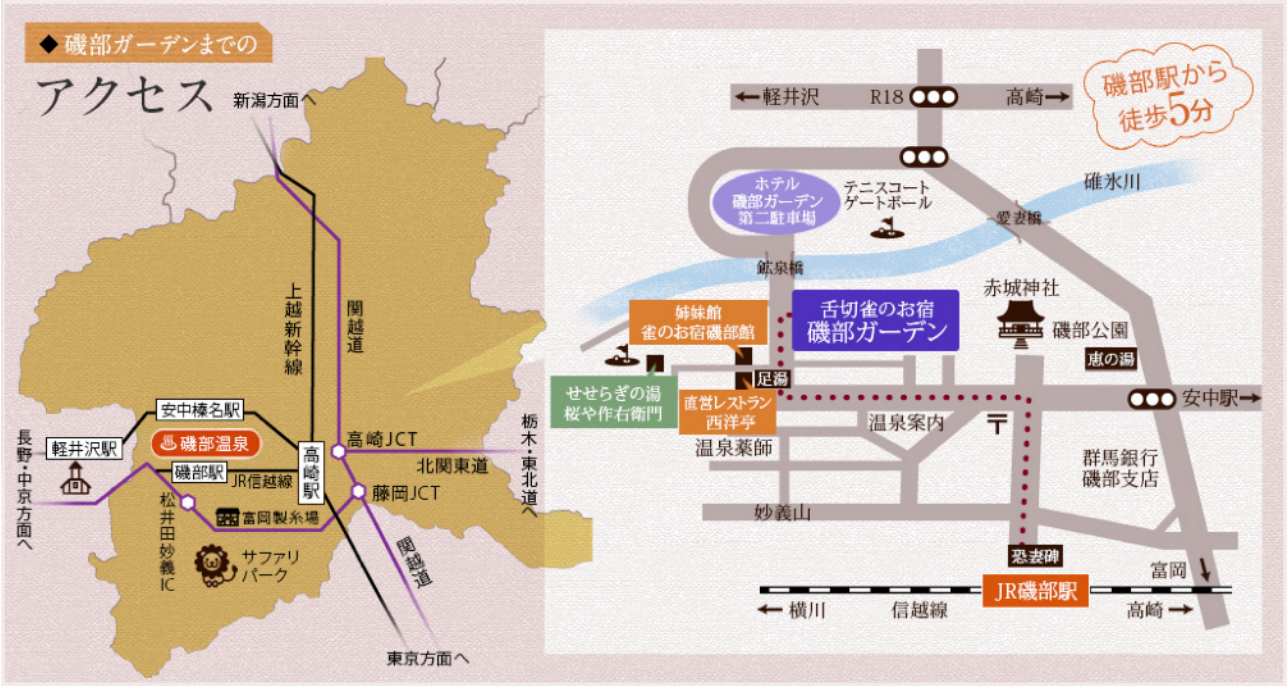
第126回 音声言語情報処理研究会 (SIG-SLP) http://sig-slp.jp/

日程：2019/02/27(水)~28(木)

会場：舌切雀のお宿　ホテル磯部ガーデン（磯部温泉）
4F　桜の間 (411)

〒379-0127群馬県安中市磯部1-12-5
【TEL】027-385-0085
【FAX】027-385-0055
URL http://www.isobesuzume.co.jp

会場への交通アクセス：



東京→上越新幹線（50分）→ 高崎→ 信越本線（18分）→磯部
→徒歩（5分）→磯部ガーデン

Technical Program

02/27	1日目	02/28	2日目
12:00~	受付開始 (4F　桜の間 411)	7:00〜 9:00	朝食 (2F　鳳凰の間)
13:00~ 14:15	一般講演１（英語） (座長：山岸順一)	10:00- 10:50	一般講演２ (日本語) (座長：篠崎隆宏)
14:30- 15:30	**招待講演１** **Prof. Zhenhua Ling** (座長：Xin Wang)	11:00~ 12:00	**【企画】** **Interspeech2018** **報告会**
15:40- 16:40	**招待講演２** **Dr. Wei Ping** (座長：Xin Wang)	12:00- 13:00	昼食 （4F　桜の間 412）
16:50- 17:50	**招待講演３** **Dr. Kong Aik Lee** (座長：Xin Wang)	13:00- 14:00	**招待講演４** **大木 哲史 博士** (座長：塩田さやか)
18:30- 20:00	夕食 （会場：3F　竹の間）	14:00- 14:30	SIG-SLP企業賞発表と紹介 Fairly Devices賞 Yahoo! JAPAN賞 (座長：西村 雅史)
21:00- 22:00	**【企画】** **国際学会既発表セッション** **(座長：高木信二)**	15:00	富岡製糸場見学 （希望者のみ、私費）

都合により、アナウンスなくスケジュールが変更になることがあります。

招待講演1

Prof. Zhenhua Ling

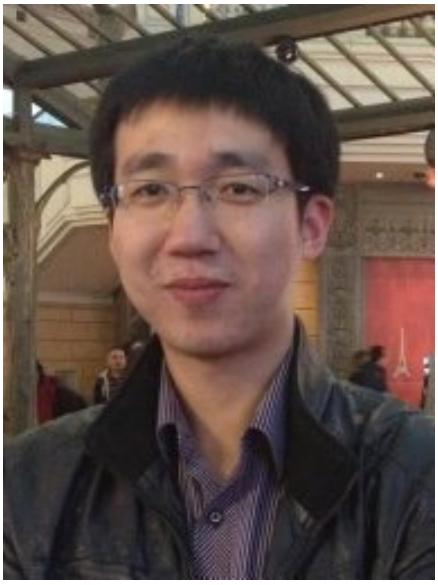
University of Science and Technology of China, China

Deep Learning-Based Voice Conversion

Abstract:
I will introduce our recent work on applying deep learning techniques to voice conversion in this talk. Several methods have been proposed to improve different components in the pipeline of a statistical parametric voice conversion system, including deep neural networks with layer-wise generative training for acoustic modeling, deep autoencoders with binary distributed hidden units for feature representation, and WaveNet vocoder with limited training data for waveform reconstruction. Then, I will introduce our system designed for Voice Conversion Challenge 2018, which achieved the best performance under both parallel and non-parallel conditions in this evaluation. After this, I will present our recent progress on sequence-to-sequence acoustic modeling for voice conversion, which converts the acoustic features and durations of source utterances simultaneously using a unified acoustic model. Finally, some discussions on the future development of voice conversion techniques will be given.

Bio:
Zhenhua Ling received the B.E. degree in electronic information engineering, the M.S. and Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008, respectively. From October 2007 to March 2008, he was a Marie Curie Fellow with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K. From July 2008 to February 2011, he was a joint Postdoctoral Researcher with the University of Science and Technology of China, Hefei, China, and iFLYTEK Co., Ltd., China. He is currently an Associate Professor with the University of Science and Technology of China. He also worked at the University of Washington, Seattle, WA, USA, as a Visiting Scholar from August 2012 to August 2013. His research interests include speech processing, speech synthesis, voice conversion, and natural language processing. He was the recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2010. He is now an Associate Editor for the IEEE/ACM transactions on audio, speech, and language processing.

招待講演2



Dr. Wei Ping

Baidu Research, USA

End-to-End Neural TTS and Parallel Wave Generation

Abstract:

There are two phenomenal trends in speech synthesis research: 1) directly generating waveform samples through state-of-the-art generative models, and 2) building end-to-end TTS systems without too much expert knowledge. In this talk, I will first present our recent work on parallel wave generation, which removes the autoregressive inference bottleneck in WaveNet. I will also compare various non-autoregressive generative models for waveform synthesis. In addition, “end-to-end"speech synthesis actually refers to the text-to-spectrogram models with a separate vocoder in previous studies. I will introduce the first text-to-wave neural architecture for TTS, which is fully convolutional and enables truly end-to-end training from scratch.

Bio:

Wei Ping obtained his Ph.D. in Computer Science from University of California, Irvine (UCI) in 2016. Before that, he received his B.E. and M.E. degrees from Harbin Institute of Technology and Tsinghua University in 2008 and 2011, respectively. He is currently a Senior Research Scientist at Baidu Research (USA), leading their team on speech synthesis. His research area is machine learning and speech synthesis, with interests spread over deep learning, generative models, graphical models and variational inference.

He has published a series of advanced deep learning papers on text-to-speech synthesis, including Deep Voice 2 (NIPS’17), Deep Voice 3 (ICLR’18), Neural Voice Cloning (NIPS’18) and ClariNet (Submitted to ICLR 2019), which are well-known among speech synthesis researchers.

招待講演３

Dr. Kong Aik Lee

NEC Corp, Central Research Laboratories, Japan

Speaker Embedding and Recent Advances in Speaker Recognition Evaluation

Abstract:

Automatic speaker recognition is the task of identifying or verifying an individual's identity from their voice samples using machine learning algorithms, without any human intervention. Voice is a combination of physical and behavioral biometrics characteristics. The physical features of an individual's voice are based on the shape and size of the vocal tract, mouth, nasal cavities, and lips that involve in producing speech sound. The behavioral aspects, which include the use of a particular accent, intonation style, pronunciation pattern, choice of vocabulary and so on, are associated more with the words or lexical content of the spoken utterances. Speaker recognition has seen significant advancements over the past few decades, giving rise to the successful introduction of commercial products. With the advent of Big Data and the resurrection of data-hungry modeling techniques such as artificial neural networks, more recently the research focus has shifted from a more controlled scenario towards larger and more realistic speaker in the wild scenarios. The latest cycle of NIST evaluations (SRE'18), which in addition to traditional conversational telephony speech (CTS) involves voice over IP (VOIP) data as well as audio extracted from online videos. This talk aim to present recent technical advances in speaker recognition and NIST SRE'18 from NEC perspective.

Bio:

Kong Aik Lee is a Senior Principal Researcher, Biometrics Research Laboratories, NEC Corp., Japan. He received his B.E. (First Class Honours) in Electrical Engineering from University Technology Malaysia (UTM), Malaysia in 1999 and Ph.D. from Nanyang Technological University (NTU), Singapore in 2005. From 2006 to 2018, He was a research scientist at the Human Language Technology Department, Institute for Infocomm Research (I2R), A*STAR, Singapore. His current research interests include speaker recognition and characterization, multilingual recognition and identification, speech analysis and processing, machine learning and digital signal processing. He also serves as an Editorial Board Member for Elsevier Computer Speech and Language, and Associate Editor for IEEE/ACM Transactions on Audio, Speech and Language Processing. He is a senior member of IEEE.

招待講演4

大木 哲史 博士

静岡大

生体認証に対するなりすまし攻撃とその対策

概要:

機械学習アルゴリズムに関する研究の発展は目覚ましいが，近年ではその機能性のみならず，アルゴリズムのセキュリティへの配慮の重要性に対する認識が高まっている．本講演では，機械学習アルゴリズムに対する攻撃手法を紹介するとともに，話者認識におけるMAP適応を例にあげて，その理論的脆弱性の解析を元に，登録された全ての話者に対して，漸近的に100%の確率でなりすますことができるウルフ攻撃の実現可能性について紹介する．

　また，このような攻撃を含む未知のなりすまし攻撃を検知するための取り組みについて紹介する．

経歴

2010年早稲田大学博士（工学）。早稲田大学理工学研究所嘱託研究員、同研究所招聘研究員、同研究所次席研究員（研究員助教）を経て、2017年4月より静岡大学情報学部情報科学科/総合科学技術研究科　情報学専攻　講師。生体認証、ネットワークセキュリティ、機械の人間との間に生じる関係とそのセキュリティ問題に関する研究に従事。

企画１：国際学会既発表セッション(ポスター発表)

Yusuke Yasuda, Xin Wang, Shinji Takaki, Junichi Yamagishi (NII),
“Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language”
Abstract: End-to-end speech synthesis is a promising approach that directly converts raw text to speech. Japanese could be one of the most difficult languages for which to achieve end-to-end speech synthesis in two reasons. One reason is its character diversity, and the other is its pitch accents. Therefore, state-of-the-art systems are still based on a traditional pipeline framework that requires a separate text analyzer and duration model. As a first step towards end-to-end speech synthesis, this research focus on pitch accent in Japanese language. We propose a new architecture that extends Tacotron with self-attention to capture long-term dependencies related to pitch accents. A large-scale listening test show the proposed system outperforms baseline Tacotron. In addition, we investigated the impacts of the presence of accentual-type labels, the use of force or predicted alignments, and acoustic features used as local condition parameters of the Wavenet vocoder. Our results reveal that although the proposed systems still do not match the quality of a top-line pipeline system for Japanese, we show important stepping stones towards end-to-end Japanese speech synthesis. **(Accepted for ICASSP 2019)**

Yi Zhao, Shinji Takaki, Hieu-Thi Luong, Junichi Yamagishi (NII), Daisuke Saito, Nobuaki Minematsu (University of Tokyo)
“Wasserstein GAN and Waveform Loss-based Multi-speaker Acoustic Model Training for Text-to-Speech System”
Abstract: Recent neural networks directly learned from speech waveform samples such as WaveNet and sampleRNN have achieved very high-quality synthetic speech in terms of both naturalness and speaker similarity even in multi-speaker text-to-speech synthesis systems. Such the neural networks are used as a replacement of vocoder and hence they are often called neural vocoder. The neural vocoder uses acoustic features as local condition parameters, which need to be accurately predicted by another acoustic model. However it is not fully investigated how we should train the acoustic model that predicts the local condition parameters and the final quality of synthetic speech are significantly affected by the performance of the acoustic model. The significant degradation happens especially when predicted acoustic features has mismatched characteristics compared to natural ones. In order to reduce the mismatched characteristics between natural and generated acoustic features, this paper proposes frameworks which incorporates either conditional generative adversarial networks (GANs) or its variant called Wasserstein GAN with gradient penalty (WGAN-GP) into multi-speaker speech synthesis that uses the Wavenet vocoder. Furthermore, this paper extends the GAN frameworks and uses the discretized mixture logistic loss of a well-trained WaveNet as well as mean squared error and adversarial losses as parts of objective functions. Experimental results show that acoustic models trained using the WGAN-GP framework using back propagated DML loss can achieve highest subjective evaluation scores in terms of both quality and speaker similarity. **(Published at IEEE Access)**

Tifani Warnita, Nakamasa Inoue, Koichi Shinoda (Tokyo Institute of Technology),
“Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data”
Abstract: Early prediction of Alzheimer's disease has the major importance for taking an appropriate and quick treatment in order to prevent the disease become worse. In this paper, we propose our linguistic-independent approach for detecting dementia by utilizing only speech data whereas most of the previous study using linguistic information in their approach. We extract the paralinguistic feature set for each utterance of the patient then do the classification using Gated Convolutional Neural Network (GCNN) with a majority voting mechanism for the final verdict. We evaluated our method using Pitt Corpus and yield the accuracy of 73.6%, which is better than the conventional sequential minimal optimization (SMO) by 7.6 points. This GCNN can be trained with

a relatively small amount of data and can capture the temporal information in audio paralinguistic features. Furthermore, since we do not use any linguistic features, our approach has the advantage of being more easily applicable in various languages. **(Published at Interspeech 2018)**

Fuming Fang, Xin Wang, Junichi Yamagishi , Isao Echizen (NII),
"Audiovisual speaker conversion: jointly and simultaneously transforming facial expression and acoustic characteristics"
Abstract: We present an audiovisual speaker conversion method, which jointly and simultaneously transforms both facial expressions and voice of a source speaker into those of a target speaker. Since facial and acoustic features are highly correlated together by the proposed method, it would allow them to compensate for each other and thus the converted target speaker appears and sounds natural. We used a neural network to convert facial and acoustic features and then used a WaveNet and an image-reconstruction network to generate waveform and RGB image from both the converted features. Experimental results showed that the proposed method achieved better naturalness and speaker similarity compared with one that separately transformed facial and acoustic features. **(Accepted for ICASSP 2019)**

Yilong Peng, Hayato Shibata, Takahiro Shinozaki (Tokyo Institute of Technology)
"Reward Only Training of Encoder-Decoder Digit Recognition Systems Based on Policy Gradient Methods"
Abstract: Recently, zero resource speech recognition has gotten more popular not only in realistic engineering but also in scientific purposes. However, existing unsupervised learning methods that use speech input only are unable to associate speech to its corresponding text. In this paper, we propose an approach that assumes a scalar reward is given for each decoded result, use it to train the system in reinforcement learning. Focusing on encoder-decoder based speech recognition neural network system, we find the difficulty is to obtain a convergence without the help of supervised learning. Towards this problem, we explore on several neural network architectures, optimization methods and reward definitions, seeking a suitable configuration for policy gradient reinforcement learning. We performed experiments on connected digit utterances from the TIDIGITS corpus and reduce the digit error rate to 13.6% on our best performed digit recognition system, reveal the appropriate condition for unsupervised reinforcement learning and shows it is largely different from supervised training. **(Published at APSIPA 2018)**

Tomohiro Tanaka and Takahiro Shinozaki (Tokyo Institute of Technology)
"End-to-End Training of Keyword Detection Neural Networks Using F-Measure Objective and 2D-RNN"
Abstract: Acoustic embedding based keyword detection neural network models have been showing excellent performance. In this work, we propose two end-to-end continuous keyword detectors that work without assuming the segment boundaries of keywords in the input sequence. One is an extension of the conventional long short term memory (LSTM) based embedding approach removing the segmentation assumption, and the other is an extension of continuous dynamic programming (DP) matching to an end-to-end neural network by using a two-dimensional recurrent neural network (2D-RNN). For the training, we propose and investigate a soft decision version of the F-measure as the objective function in addition to the cross-entropy measure to use a consistent evaluation measure in the training and the evaluation. Experiments using the WSJ corpus show the 2D-RNN based continuous DP matching has much higher performance than the embedding based detector and posteriogram feature based conventional continuous DP matching. **(Published at APSIPA 2018)**

Bairong Zhuang, Wenbo Wang and Takahiro Shinozaki (Tokyo Institute of Technology)
"Investigation of Attention-Based Multimodal Fusion and Maximum Mutual Information Objective for DSTC7 Track3"
Abstract: In this paper, we show our investigation on the Audio Visual Scene-aware dialog (AVSD) task which is proposed in DSTC7. We investigate the effectiveness of different modality fusion methods as well as different input modalities. We also employ the Maximum Mutual Information(MMI) objective as the objective for the AVSD system. Our experiments shows the

system that uses MMI as the objective obtains 6.6% relative improvement over the baseline system on BLEU. **(Published at DSTC7 workshop: Dialog System Technology Challenges)**

Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, and Tomoki Toda (Nagoya University)
"Collapsed speech segment detection and suppression for WaveNet vocoder"
Abstract: In this paper, we propose a collapsed waveform detection and refinement framework for the WaveNet vocoder which is one of the state-of-the-art neural network-based vocoders. Although the WaveNet vocoder generates speech with high-fidelity conditioning on the natural acoustic features, it is hard to deal with the unseen acoustic features. That is, the WaveNet vocoder sometimes generates very noisy speech segments when conditioning on the outside testing features such as voice converted or speech enhanced ones. To address this problem, we first design a defective speech detector, which uses a waveform envelope detection technique to detect the collapsed speech segments. Then the WaveNet vocoder regenerate this unexpected segments with the proposed linear predictive coding (LPC) coefficients-constraint, which refine the output distortion from the WaveNet vocoder to avoid generating the collapsed speech. The verification objective evaluation results indicates the effectiveness of the proposed detection method which achieves about 10% equal error rate. Furthermore, the quality and speaker similarity subjective test are also conducted, and the results demonstrate the proposed framework can improve the speech quality while maintain the same speaker similarity as the original WaveNet vocoder. **(Published at Interspeech 2018)**

Hitoshi Suda, Gaku Kotani, Shinnosuke Takamichi, and Daisuke Saito (Tokyo University)
"A Revisit to Feature Handling for High-quality Voice Conversion Based on Gaussian Mixture Model"
Abstract: This paper discusses influences of handling acoustic features on the quality of generated sounds in voice conversion systems based on Gaussian mixture models. This paper also introduces an alternative wave generation method, which is named SP-WORLD, inspired by WORLD vocoder framework, and which outperforms conventional MLSA filtering in some cases. **(Published at APSIPA 2018)**

Hieu-Thi Luong and Junichi Yamagishi (NII)
"Scaling and bias codes for modeling speaker-adaptive DNN-based speech synthesis systems"
Abstract: Augmenting a speaker embedding vector to the linguistic input of a neural network is a popular method for modeling multi-speaker speech synthesis systems. This setup also allow the model to be quickly adapted to unseen speakers. However by first systematically reviewing the core principles of neural-network based speaker-adaptive models, we show that the speaker embedding method is constrained by its own nature as bias adaptation. Furthermore we propose to expand the concept of the speaker embedding to scaling and bias operations in order to add a new degree of freedom for adaptation process. The experiment results showed that the proposed method improved the performance of speaker adaptation to the unseen speaker compared with the conventional input code method. **(Published at SLT 2018)**

Patrick Lumban Tobing, Tomoki Hayashi, Yi-Chiao Wu, Kazuhiro Kobayashi, and Tomoki Toda (Nagoya University)
"Voice Conversion with Fine-Tuned WaveNet based on Concatenated Spectral Mappings using Recurrent Neural Network"
Abstract: In this work, we propose a voice conversion (VC) framework with the use of concatenated recurrent neural network (RNN)-based spectral mappings and finely-tuned WaveNet vocoder. It is well known that with the use of distorted (oversmoothed) speech features, such as spectral parameters estimated from a statistical mapping model, WaveNet suffers from quality degradation. This is due to the mismatches between the natural spectral parameters used in developing the WaveNet model and the estimated features used in the generation time. In VC, it is not straightforward to use oversmoothed features in WaveNet development because the time-sequence alignment of the speech signals between the source and the target speakers is different.

To overcome this issue, we propose to develop RNN-based spectral mapping models for each of the target-to-source mapping and the source-to-target mapping. Hence, to obtain the over-smoothed features for WaveNet development, the target-to-source and the source-to-target mapping models are concatenated to produce estimated target features with the alignment of the target speaker. A pre-trained WaveNet model is then fine-tuned to be adapted with the over-smoothed target spectral features. In the generation time, the source-to-target mapping model is used to generate estimated spectral features to be fed into the fine-tuned WaveNet vocoder. The experimental results demonstrate the effectiveness of the proposed method in improving the naturalness of the converted waveform, even if compared with the use of a post-conversion processing, based on spectrum differential and global variance, which is used to alleviate the over-smoothing. **(Published at SLT 2018)**

Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)
"Exemplar-based Lip-to-Speech Synthesis Using Convolutional Neural Networks"
Abstract: This paper proposes a neural network-based lip-to-speech synthesis approach that converts "unvoiced" lip movements to "voiced" utterances. We build on our recently proposed exemplar-based non-negative matrix factorization approach by addressing several of its shortcomings. First, the original model imposes unnatural constraints on the preprocessing of visual features in order to satisfy the non-negativity constraint of NMF. Second, there is a possibility that an activity matrix cannot be shared between the visual and the audio feature in an NMF-based approach. To tackle these problems, in this paper, we propose a new method that employs convolutional neural networks to convert visual features into audio features, and also integrates an exemplar-based approach into the neural networks in order to combine the advantages of our proposed approach with the flexibility of neural network approaches. Experimental results showed that our proposed method produced more proper spectra than conventional NMF-based methods. **(Accepted for IW-FCV 2019)**

企画２：Interspeech2018報告会 (口頭発表)

SIG-SLPの2月の研究会では、ここ数年、INTERSPEECHの論文紹介セッションを開催しています。2019年の2月研究会では、９月にインドにて開催された国際会議Interspeech2018において発表された以下の論文を紹介する予定です。

・音源分離

担当：俵直弘(早稲田大学)

紹介予定論文

1. Z. Wang et al, End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction
2. Z. Meng et al, Cycle-consistent speech enhancement
3. Z. Wang et al, Integrating spectral and spatial features for multi-channel speaker separation

・音声認識

担当：増村亮(日本電信電話株式会社)

鈴木雅之(日本IBM)

福田隆(日本IBM)

紹介予定論文

4. S. Kim et. al, Improved Training for Online End-to-end Speech Recognition Systems
5. R. Pang et. al, Compression of End-to-End Models
6. A. Zeyer et. al, Improved Training of End-to-end Attention Models for Speech Recognition
7. A. Sriram et al, Cold Fusion: Training Seq2Seq Models Together with Language Models
8. KC. Sim et al, Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition
9. J. Andrés-Ferrer et al, Efficient language model adaptation with Noise Contrastive Estimation and Kullback-Leibler regularization
10. B. Milde et al, Unspeech: Unsupervised Speech Context Embeddings
11. D. Liu et al, Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings
12. .Y Chung et al, Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech

・医療・支援技術

担当：越智　景子(東京工科大学)

紹介予定論文

13. P. Kothalker et al, Fusing text-dependent word-level i-vector models to screen 'at risk' child speech
14. B. Mirheidari et al, Detecting signs of dementia using word vector representations

・話者照合，感情認識

担当：安藤厚志(日本電信電話株式会社),
塩田さやか(首都大学東京)

紹介予定論文

15. Y. Zhu et al, Self-attentive Speaker Embeddings for Text-Independent Speaker Verification,
16. Z. Huang et al, Angular Softmax for Short-Duration Text-independent Speaker Verification
17. W. Ding et al, MTGAN: Speaker Verification through Multitasking Triplet Generative Adversarial Networks
18. H. Murthy et al, Decision-level featureswitching as a paradigm for replay attack detection
19. C. Wijenayake et al, Modulation Dy-namic Features for the Detection of Replay Attacks
20. E. Ambikairajah et al, Deep Siamese Architecture Based Re-play Detection for Secure Voice Biometric,
21. M. Sarma et al, Emotion Identification from Raw Speech Signals Using DNNs,
22. P. Yenigalla et al, Speech Emotion Recognition Using Spectrogram & Phoneme Embedding
23. S. Sahu et al, On Enhancing Speech Emotion Recognition using Generative Adversarial Networks
24. S. Parthasarathy et al, Ladder Networks for Emotion Recognition: Using Unsupervised Auxilary Tasks to Improve Predictions of Emotional Attributes

・音声合成

担当：高木信二(NII)
玉森聡(愛知工業大学),
沢田慶(マイクロソフトディベロップメント株式会社)

紹介予定論文

25. B. Sisman et al, A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder
26. L. Liu et al, WaveNet Vocoder with Limited Training Data for Voice Conversion
27. Y. Wu et al, Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder
28. L. Juvela et al, Speaker-independent Raw Waveform Model for Glottal Excitation
29. Z. Hodari et al, Learning Interpretable Control Dimensions for Speech Synthesis by Using External Data,
30. K. Akuzawa et al, Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder,
31. H. Li et al, EMPHASIS: An Emotional Phoneme-based Acoustic Model for Speech Synthesis System
32. A. Baird et al, Perception and analysis of the likeability and human likeness of synthesized speech
33. Ruibo Fu et al, Transfer Learning Based Progressive Neural Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis
34. K. Chen et al, High quality voice conversion using spectrogram based Wavenet

都合により、アナウンスなく紹介する論文は変更になることがあります。

一般講演 1 (口頭発表, 英語)

発表時間：2/27 13:00-13:25

郡山　知樹 (東京工業大学), 高道 慎之介 (東京大学大学院), 小林 隆夫 (東京工業大学)

“GMMNに基づく音声合成におけるグラム行列のスパース近似の検討”

Abstract: 生成的モーメントマッチングネット(GMMN)では音声のランダムな特徴を合成できるが, 目的関数であるmaximum mean discrepancy(MMD)は計算量の点から直接の計算が不可能であり近似が必要である．しかし，MMDの近似方法について詳細な検討は行われてこなかった．本研究ではMMDに用いるグラム行列をスパース近似する手法を検討し，その性能について自然性およびランダム性の観点から評価する.

(英語による発表を予定)

発表時間：2/27 13:25-13:50

Yi Liu (Tokyo Institute of Technology), Takahiro Shinozaki (Tokyo Institute of Technology)

“I-vector Domain Adaptation Using Cycle-Consistent Adversarial Networks for Speaker Recognition”

Abstract: Speaker recognition systems often suffer from severe performance degradation due to the difference between training data and evaluation data, which is called domain mismatch problem. In this paper, inspired by adversarial strategies in deep learning techniques, we propose a cycle-consistent adversarial networks-based domain adaptation method. This method performs an i-vector domain transformation from source (out) domain to target (in) domain to reduce the domain mismatch. Additionally, it uses a cycle structure that reduces the negative influence of losing speaker information in i-vectors during the transformation and makes it possible to use unpaired datasets for training. The experimental result shows that the speaker recognition system using the proposed method obtains a better performance compared with the conventional i-vector and PLDA based speaker recognition system.

発表時間：2/27 13:50-14:15

Xin Wang (NII), Shinji Takaki (NII), Junichi Yamagishi (NII)

“Investigating neural source-filter waveform model for statistical parametric speech synthesis”

Abstract: Recently we proposed the neural source-filter model (NSF) that converts a sequence of acoustic features into a speech waveform. Similar to other recent neural waveform models, the NSF is a non-autogressive model powered by dilated CNN; however, the NSF uses the sine waveform instead of the random noise as the excitation. Furthermore, without using the normalizing flow, the NSF simply optimizes the network parameters by minimizing a spectral amplitude distance. In this work, we further investigated the three issues: whether the network structure can be further simplified; whether the NSF can be applied to multi-speaker synthesis; whether the NSF can be applied to convert the linguistic features into the speech waveforms. Our experiments showed positive results on all the three points. Particularly, we found that the WaveNet-style gated activation can be safely removed and the NSF performs quite well as a pure dilated-CNN-based network.

一般講演２(口頭発表, 日本語)

発表時間：2/28 10:00-10:25

森川 寛也（クリスタルメソッド株式会社）三橋 晟（クリスタルメソッド株式会社）河合 継（クリスタルメソッド株式会社）能勢 隆（東北大学大学院工学研究科・工学部）千葉 祐弥（東北大学大学院工学研究科・工学部）

“テキストと音声のマルチモーダル感情認識”

Abstract: 本論文では，音響情報による感情識別の結果と言語情報による感情識別の結果を結合した感情認識の手法を提案する． 具体的には音声を音響的な情報と言語的な情報に分けてそれぞれの感情識別器を作り， 2つの識別結果を結合して感情識別を行うという方法である． 音響情報の感情識別器については，文章を単語ごとに区切り１単語１セルとしてLSTMへ入力し学習させる方法と， 区切らずに学習させる方法とを比較した． 結果としては現在の時点では区切らないほうが精度が高いが， 今後データ量の増加に伴い区切った方が良い精度を出すと予想できる結果となった． また，言語と音響の識別結果を結合したモデルではテストデータに対して，75.5%の精度で感情を識別することができた． これは音響情報のみと言語情報のみの識別結果よりも精度が高かった． 今後の課題としてさらに精度を上げていくには大規模コーパスを作成か， データ拡張によってデータを増やしていく必要があることが分かった．

発表時間：2/28 10:25-10:50

高道　慎之介（東京大学）猿渡　洋（東京大学）

“正弦関数摂動von Mises分布DNNのモード近似を用いた位相復元”

Abstract: Deep neural networkを用いた振幅スペクトログラムからの位相復元を提案する．正弦関数摂動von Mises分布DNNは音声の群遅延のモデル化に適しているが，その分布のモードが解析的に求まらないため，マルチタスク学習への応用が困難である．本稿では，分布のモードを微分可能かつ解析的な形で近似することで，マルチタスク学習及び高精度な位相復元を可能にする．

2018年 SIG-SLP企業賞

2018年SIG-SLP企業賞は、昨年SLP研究会で発表された学生論文のうち、もっとも優れた論文に送られる賞です。SIG-SLPのダイヤモンドレベル企業スポンサー名をお借りし、以下の最優秀学生論文２編に企業賞を授与します。選考委員会は音声言語情報処理研究会運営委員および該当企業の担当者により構成され、選考委員会委員の投票により決定されました。

Fairly Devices賞

End-to-End音声合成を用いた単語単位End-to-End音声認識のデータ拡張

上乃 聖, 三村 正人, 坂井 信輔, 河原 達也 (京都大学情報学研究科)

2018-SLP-125

2018/12/3

Yahoo! JAPAN賞

条件付き敵対的生成ネットワークを用いたデータ拡張による
対話行為分類法の検討

河野 誠也, 吉野 幸一郎, 中村 哲 (奈良先端科学技術大学院大学)

2018-SLP-125

2018/12/3

宿泊および参加料について

研究会参加者は以下の費用を音声言語情報処理研究会へお支払いください
宿泊費および1日目の夕食および2日目の朝食・昼食代：15,000円(税込)
会議室利用料金：1,500円(税込)
合計16,500円(税込)

合宿形式での参加を推奨しますが、止むを得ず宿泊無しで二日間参加する場合は以下の金額を頂戴いたします。
2日目の昼食代：1,000円(税込)
会議室利用料金：1,500円(税込)
合計 2,500円 (税込)

初日のみ参加する場合は以下の金額を頂戴いたします。
会議室利用料金：1,500円 (税込)

また併せて、以下の研究会参加費費を情報処理研究会へお支払いください
・当研究会登録会員/ジュニア会員:無料
・情報処理学会会員:¥1,500
・情報処理学会学生会員:¥500
・非 会 員:¥2,500

どちらの費用も当時会場にて現金にてお支払いください。クレジットカードは利用できません。

★情報処理学会　音声言語情報処理研究会（SLP）

主査
西村 雅史（静岡大）

幹事
福田 隆（日本IBM）
山岸 順一（NII）
塩田 さやか（首都大東京）
俵 直弘（早稲田大）

問い合わせ先
国立情報学研究所 山岸順一
jyamagis at nii.ac.jp

SLPスポンサー

Fairy Devices

YAHOO! JAPAN

IBM Research

株式会社エーアイ