requested by for example language learners or people with aphasia or dementia. Slow speech produced by human speakers typically contain a larger number of pauses, and both pauses and speech have longer segment durations than speech produced at a standard or fast speaking rate. This paper presents several methods of prolonging speech. Two speech chunks of about 30 seconds each, read by a professional voice talent at a very slow speaking rate, were used as reference. Seven pairs of stimuli containing the same word sequences were produced, one by the same professional, reading at her standard speaking rate and six by a moderately slow synthetic voice trained on the same human voice. Different combinations of pause insertions and stretching were used to match the total length of the corresponding reference stimulus. Stretching was applied in different proportions to speech and non-speech, and pauses were inserted at punctuations, at certain phrase boundaries, between each word, or by copying the pause locations of the reference reading. 128 crowdsourced listeners evaluated the 16 stimuli. The results show that all manipulated readings are less consistent with expectations of slow speech than the reference, but that the synthesised readings are comparable to stretched human speech. Key factors are the relation between speech and silence and the duration of talkspurts.

Personality in the mix - investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis
Joakim Gustafson, Jonas Beskow & Eva Szekely

Studies on human-human interactions have shown that that the fluency of a speaker influences the perception of personality. Adding fillers and discourse markers can make the speaker seem uncertain, more casual and spontaneous. With recent TTS developments it is now possible to investigate if the same holds for artificial speakers. In a previous experiment, it was shown that local insertion of fillers in a regular TTS voice influenced the perceived personality. In the current study we extend that work in two ways: Firstly, we recreate the English experiment adding a voice trained on spontaneous speech, where adding fillers also has a global effect on the synthesized speech. We also add Swedish read and spontaneous voices. Secondly, for the Swedish voices, we investigate the effect of using a multispeaker model mixing a read speech voice and a spontaneous speech voice when generating disfluent synthetic speech.

Adaptation of Tacotron2-based Text-To-Speech for Articulatory-to-Acoustic Mapping using Ultrasound Tongue Imaging

Csaba Zainkó, László Tóth, Amin Honarmandi Shandiz, Gábor Gosztolya, Alexandra Markó, Géza Németh & Tamás Gábor Csapó

For articulatory-to-acoustic mapping, typically only limited parallel training data is available, making it impossible to apply fully end-to-end solutions like Tacotron2. In this paper, we experimented with transfer learning and adaptation of a Tacotron2 text-to-speech model to improve the final synthesis quality of ultrasound-based articulatory-to-acoustic mapping with a limited database. We use a multi-speaker pre-trained Tacotron2 TTS model and a pre-trained WaveGlow neural vocoder. The articulatory-to-acoustic conversion contains three steps: 1) from a sequence of ultrasound tongue image recordings, a 3D convolutional neural network predicts the inputs of the pre-trained Tacotron2 model, 2) the Tacotron2 model converts this intermediate representation to an 80-dimensional mel-spectrogram, and 3) the WaveGlow model is applied for final inference. This generated speech contains the timing of the original articulatory data from the ultrasound recording, but the F0 contour and the spectral information is predicted by the Tacotron2 model. The F0 values are independent of the original ultrasound images, but represent the target speaker, as they are inferred from the pretrained Tacotron2 model. In our experiments, we demonstrated that the synthesized speech quality is more natural with the proposed solutions than with our earlier model.

Session 3: Expressive synthesis

Improving Emotional TTS with an Emotion Intensity Input from Unsupervised Extraction
Bastian Schnell & Philip N. Garner

We aim to provide controls for emotion in synthetic speech. Many emotions are not displayed continuously in an otherwise emotional utterance; rather, the intensity varies with time. We show that an emotion recogniser is capable of producing a measure of emotion intensity via attention or saliency; this measure is appropriate to label utterances subsequently used to train a speech synthesiser. We evaluate novel and published means to do this showing that, whilst it is no longer state of the art for emotion recognition, attention is a good way to indicate emotion intensity for speech synthesis.

Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis
Slava Shechtman & Avrech Ben-David

Sequence-to-Sequence Text-to-Speech (S2S TTS) architectures that directly generate low level acoustic features from phonetic sequence are known to produce natural and expressive speech, when provided with moderate-to-large amounts of high quality training data. When exposed to a sequence of coarse speakeragnostic prosodic descriptors, such systems become prosodycontrollable and can learn and transfer desired prosodic patterns (e.g. word-emphasis or speaking style) from one seen speaker to another (in multi-speaker settings). But what if a high quality speech corpus for a desired speaking style is not available? In this work we explore the feasibility of teaching a neutral pre-trained prosody-controllable S2S TTS voice to speak with a conversational speaking style, as learnt from a low-quality multi-speaker spontaneous dialog corpus (originally intended for Automatic Speech Recognition). We have found that it is absolutely necessary to incorporate word semantics for that task. We fine-tune BERT network to predict the prosodic descriptors from the input text, based on that corpus, and apply them to the prosody-controllable S2S TTS at inference time. The subjective listening tests revealed that the learnt conversational style rated higher than baseline for 68

EmoCat: Language-agnostic Emotional Voice Conversion
Bastian Schnell, Goeric Huybrechts, Bartek Perz, Thomas Drugman & Jaime Lorenzo-Trueba

Emotional voice conversion models adapt the emotion in speech without changing the speaker identity or linguistic content. They are less data hungry than text-to-speech models and allow to generate large amounts of emotional data for downstream tasks. In this work we propose EmoCat, a language-agnostic emotional voice conversion model. It achieves high-quality emotion conversion in German with less than 45 minutes of German emotional recordings by exploiting large amounts of emotional data in US English. EmoCat is an encoder-decoder model based on CopyCat, a voice conversion system which transfers prosody. We use adversarial training to remove emotion leakage from the encoder to the decoder. The adversarial training is improved by a novel contribution to gradient reversal to truly reverse gradients. This allows to remove only the leaking information and to converge to better optima with higher conversion performance. Evaluations show that Emocat can convert to different emotions but misses on emotion intensity compared to the recordings, especially for very expressive emotions. EmoCat is able to achieve audio quality on par with the recordings for five out of six tested emotion intensities.

Enhancing audio quality for expressive Neural Text-to-Speech
Abdelhamid Ezzerg, Adam Gabrys, Bartosz Putrycz, Daniel Korzekwa, Daniel Saez-Trigueros, David McHardy, Kamil Pokora, Jakub Lachowicz, Jaime Lorenzo-Trueba & Viacheslav Klimkov

Artificial speech synthesis has made a great leap in terms of naturalness as recent Text-to-Speech (TTS) systems are capable of producing speech with similar quality to human recordings. However, not all speaking styles are easy to model: highly expressive voices are still challenging even to recent TTS architectures since there seems to be a trade-off between expressiveness in a generated audio and its signal quality. In this paper, we present a set of techniques that can be leveraged to enhance the signal quality of a highly-expressive voice without the use of additional data. The proposed techniques include: tuning the autoregressive loop's granularity during training; using Generative Adversarial Networks in acoustic modeling; and the use of Variational Auto-Encoders in both the acoustic model and the neural vocoder. We show that, when combined, these techniques greatly closed the gap in perceived naturalness between the baseline system and recordings by 39

Are we truly modeling expressiveness? A study on expressive TTS in Brazilian Portuguese for real-life application styles

Lucas H. Ueda, Paula D. P. Costa, Flavio O. Simoes & Mário U. Neto

This paper presents a study of expressive speech synthesis applied to real-life application styles in Brazilian Portuguese. We explore the use of data with different recording conditions in state-of-the-art architectures in expressive TTS. Our results suggest that the variability of recording conditions of the same style, combined with a guided training of the latent representation space of the Reference Encoder, assists in the modeling of non-archetypal expressivities. Additionally, we propose an alternative to evaluating the model's ability to generate expressive speech during preliminary results, based on a classifier using GeMAPS features.

Session 4: Articulation and Naturalness

Vocal tract area function extraction using ultrasound for articulatory speech synthesis
Debasish Ray Mohapatra, Pramit Saha, Yadong Liu, Bryan Gick & Sidney Fels

This paper studies the feasibility of an articulatory speech synthesizer by extracting the mid-sagittal tongue and palate contours using the ultrasound (US) imaging modality. The extracted contours are then used to compute the vocal tract crosssectional areas (i.e., area function) during phonation, which then drives an articulary speech synthesizer. Using this approach, we synthesized four phonetic vowel sounds (/a/, /i/, /e/ and /o/). The derived vocal tract (VT) transfer functions are shown to match over multiple utterances for a single vowel, thereby confirming reliable and accurate area function derivation using the US. The acoustic formants of simulated vowels using the proposed method show a modest deviation from the speaker's recorded speech signal since the current articulatory model does not include the mouth radiation mechanism. Furthermore, the higher formants' positions (F5-F8) are approximately equivalent to the high-quality standard MRI-based acoustic results and have an average error of 3.90

Non-Autoregressive TTS with Explicit Duration Modelling for Low-Resource Highly Expressive Speech
Raahil Shah, Kamil Pokora, Abdelhamid Ezzerg, Viacheslav Klimkov, Goeric Huybrechts, Bartosz Putrycz, Daniel Korzekwa & Thomas Merritt

Whilst recent neural text-to-speech (TTS) approaches produce high-quality speech, they typically require a large amount of recordings from the target speaker. In previous work [1], a 3- step method was proposed to generate high-quality TTS while greatly reducing the amount of data required for training. However, we have observed a ceiling effect in the level of naturalness achievable for highly expressive voices when using this approach. In this paper, we present a method for building highly expressive TTS voices with as little as 15 minutes of speech data from the target speaker. Compared to the current state-of-the-art approach, our proposed improvements close the gap to recordings by 23.3

Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies
Paul Konstantin Krug, Simon Stone & Peter Birkholz

In this work, the current state-of-the-art of articulatory speech synthesis (VOCALTRACTLAB) is compared to a wide range of different text-to-speech systems that once represented or still represent the continuously evolving state-of-the-art of speech synthesis technology. The comparison systems include neural and concatenative synthesis by Google and Microsoft, as well as Hidden Markov Model-based, unit-selection and diphone synthesis developed at universities (using MARYTTS, MBROLA and DRESS). A small corpus of 15 German sentences was synthesized using the text-to-speech (and, if available, re-synthesis) functionalities of each system. The intelligibility of the synthesized utterances was evaluated in an ASR experiment. The naturalness of the utterances was evaluated in a multi-stimulus Likert test by 50 German native speakers. As an additional reference, recordings of natural speech were used in the experiments. It was found that the articulatory synthesis can achieve a performance on par with the non-commercial synthesis systems in terms of intelligibility and naturalness, while being significantly outperformed by the commercial synthesis systems.

Perception of smiling voice in spontaneous speech synthesis
Ambika Kirkland, Marcin Włodarczak, Joakim Gustafson & Eva Szekely

Smiling during speech production has been shown to result in perceptible acoustic differences compared to non-smiling speech. However, there is a scarcity of research on the perception of "smiling voice" in synthesized spontaneous speech. In this study, we used a sequence-to-sequence neural text-tospeech system built on conversational data to produce utterances with the characteristics of spontaneous speech. Segments of speech following laughter, and the same utterances not preceded by laughter, were compared in a perceptual experiment after removing laughter and/or breaths from the beginning of the utterance to determine whether participants perceive the utterances preceded by laughter as sounding as if they were produced while smiling. The results showed that participants identified the post-laughter speech as smiling at a rate significantly greater than chance. Furthermore, the effect of content (positive/neutral/negative) was investigated. These results show that laughter, a spontaneous, non-elicited phenomenon in our model's training data, can be used to synthesize expressive speech with the perceptual characteristics of smiling.

Voicy: Zero-Shot Non-Parallel Voice Conversion in Noisy Reverberant Environments

Alejandro Mottini, Jaime Lorenzo-Trueba, Sri Vishnu Kumar Karlapati & Thomas Drugman

Voice Conversion (VC) is a technique that aims to transform the non-linguistic information of a source utterance to change the perceived identity of the speaker. While there is a rich literature on VC, most proposed methods are trained and evaluated on clean speech recordings. However, many acoustic environments are noisy and reverberant, severely restricting the applicability of popular VC methods to such scenarios. To address this limitation, we propose Voicy, a new VC framework particularly tailored for noisy speech. Our method, which is inspired by the de-noising auto-encoders framework, is comprised of four encoders (speaker, content, phonetic and acoustic-ASR) and one decoder. Importantly, Voicy is capable of performing non-parallel zero-shot VC, an important requirement for any VC system that needs to work on speakers not seen during training. We have validated our approach using a noisy reverberant version of the LibriSpeech dataset. Experimental results show that Voicy outperforms other tested VC techniques in terms of naturalness and target speaker similarity in noisy reverberant environments.

Keynote 2: Early Development of Infantile Communication by Sound

---

Early Development of Infantile Communication by Sound
István Winkler

István Winkler, PhD, DSc, electrical engineer, psychologist. He received his PhD in 1993 at the University of Helsinki, studying auditory sensory memory by electroencephalographic measures. He defended his Doctor of Science thesis in 2005 at the Hungarian Academy of Sciences on auditory deviance detection. His current fields of interest are predictive processing in the auditory deviance detection, auditory scene analysis, communication by sound, and the development of these functions in infancy. During his career, he has authored/coauthored over 250 publications, which received over 11000 references. Currently he is the director of the Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Budapest, Hungary and the head of the Sound and Speech Perception research group (http://www.ttk.hu/kpi/en/sound-and-speech-perception/).

Session 5: Emotion, singing and voice conversion

Rapping-Singing Voice Synthesis based on Phoneme-level Prosody Control
Konstantinos Markopoulos, Nikolaos Ellinas, Alexandra Vioni, Myrsini Christidou, Panos Kakoulidis, Georgios Vamvoukakis, June Sig Sung, Hyoungmin Park, Pirros Tsiakoulis, Aimilios Chalamandaris & Georgia Maniati

In this paper, a text-to-rapping/singing system is introduced, which can be adapted to any speaker's voice. It utilizes a Tacotron-based multi-speaker acoustic model trained on readonly speech data and which provides prosody control at the phoneme level. Dataset augmentation and additional prosody manipulation based on traditional DSP algorithms are also investigated. The neural TTS model is fine-tuned to an unseen speaker's limited recordings, allowing rapping/singing synthesis with the target's speaker voice. The detailed pipeline of the system is described, which includes the extraction of the target pitch and duration values from an a capella song and their conversion into target speaker's valid range of notes before synthesis. An additional stage of prosodic manipulation of the output via WSOLA is also investigated for better matching the target duration values. The synthesized utterances can be mixed with an instrumental accompaniment track to produce a complete song. The proposed system is evaluated via subjective listening tests as well as in comparison to an available alternate system which also aims to produce synthetic singing voice from read-only training data. Results show that the proposed approach can produce high quality rapping/singing voice with increased naturalness.

Exploring Disentanglement with Multilingual and Monolingual VQ-VAE
Jennifer Williams, Jason Fong, Erica Cooper & Junichi Yamagishi

This work examines the content and usefulness of disentangled phone and speaker representations from two separately trained VQ-VAE systems: one trained on multilingual data and another trained on monolingual data. We explore the multi- and monolingual models using four small proof-of-concept tasks: copysynthesis, voice transformation, linguistic code-switching, and content-based privacy masking. From these tasks, we reflect on how disentangled phone and speaker representations can be used to manipulate speech in a meaningful way. Our experiments demonstrate that the VQ representations are suitable for these tasks, including creating new voices by mixing speaker representations together. We also present our novel technique to conceal the content of targeted words within an utterance by manipulating phone VQ codes, while retaining speaker identity and intelligibility of surrounding words. Finally, we discuss recommendations for further increasing the viability of disentangled representations.

Text-to-Speech Synthesis Techniques for MIDI-to-Audio Synthesis
Erica Cooper, Xin Wang & Junichi Yamagishi

Speech synthesis and music audio generation from symbolic input differ in many aspects but share some similarities. In this study, we investigate how text-to-speech synthesis techniques can be used for piano MIDI-to-audio synthesis tasks. Our investigation includes Tacotron and neural sourcefilter waveform models as the basic components, with which we build MIDI-to-audio synthesis systems in similar ways to TTS frameworks. We also include reference systems using conventional sound modeling techniques such as samplebased and physical-modeling-based methods. The subjective experimental results demonstrate that the investigated TTS components can be applied to piano MIDI-to-audio synthesis with minor modifications. The results also reveal the performance bottleneck – while the waveform model can synthesize high quality piano sound given natural acoustic features, the conversion from MIDI to acoustic features is challenging. The full MIDI-to-audio synthesis system is still inferior to the sample-based or physical-modeling-based approaches, but we encourage TTS researchers to test their TTS models for this new task and improve the performance.

Preliminary study on using vector quantization latent spaces for TTS/VC systems with consistent performance
Hieu-Thi Luong & Junichi Yamagishi

Generally speaking, the main objective when training a neural speech synthesis system is to synthesize natural and expressive speech from the output layer of the neural network without much attention given to the hidden layers. However, by learning useful latent representation, the system can be used for many more practical scenarios. In this paper, we investigate the use of quantized vectors to model the latent linguistic embedding and compare it with the continuous counterpart. By enforcing different policies over the latent spaces in the training, we are able to obtain a latent linguistic embedding that takes on different properties while having a similar performance in terms of quality and speaker similarity. Our experiments show that the voice cloning system built with vector quantization has only a small degradation in terms of perceptive evaluations, but has a discrete latent space that is useful for reducing the representation bit-rate, which is desirable for data transferring, or limiting the information leaking, which is important for speaker anonymization and other tasks of that nature.

Low-latency real-time non-parallel voice conversion based on cyclic variational autoencoder and multiband WaveRNN with data-driven linear prediction
Patrick Lumban Tobing & Tomoki Toda

This paper presents a low-latency real-time (LLRT) non-parallel voice conversion (VC) framework based on cyclic variational autoencoder (CycleVAE) and multiband WaveRNN with datadriven linear prediction (MWDLP). CycleVAE is a robust nonparallel multispeaker spectral model, which utilizes a speakerindependent latent space and a speaker-dependent code to generate reconstructed/converted spectral features given the spectral features of an input speaker. On the other hand, MWDLP is an efficient and a high-quality neural vocoder that can handle multispeaker data and generate speech waveform for LLRT applications with CPU. To accommodate LLRT constraint with CPU, we propose a novel CycleVAE framework that utilizes mel-spectrogram as spectral features and is built with a sparse network architecture. Further, to improve the modeling performance, we also propose a novel fine-tuning procedure that refines the frame-rate CycleVAE network by utilizing the waveform loss from the MWDLP network. The experimental results demonstrate that the proposed framework achieves highperformance VC, while allowing for LLRT usage with a singlecore of 2.1–2.7 GHz CPU on a real-time factor of 0.87–0.95, including input/output, feature extraction, on a frame shift of 10 ms, a window length of 27.5 ms, and 2 lookup frames.

Session 6: Multilingual and evaluation

Factors Affecting the Evaluation of Synthetic Speech in Context
Johannah O'Mahony, Pilar Oplustil-Gallegos, Catherine Lai & Simon King

Realizing text-to-speech (TTS) system of dialects is useful for personalizing TTS systems. However, TTS for many dialects of pitch accent languages is not realized because of lowresourced problem. Among many dialects of pitch accent languages, this paper focuses on Osaka dialect of Japanese, one of the most challenging pitch accent languages. For Japanese TTS system, accent labels are known to be necessary as input to synthesize natural speech. In rich-resourced dialect, humanresourced approaches and dictionary-based approaches are often used to annotate accent labels for training and inference, but such approaches are unfeasible and time-consuming for lowresourced dialects. In this paper, we propose accent extraction model that utilizes vector quantized variational autoencoder (VQ-VAE) to prepare accent information from speech, and accent prediction models that utilize decision tree and deep learning techniques to predict accent information from the input text. The models were examined with corpus of Osaka dialect, whose accent labels do not exist. The results showed that accent extraction model succeeded in extracting accent information of Osaka dialect from speech utterances as latent variable. It also showed that the accent of synthesized speech by accent prediction models were not better than baseline, but it had advantages such as interpretability.

Non-native English lexicon creation for bilingual speech synthesis
Arun Baby, Pranav Jawale, Saranya Vinnaitherthan, Sumukh Badam, Nagaraj Adiga & Sharath Adavane

Bilingual English speakers speak English as one of their languages. Their English is of a non-native kind, and their conversations are of a code-mixed fashion. The intelligibility of a bilingual text-to-speech (TTS) system for such non-native English speakers depends on a lexicon that captures the phoneme sequence used by non-native speakers. However, due to the lack of non-native English lexicon, existing bilingual TTS systems employ native English lexicons that are widely available, in addition to their native language lexicon. Due to the inconsistency between the non-native English pronunciation in the audio and native English lexicon in the text, the intelligibility of synthesized speech in such TTS systems is significantly reduced. This paper is motivated by the knowledge that the native language of the speaker highly influences non-native English pronunciation. We propose a generic approach to obtain rules based on letter to phoneme alignment to map native English lexicon to their non-native version. The effectiveness of such mapping is studied by comparing bilingual (Indian English and Hindi) TTS systems trained with and without the proposed rules. The subjective evaluation shows that the bilingual TTS system trained with the proposed non-native English lexicon rules obtains a 6

Cross-lingual Transfer of Phonological Features for Low-resource Speech Synthesis
Dan Wells & Korin Richmond

Previous work on cross-lingual transfer learning in text-tospeech has shown the effectiveness of fine-tuning phonemic representations on small amounts of target language data. In other contexts, phonological features (PFs) have been suggested as a more suitable input representation than phonemes for sharing acoustic information between languages, for example in multilingual model training or for code-switching synthesis where an utterance may contain words from multiple languages. Starting from a model trained on 14 hours of English, we find that cross-lingual fine-tuning with 15 minutes of German data can produce speech with subjective naturalness ratings comparable to a model trained from scratch on 4 hours of German, using either phonemes or PFs. We also find a modest but statistically significant improvement in naturalness ratings using PFs over phonemes when training from scratch on 4 hours of German.

Mind your p's and k's – Comparing obstruents across TTS voices of the Blizzard Challenge 2013

Ayushi Pandey, Sébastien Le Maguer, Julie Berndsen & Naomi Harte

Obstruent consonants have been investigated in speech quality assessment studies of natural speech, where enhancing their perception has improved overall speech quality. This paper presents a comparative analysis of acoustic-phonetic features of obstruent consonants in synthetic speech. Features for obstruent consonants are identified where TTS systems differ significantly from a natural human voice, as a function of quality. The synthetic speech voices from the Blizzard Challenge of 2013 are used for this investigation. TTS systems were first assigned groups based on their MOS rating (quality) and shared TTS technique (family). Then, acoustic-phonetic features characteristic of contrastive properties in obstruents, were extracted from all systems. While quality differences between low-rated systems and high-rated systems were observed in a large number of features, we report those where statistically significant differences (p-val < 0.001) were observed between the systems. Where quality effects were not found, we investigated whether systems of the same family exhibit similar behaviour. Finally, individual systems within a group were examined for their differing influence on the acoustic-phonetic feature set of obstruents. Here, we found that HMM systems with similar MOS ratings do not differ in their acoustic realization of obstruents, while Unit Selection systems showed stronger individual system variability. A comparative analysis of obstruent consonants across TTS systems applies techniques from the domain of corpusphonetics to the task of speech synthesis evaluation. Identifying phonologically relevant acoustic features, may indicate the underlying articulatory process compromised in those systems, that correlates with the distorted acoustics.

Improving Polyglot Speech Synthesis through Multi-task and Adversarial Learning

Jason Fong, Jilong Wu, Prabhav Agrawal, Andrew Gibiansky, Thilo Koehler & Qing He

It is still quite challenging for polyglot speech synthesis systems to synthesise speech with the same pronunciations and accent as a native speaker, especially when there are fewer speakers per language. In this work, we target an extreme version of the polyglot synthesis problem, where we have only one speaker per language, and the system has to learn to disentangle speaker from language features from just one speakerlanguage pair. To tackle this problem, we propose a novel approach based on a combination of multi-task learning and adversarial learning to help the model produce more realistic acoustic features for speaker-language combinations for which we have no data. Our proposed system improves the overall naturalness of synthesised speech achieving upto 4.2

Session 7: Modeling and evaluation

Multi-Scale Spectrogram Modelling for Neural Text-to-Speech
Ammar Abbas, Bajibabu Bollepalli, Alexis Moinet, Arnaud Joly, Penny Karanasou, Peter Makarov, Simon Slangens, Sri Karlapati & Thomas Drugman

We propose a novel Multi-Scale Spectrogram (MSS) modelling approach to synthesise speech with an improved coarse and fine-grained prosody. We present a generic multi-scale spectrogram prediction mechanism where the system first predicts coarser scale mel-spectrograms that capture the suprasegmental information in speech, and later uses these coarser scale melspectrograms to predict finer scale mel-spectrograms capturing fine-grained prosody. We present details for two specific versions of MSS called Word-level MSS and Sentence-level MSS where the scales in our system are motivated by the linguistic units. TheWord-level MSS models word, phoneme, and framelevel spectrograms while Sentence-level MSS models sentencelevel spectrogram in addition. Subjective evaluations show that Word-level MSS performs statistically significantly better compared to the baseline on two voices.

How do Voices from Past Speech Synthesis Challenges Compare Today?
Erica Cooper & Junichi Yamagishi

Shared challenges provide a venue for comparing systems trained on common data using a standardized evaluation, and they also provide an invaluable resource for researchers when the data and evaluation results are publicly released. The Blizzard Challenge and Voice Conversion Challenge are two such challenges for text-to-speech synthesis and for speaker conversion, respectively, and their publicly-available system samples and listening test results comprise a historical record of stateof- the-art synthesis methods over the years. In this paper, we revisit these past challenges and conduct a large-scale listening test with samples from many challenges combined. Our aims are to analyze and compare opinions of a large number of systems together, to determine whether and how opinions change over time, and to collect a large-scale dataset of a diverse variety of synthetic samples and their ratings for further research. We found strong correlations challenge by challenge at the system level between the original results and our new listening test. We also observed the importance of the choice of speaker on synthesis quality.

Accent Modeling of Low-Resourced Dialect in Pitch Accent Language Using Variational Autoencoder
Kazuya Yufune, Tomoki Koriyama, Shinnosuke Takamichi & Hiroshi Saruwatari

In this paper, we propose an emotion-controllable text-tospeech (TTS) model that allows both emotional-level (i.e., coarse-grained) control and prosodic-feature-level (i.e., finegrained) control of speech using both emotional soft-labels and prosodic features. Conventional methods control speech only by using emotional labels or prosodic features (e.g., mean and standard deviation of pitch), which cannot express diverse emotions. Our model is based on a prosodic feature generator that decodes emotion soft-labels into prosodic features. It allows controlling the emotion of synthetic speech by both emotion labels and prosodic features. The experiment results show 1) the emotion-perceptual accuracy of synthetic speech reaches 66

Liaison and Pronunciation Learning in End-to-End Text-to-Speech in French
Jason Taylor, Sébastien Le Maguer & Korin Richmond

Sequence-to-sequence (S2S) TTS models like Tacotron have grapheme-only inputs when trained fully end-to-end. Grapheme inputs map to phone sounds depending on context, which traditionally is handled by extensive preprocessing in the TTS front-end. However, French orthography does not provide a clear one-to-one mapping between graphemes and sounds, and in English, which similarly has rather non-phonetic orthography, pronunciations are a significant cause of error in S2STTS with grapheme-inputs. In this paper, we test implicit pronunciation knowledge where graphemes do not map directly to phones. Implicit pronunciation knowledge learnt in S2S-TTS is similar to a standalone grapheme-to-phoneme (G2P) model, which makes explicit phone predictions at the sequential level. We find grapheme-input S2S-TTS makes implicit pronunciation errors similar to explicit G2P models - notably for foreign names. In a traditional front-end pipeline, there are also postlexical rules which modify G2P output at the sequential level. In French, post-lexical rules require a deep knowledge of linguistic structure in a process called Liaison. Without explicit rules, we find S2S-TTS with grapheme-inputs over-inserts Liaison sounds, leading to a significant preference for a phonebased equivalent. By testing with linguistically-motivated stimuli, we observe differences that would otherwise go undetected.

FeatherTTS: Robust and Efficient attention based Neural TTS

Qiao Tian, Chao Liu, Zewang Zhang, Heng Lu, Linghui Chen, Bin Wei, Pujiang He & Shan Liu

Attention based neural TTS is elegant speech synthesis pipeline and has shown a powerful ability to generate natural speech. However, it is still not robust enough to meet the stability requirements for industrial products. Besides, it suffers from slow inference speed owning to the autoregressive generation process. In this work, we propose FeatherTTS, a robust and efficient attention-based neural TTS system. Firstly, we propose a novel Gaussian attention which utilizes interpretability of Gaussian attention and the strict monotonic property in TTS. By this method, we replace the commonly used stop token prediction architecture with attentive stop prediction. Secondly, we apply block sparsity on the autoregressive decoder to speed up speech synthesis. The experimental results show that our proposed FeatherTTS not only nearly eliminates the problem of word skipping, repeating in particularly hard texts and keep the naturalness of generated speech, but also speeds up acoustic feature generation by 3.5 times over Tacotron. Overall, the proposed FeatherTTS can be 35x faster than real-time on a single CPU.

Keynote 3: Deep Audio Conversion Technologies and Their Applications in Speech, Singing, and Music

---

Expressive Neural TTS
Thomas Drugman

Lior Wolf is a research scientist at Facebook AI Research and a full professor in the School of Computer Science at Tel-Aviv University, Israel. He conducted postdoctoral research at prof. Poggio's lab at the Massachusetts Institute of Technology and received his PhD degree from the Hebrew University, under the supervision of Prof. Shashua. He is an ERC grantee and has won the ICCV 2001 and ICCV 2019 honorable mention, and the best paper awards at ECCV 2000 and ICANN 2016. His research focuses on computer vision, audio synthesis, and deep learning.

Session 8: Synthesis and Context

Comparing acoustic and textual representations of previous linguistic context for improving Text-to-Speech
Pilar Oplustil-Gallegos, Johannah O'Mahony & Simon King

Text alone does not contain sufficient information to predict the spoken form. Using additional information, such as the linguistic context, should improve Text-to-Speech naturalness in general, and prosody in particular. Most recent research on using context is limited to using textual features of adjacent utterances, extracted with large pre-trained language models such as BERT. In this paper, we compare multiple representations of linguistic context by conditioning a Text-to-Speech model on features of the preceding utterance. We experiment with three design choices: (1) acoustic vs. textual representations; (2) features extracted with large pre-trained models vs. features learnt jointly during training; and (3) representing context at the utterance level vs. word level. Our results show that appropriate representations of either text or acoustic context alone yield significantly better naturalness than a baseline that does not use context. Combining an utterance-level acoustic representation with a word-level textual representation gave the best results overall.

Audiobook Speech Synthesis Conditioned by Cross-Sentence Context-Aware Word Embeddings
Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, Naoko Tanji, Yusuke Ijima, Ryo Masumura & Hiroshi Saruwatari

This paper proposes an audiobook speech synthesis method that considers a wider range of contexts than a sentence level. The style of the audiobook speech depends not only on the current sentence to be synthesized but also on its neighboring sentences. Therefore, unlike conventional text-to-speech synthesis for isolated sentences, it is necessary to consider the context of the neighboring sentences. Our method utilizes cross-sentence context-aware word embedding, which is obtained by inputting the neighboring and current sentences into BERT. The speech synthesis model, Tacotron2, is conditioned by this word embedding in addition to the current sentence. Experimental results show that taking neighboring sentences into account significantly improves synthetic speech quality.

Lipsyncing efforts for transcreating lecture videos in Indian languages
Mano Ranjith Kumar M, Jom Kuriakose, Karthik Pandia D S & Hema A Murthy

This paper proposes a novel lip-syncing module for the transcreation of lecture videos from English to Indian languages. The audio from the lecture is transcribed using automatic speech recognition. The text is translated and manually curated before and after translation to avoid mistakes. The curated text is synthesized using the Indian language end-to-endbased text-to-speech synthesis systems. The synthesized audio and video are out-of-sync. This paper attempts to automate this process of producing video lectures lip-synced into Indian languages using different techniques. Lip-syncing an educational video with the Indian language audio is challenging owing to (a) the duration of Indian language audio being considerably longer or shorter than that of the original audio, (b) the extempore speech causes the audio in the source videos to have long silences. Any modification to the speed of audio can be unpleasant to listeners. The proposed system non-uniformly re-samples the video to ensure better lip-syncing. The novelty of this paper is in the use of HMMGMM alignments in tandem with syllable segmentation using group delay, as visemes are closer to syllables. The proposed lip-syncing techniques are evaluated using subjective evaluation methods. Results indicate that accurate alignment at the syllable level is crucial for lip-syncing.

Homograph disambiguation with contextual word embeddings for TTS systems
Marco Nicolis & Viacheslav Klimkov

We describe a heterophone homograph (simply 'homograph' henceforth) disambiguation system based on per-case classifiers, trained on a small amount of labelled data. These classifiers use contextual word embeddings as input features and achieve state-of-the-art accuracy of 0.991 on the English homographs on a publicly available dataset, without any additional rule system being necessary. We show that as little as 100 sentences are sufficient to train a light-weight dedicated classifier, provided the dataset is sufficiently balanced, i.e. all versions of the homograph are adequately represented. We further add data in cases where the original dataset is deeply unbalanced (i.e. one homograph version overwhelmingly represented). This is effectively a special case of active learning, by which we select additional cases of the under-represented homograph versions and show an 11

Analysing Temporal Sensitivity of VQ-VAE Sub-Phone Codebooks
Jason Fong, Jennifer Williams & Simon King

In this work we present an analysis of temporal sensitivity of VQ-VAE sub-phone token sequences. Previous work has demonstrated that VQ-VAE systems learn a type of sub-phone representation. However, a detailed examination of the representations themselves is currently lacking. We address this gap by exploring linguistic unit reorganisation. Our experiments show that sub-phone codebook sequences are temporally correlated enough to identify VQ codes that correspond to distinct linguistic units. We found that it is possible to extract VQ codes and re-arrange these linguistic units in a meaningful way (i.e. changing the word-order of a sentence). This work puts us one step closer to understanding how to modify pronunciations at a fine granularity, such as below the phone-level unit.

INDEX