Vol.2019-SLP-126 No.7
2019/2/27

SLP研究会の新たな試み：国際会議既発表セッション

山岸　順一[1]　安田　裕介[1]　Yi Zhao[1]　Tifani Warnita[2]　Fuming Fang[1]　Yilong Peng[2]
田中 智宏[2]　Bairong Zhuang[2]　Yi-Chiao Wu[3]　須田仁志[4]　Hieu-Thi Luong[1]
Patrick Lumban Tobing[3]　高島悠樹[5]

概要：情報処理学会 音声言語情報処理研究会では、原稿の執筆等をせずに気軽に発表ができることを目指し、「国際会議既発表セッション」という新たな取り組みを試験的に開始した。国内学会・研究会では未発表だが、最近の国際会議 (Interspeech 2018，APSIPA 2018，SLT 2018，ICASSP 2019 等）やジャーナルや Arxiv 等で発表済み，投稿済みである論文を紹介する位置付けである。

1.　はじめに

情報処理学会 音声言語情報処理研究会では、原稿の執筆等をせずに気軽に研究紹介ができることを目指し、「国際会議既発表セッション」という新たな取り組みを試験的に開始した。国内学会・研究会では未発表だが、最近の国際会議 (Interspeech 2018，APSIPA 2018，SLT 2018，ICASSP 2019 等）やジャーナルや Arxiv 等で発表済み，投稿済みである論文を紹介する位置付けである。

2019 年 2 月に開催する情報処理学会 音声言語情報処理研究会で紹介する論文は以下の通りである。

2.　紹介論文 1

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)

紹介者：安田裕介

[1] 国立情報学研究所
[2] 東京工業大学
[3] 名古屋大学
[4] 東京大学
[5] 神戸大学

3. 紹介論文 2

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)

紹介者：Yi Zhao

4. 紹介論文 3

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)

紹介者：Tifani Warnita

5. 紹介論文 4

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)

紹介者：Fuming Fang

6. 紹介論文 5

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)

紹介者：Yilong Peng

7. 紹介論文 6

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)

紹介者: 田中 智宏

8. 紹介論文 7

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)

紹介者: Bairong Zhuang

9. 紹介論文 8

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)

紹介者: Yi-Chiao Wu

10. 紹介論文 9

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)

紹介者: 須田仁志

11. 紹介論文 10

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)

紹介者: Hieu-Thi Luong

12. 紹介論文 11

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)

紹介者: Patrick Lumban Tobing

13. 紹介論文 12

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)

紹介者: 高島悠樹