様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ２ 日現在

機関番号：６２６１５
研究種目：研究活動スタート支援
研究期間：2019～2020
課題番号：１９Ｋ２４３７２

研究課題名（和文）Encoder Factorization for Capturing Dialect and Articulation Level in End-to-End Speech Synthesis

研究課題名（英文）Encoder Factorization for Capturing Dialect and Articulation Level in End-to-End Speech Synthesis

研究代表者

Ｃｏｏｐｅｒ　Ｅｒｉｃａ（Cooper, Erica）

国立情報学研究所・コンテンツ科学研究系・特任助教

研究者番号：３０８４３１５６

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：現在の音声合成技術は自然な音声を生成することが可能であるが、方言など目標話者の特性を完全に再現する事は困難である。本研究では、どの種類の話者埋め込み表現が最も効果的に話者性を再現するかについて調査を行い、Learnable DictionaryEncodingが最もうまく機能することを確認した。 同様の方言埋め込み表現が、合成音声の方言を改善するのに役立つことも確認。 最後に、人工的に作成した学習デ ータと理想的ではない録音条件の音声データの両方を使用したデータ拡張方法についても調査し、これを用いることでモデルから予測された合成音声の自然さがさらに改善することも示した。

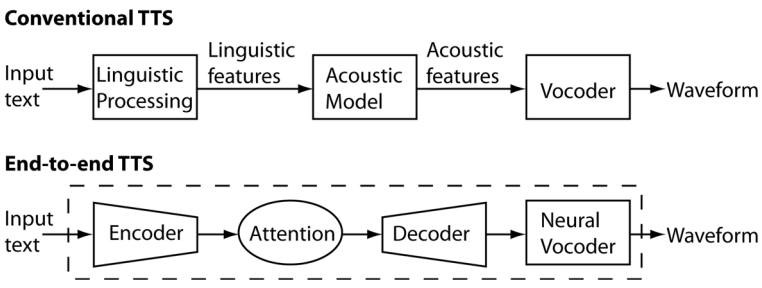
研究成果の学術的意義や社会的意義
本課題では、end-to-end音声合成における合成音声の話者性や方言再現性の向上のため、エンコーダの因子を制御する方法を調査した。話者の個性や特性をより適切に再現することにより、より多くの目標話者を音声合成システムにおいて利用することが可能になり、技術の応用先が広がると期待される。

研究成果の概要（英文）：Synthesizing speech in many voices and styles has long been a goal in speech research. While current state-of-the-art synthesizers can produce very natural sounding speech, matching the voice of a target speaker when only a small amount of that speaker's data is available is still a challenge, especially for characteristics such as dialect. We conducted experiments to determine what kind of speaker embeddings work best for synthesis in the voice of a new speaker, and found that Learnable Dictionary Encoding (LDE) based speaker representations worked well, based on a crowdsourced listening test. We also found that similarly obtaining LDE-based dialect representations helped to improve the dialect of the synthesized speech. Finally, we explored data augmentation techniques using both artificially modified data as well as real data from non-ideal recording conditions, and found that including the found data in model training could further improve naturalness of synthesized speech.

研究分野：Text-to-speech synthesis

キーワード：Speech synthesis　Speaker modeling　Deep learning　Neural network

１．研究開始当初の背景　Research Background



Text-to-speech (TTS) synthesis is the task of predicting and generating speech audio from input text, and end-to-end neural methods have become the state of the art in producing high-quality, natural-sounding synthetic speech. Common end-to-end TTS architectures use an encoder-decoder model with attention, in contrast to the modular approach which uses separate models for linguistic processing, acoustic modeling, and waveform generation. While traditional approaches to TTS typically require large amounts of high-quality, neutral-style speech from a single professional speaker, end-to-end TTS allows for the use of more diverse sources of speech audio as training data, opening possibilities for new TTS applications in a much wider variety of voices and styles.

Multi-speaker modeling for TTS has long been an area of active research. At the time when this project was proposed in 2019, an approach called *decoder factorization*, wherein a speaker-specific representation such as a speaker embedding is included as input to the decoder, was a successful approach for multi-speaker modeling for end-to-end TTS. However, it was observed that while acoustic speaker characteristics could be successfully captured in this manner, nuances such as dialect and characteristic prosody were not captured. This is because the decoder handles low-level acoustic characteristics whereas dialect and prosody are longer-range linguistic effects. Thus, we proposed *encoder factorization* to better model these phenomena and to better model speakers and their characteristic speaking styles.

２．研究の目的　Research Motivation and Goals

We had three main goals that all relate to the overall goal of improving speaker modeling for multi-speaker TTS:

Goal 1: Encoder factorization. We hypothesize that inputting a speaker representation at the encoder instead of or in addition to the decoder will improve speaker similarity by better capturing the phonetic and longer-range aspects of speaker identity. We also hypothesize that we can use state-of-the-art speaker identification models to extract speaker embeddings that capture speaker characteristics well.

Goal 2: Dialect modeling. We aim to better capture longer-range speaker characteristics such as dialect by creating dialect representations in a similar manner to how the speaker representations are created and using them as additional information about speaker characteristics to input to TTS.

Goal 3: Improved synthesis of unseen speakers. Although we originally proposed to model level of articulation as one more characteristic of speech, we discovered during this project that a more interesting challenge is to improve zero-shot synthesis of unseen speakers, where the similarity of synthesized speech does not reach the level of synthesis of the speakers that were seen during training. We hypothesize that this is due to overfitting to seen speakers during TTS training, and that data augmentation may mitigate this.

３．研究の方法　Research Methods

For goal 1, we trained multi-speaker end-to-end TTS models with speaker embeddings input at three different locations: at the encoder, concatenated with the self-

attention and CBH-LSTM outputs; at the prenet to the decoder; and at the postnet. We compared these approaches to the standard method of inputting them only at the prenet to the decoder.

For goal 2, we trained dialect embedding models in the same manner as our speaker embedding models and input the resulting dialect embeddings to the TTS models along with the speaker embeddings to improve synthesis of target speakers' particular dialects. We also included channel labels to model the worse recording conditions and to select the best channel settings at synthesis time.

For goal 3, we tried a synthetic data augmentation approach wherein the original TTS training data was perturbed to create new artificial speaker identities, and also a "found data" augmentation approach wherein we trained the TTS model on a larger set of speakers and dialects by adding data of lower quality than that which is typically used for training synthesis models, and also by modeling the channel factor in order to avoid hurting synthesis quality by training on the lower-quality data.

4. 研究成果 Research Outcomes

Goal 1: Encoder factorization using learnable dictionary encoding (LDE) embeddings can produce natural-sounding synthesized speech with good similarity to the target speaker (results published at ICASSP 2020)

In zero-shot speaker adaptation for end-to-end TTS, we extract speaker embeddings from a pre-trained speaker identification model and input them to the TTS model as a speaker representation. Our end-to-end multi-speaker TTS model architecture is based on Tacotron [1], extended with self-attention as described in [2] to better capture long-range dependencies. Figure 1 shows our multi-speaker Tacotron architecture with our three proposed input locations for speaker embeddings. Our model takes a sequence of phonemes or graphemes as input, and outputs a Mel spectrogram.

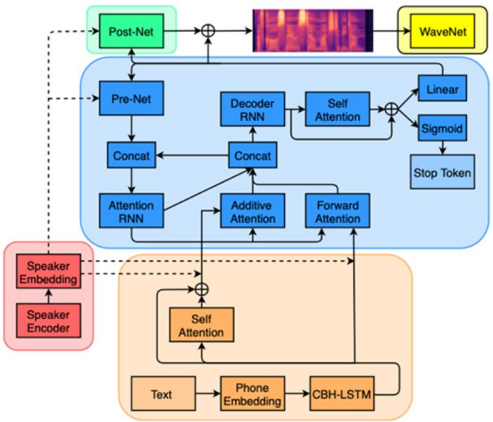


Figure 1: Architecture of Multi-Speaker Tacotron

We trained multi-speaker end-to-end TTS models with speaker embeddings input at three different locations: concatenated with the two encoder outputs, at the prenet to the decoder, and at the postnet. We found that inputting speaker embeddings at the encoder in addition to the decoder prenet produced synthesized speech with better speaker similarity for unseen speakers than inputting only to the decoder prenet, as measured objectively by cosine similiarity to the reference speaker. We also found that training gender-dependent models helped to improve speaker similarity. Cosine similarities for seen (train) and unseen (dev) speakers for different training configurations and embedding input locations can be seen in the table in Figure 2.

Input location	Gender-ind		Gender-dep	
	train	dev	train	dev
pre	0.357	0.402	0.438	0.361
attn	0.709	0.490	0.711	0.476
pre+attn	0.676	0.489	0.708	**0.533**
pre+attn+post	0.684	0.480	0.717	0.477

Figure 2: Cosine similarities to the target speaker using different training approaches and speaker embedding input locations

We also experimented with different types of LDE-based [3] speaker embeddings and compared them with the popular x-vector representation [4]. We explored different model settings and training criteria such as whether to use regular softmax or angular softmax as the loss function, whether to use mean pooling only or standard deviation pooling as well, different dimension settings for the extracted embeddings, and whether to normalize the embeddings as a postprocessing step. We found that 512-dimensional embeddings extracted from a model using mean pooling only, angular softmax, and no normalization not only had the best results of the models we tried by speaker identification metrics such as equal error rate and minimum detection cost, but that

these embeddings also produced synthesized speech with the best speaker similarity for unseen speakers, indicating their effectiveness as speaker representations for multiple tasks.

Goal 2: Dialect embeddings can improve dialect similarity of a target speaker (results published at Interspeech 2020)

We trained dialect identification models in the same manner as our LDE speaker identification models, using data that was labeled for English dialects. Then, for a given target speaker, we extract both their speaker embedding and their dialect embedding from their enrollment utterance, and input both to the TTS model. Figure 3 shows the TTS model that uses both dialect and speaker embeddings. We found that the use of dialect embeddings did help the synthesized speech to better match the perceived dialect of the target speaker, according to a crowdsourced listening test.

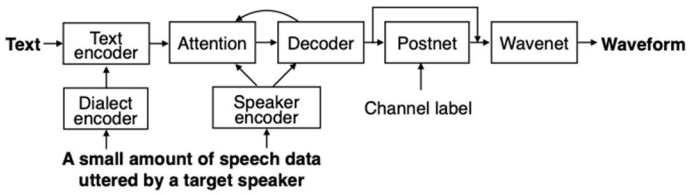


Figure 3: End-to-end text-to-speech architecture with both dialect and speaker embeddings

Goal 3: Data augmentation using found data and channel modeling can improve synthesis quality (results published at Interspeech 2020)

We tried two different methods of data augmentation to create more speaker variety for training TTS models and reduce overfitting to a smaller number of training speakers. The first approach we tried was a type of artificial speaker augmentation based on vocal tract length perturbation (VTLP) [5] wherein we simply speed up and slow down the training data by resampling. The resulting signals have different fundamental frequency, speaking rate, formants, and spectra, and thus effectively sound like a different speaker. Our second method of data augmentation was to include more data in training by using lower-quality data collected for purposes other than TTS, such as speech recognition. This data does not meet our usual requirements for TTS: it may contain background noise, reverberation, or other problems. So, to avoid degrading the quality of the output synthesized speech by training the model on this kind of data, we included a one-hot channel label that indicates which dataset an utterance comes from, and this label is input to Tacotron's postnet, which controls spectral shaping and enhancement. At synthesis time, we choose the highest-quality channel setting to produce clear audio. Results from a crowdsourced listening test revealed that using the low-quality data for augmented training was effective, but contrary to our expectations, *naturalness of seen speakers* was improved instead of speaker similarity of unseen speakers. This suggests that improving speaker similarity of unseen speakers remains a challenge.

Open-source code and audio samples:

Our open-source multi-speaker Tacotron implementation is available here:
https://github.com/nii-yamagishilab/multi-speaker-tacotron

Audio samples for multi-speaker Tacotron using different types of speaker embeddings:
https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/

Audio samples for multi-speaker Tacotron with data augmentation and dialect embeddings:
https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/augment.html

Other research outcomes are detailed in our published papers.

[1] Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," INTERSPEECH 2017.
[2] Y. Yasuda et al., "Investigation of Enhanced Tacotron Text-to-Speech Synthesis Systems with Self-Attention for Pitch Accent Language," ICASSP 2019.
[3] W. Cai et al., "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," Speaker Odyssey 2018.

[4] D. Snyder et al., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP 2018.
[5] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition," ICML Workshop on Deep Learning for Audio, Speech and Language, 2013.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　3件／うちオープンアクセス　3件）

１．著者名 Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, Junichi Yamagishi	４．巻 -
２．論文標題 Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings	５．発行年 2020年
３．雑誌名 ICASSP 2020	６．最初と最後の頁 6184-6188
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICASSP40776.2020.9054535	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Cooper Erica、Lai Cheng-I、Yasuda Yusuke、Yamagishi Junichi	４．巻 -
２．論文標題 Can Speaker Augmentation Improve Multi-Speaker End-to-End TTS?	５．発行年 2020年
３．雑誌名 Interspeech 2020	６．最初と最後の頁 3979-3983
掲載論文のDOI（デジタルオブジェクト識別子） 10.21437/Interspeech.2020-1229	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Kato Shuhei、Yasuda Yusuke、Wang Xin、Cooper Erica、Takaki Shinji、Yamagishi Junichi	４．巻 8
２．論文標題 Modeling of Rakugo Speech and Its Limitations: Toward Speech Synthesis That Entertains Audiences	５．発行年 2020年
３．雑誌名 IEEE Access	６．最初と最後の頁 138149～138161
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2020.3011975	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Multi-speaker Tacotron Code
https://github.com/nii-yamagishilab/multi-speaker-tacotron
Audio Samples for Multi-Speaker Tacotron
https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/
Audio sample page for Interspeech 2020 paper
https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/augment.html

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
米国	Massachusetts Institute of Technology			