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　６ 年　 ５ 月 １４ 日現在

機関番号：６２６１５
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１９５１

研究課題名（和文）Language-independent, multi-modal, and data-efficient approaches for speech synthesis and translation

研究課題名（英文）Language-independent, multi-modal, and data-efficient approaches for speech synthesis and translation

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・特任准教授

研究者番号：３０８４３１５６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：音声合成(TTS)、データ効率の良いTTS、TTS品質予測のためのプルーニングについて検討した。出力品質を低下させることなく、TTSモデルの重みの90%までを刈り込んだ。ポッドキャストデータを用いた低リソース言語のTTSコーパス構築のためのデータ処理を開発し、高品質な一般公開データセットを得た。また、このデータを利用したTTSシステムを開発し、同様のデータを持つあらゆる言語に再利用できるようにした。 新しい言語に微調整できる多言語TTSのための自己教師付き音声表現を研究した。自動TTS評価のための一連のチャレンジを開始し、多くの参加者を集め、この分野を発展させた。

研究成果の学術的意義や社会的意義
We developed TTS trainable on small amounts of data and lightweight TTS models. We also advanced the field of TTS evaluation. This benefits researchers and society by reducing barriers of entry to creating TTS for low-resource languages, expanding accessibility benefits of TTS to a broader audience.

研究成果の概要（英文）：We explored pruning for lightweight text-to-speech synthesis (TTS), developed data-efficient TTS for low-resource languages, and advanced the field of automatic quality prediction for TTS. We found that up to 90% of TTS model weights can be pruned without reducing output quality. We developed a data processing pipeline for building TTS corpora for low-resource languages using podcast data, resulting in a large-scale, high-quality, publicly-available dataset. We also developed a TTS system using this data that can be repurposed for any language with similar data. As self-supervised speech representations have been effective for many downstream tasks, we next investigated these as an intermediate representation for TTS trained on multilingual data, which can be fine-tuned to a new language. Finally, we identified automatic evaluation of TTS as a critical topic. We launched a series of challenges for this task in 2022 and 2023 which attracted many participants and advanced the field.

研究分野：Text-to-speech synthesis

キーワード：Text-to-speech synthesis　Low-resource languages　Neural network pruning　Evaluation

１．研究開始当初の背景

End-to-end neural models for speech and language tasks have had tremendous success in advancing the state of the art, however the quantity and quality of data required to train such models well can be prohibitive to those who wish to create systems for new domains, languages, dialects, or styles. For instance, very high-quality end-to-end text-to-speech (TTS) synthesizers have typically been trained on GPUs on 20 or more hours of high-quality recordings of one professional speaker. Furthermore, evaluating speech synthesis models typically requires conducting listening tests with native listeners of the language, a costly and time-consuming procedure which can be especially difficult for the case of low-resource languages. In order to encourage the creation of high-quality natural language tools for more diverse use cases such as under-resourced languages and dialects, it is important to lower the barrier to creation of these systems by reducing the difficulties of data requirements, computational resources, and evaluation.

２．研究の目的

The purpose of this project was to develop speech and language technologies that can make use of smaller amounts of training data from less traditional but more readily available data sources, and to find ways to make these models smaller and more lightweight to reduce the computational requirements to run them. During the course of this project, we also identified evaluation of TTS output as a critical bottleneck for research, so we additionally developed more automatic speech quality evaluation methodologies and advanced the field through a series of shared-task challenges. The outcomes are expected to be beneficial to both researchers and society by reducing the barrier to entry of creating speech technologies for new applications in lower-resourced languages and dialects, thus expanding the accessibility benefits of these technologies to a broader audience.

３．研究の方法

(1) We investigated neural network pruning techniques to develop more lightweight speech synthesis architectures.

(2) We explored the use of podcast data to build TTS datasets and systems, as well as the use of representations from self-supervised learning based (SSL) models for speech as an intermediate representation for building multilingual TTS systems that can be adapted to a new language using only a small amount of data.

(3) We further explored the use of SSL models for the task of opinion score prediction of synthesized speech and ran a series of shared-task challenges on this topic to further advance the field.

４．研究成果

(1) **Neural network pruning of TTS models:** We explored the effects of pruning end-to-end neural network based models for single-speaker text-to-speech (TTS) synthesis and evaluated the effects of different levels of pruning on naturalness, intelligibility, and prosody of the synthesized speech. The pruning method used follows a “prune, adjust, and re-prune” (PARP) approach that had previously been found to be effective for self-supervised speech models used in automatic speech recognition. We found that both neural vocoders and neural text-to-mel speech synthesis models are highly over-parametrized, and that up to 90% of the model weights can be pruned without detriment to the quality of the output synthesized speech. This was determined based on experiments pruning multiple

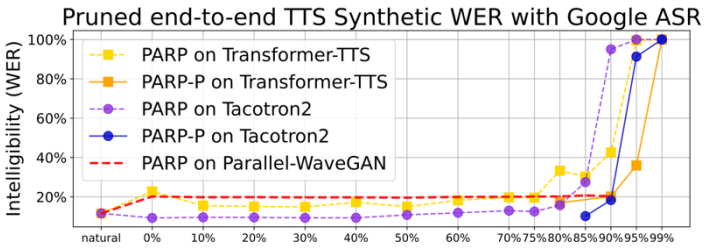

Figure 1: WERs of synthesized speech over sparsities for different model combinations.

types of neural acoustic models including Tacotron and Transformer TTS, as well as the Parallel WaveGAN neural vocoder, and evaluated both using objective metrics such as Word Error Rate from an automatic speech recognizer and measures of f0 and utterance duration, as well as by subjective Mean Opinion Score and A/B comparison listening tests. Figure 1 shows that word error rates (WER) measured by an automatic speech recognizer only start to increase at around an 80-90% pruning level using PARP and a progressive pruning approach (PARP-P), showing that intelligibility only starts to degrade at very high levels of pruning. Pruned models are smaller and thus more computationally- and space-efficient, and our results address one of our proposed aims to find more efficient neural TTS models. This work resulted in one peer-reviewed conference publication at ICASSP 2022 [1].

(2) **Methods for building TTS corpora and developing synthesis models for low-resource languages using podcast data:** Using the Hebrew language as a case study, we developed a data processing pipeline and identified suitable model architectures for developing TTS models using podcast data, resulting in a large-scale, high-quality, and publicly-available Hebrew TTS dataset and open-source code that can be repurposed and generalized to other low-resource languages. We also addressed the issue of best practices for conducting evaluation when no comparable prior TTS systems exist and when automatic speech recognizers (ASR) for the language in question may not be perfectly accurate by measuring *differences* in ASR transcription accuracy between natural speech and the same utterances produced by TTS, an approach that can generalize to any language regardless of whether state-of-the-art ASR is available. This work resulted in one peer-reviewed conference publication at Interspeech 2023 [2].

(3) **Multilingual TTS using self-supervised learning based representations that can adapt to new languages:** Self-supervised learning based (SSL) representations for speech have demonstrated remarkable usefulness for many downstream speech-related tasks, and have been shown to contain phonetic information. We therefore chose these as an intermediate representation for multilingual, multi-speaker text-to-speech synthesis trained on data from many languages, which can then be fine-tuned to a new language using only a small amount of data. Our approach produced synthesized speech with good speaker similarity and naturalness for both seen and unseen speakers for languages which were included in the training data. Our model also was able to synthesize intelligible speech with good speaker similarity to the target speaker's voice for languages which were completely unseen during training. This work is currently under review for journal publication [3], and continuing experiments to determine how the amount of adaptation data required for good synthesis relates to factors like language family and the degree of phonetic similarity between the target language and the languages included in training is ongoing work.

(4) **Automatic opinion score prediction for synthesized speech:** TTS systems are typically evaluated with human listening tests, using paradigms such as the Mean Opinion Score (MOS), in which listeners are presented with synthesized audio samples one by one and asked to rate some aspect of the audio, such as naturalness, on a numerical rating scale. Proposed synthesis systems are compared by gathering all of the ratings for all of the test samples generated by each system and computing their average. While this type of evaluation is the gold standard, it is very costly and time-consuming, representing a significant bottleneck for experimental iteration. The challenges of this kind of evaluation are amplified in the case of low-resource languages, since it may be more difficult to find native listeners of these languages to participate in listening tests. We therefore investigated automatic MOS prediction using paired data of synthesized speech samples and their ratings. Observing the effectiveness of SSL speech models for many downstream tasks, we investigated the fine-tuning of SSL models for the MOS prediction task and demonstrated their superior generalization ability to MOS prediction for synthesized speech in different languages over a prior data-driven MOS prediction approach which did not make use of SSL [4]. We further explored the speech quality information encoded in SSL models without fine-tuning that can be derived from uncertainty measures [5] and extended our prediction model to use a ranking-based approach [6]. This line of research resulted in three peer-reviewed conference publications.

(5) **The VoiceMOS Challenge shared task for MOS prediction:** Having identified the

development of more automatic evaluation methodologies for speech synthesis as critical for speech synthesis, particularly for low-resource languages, we launched a series of challenges on this topic called the VoiceMOS Challenge. We ran challenges in 2022 and 2023 during the project period, and a 2024 edition of the challenge is currently ongoing as well.

In the 2022 edition of the challenge [7], we released a large-scale dataset of English TTS and voice conversion samples along with their MOS ratings, along with three different open-source baseline systems including the SSL-based one that we previously developed [4]. We also had an "out-of-domain" track which used a much smaller amount of data from a completely separate listening test for Chinese-language TTS. The challenge attracted 22 international teams from academia and industry, and results demonstrated the overwhelming effectiveness of fine-tuning SSL models for this task in comparison with using features from frozen SSL models or not using SSL at all. The results of the OOD track also revealed that correctly predicting opinions of systems which were not present in the training data remains significantly more challenging than prediction of ratings for synthesized samples generated by systems which were included in the training data.

In the 2023 edition of the challenge [8], we shifted our focus to more diverse and challenging OOD scenarios. We collaborated with the organizers of the Blizzard Challenge 2023, which was a challenge for developing TTS using French audiobook data, and the Singing Voice Conversion Challenge, which focused on the development of singing voice conversion systems using both sung and spoken voice samples to convert speaker identity. After participants of these challenges submitted their synthesized samples, but before listening tests to rate these samples were completed, we shared these samples with VoiceMOS Challenge participants so that they could attempt to predict the outcomes of the listening test in a real-world prediction scenario. No in-domain MOS-labeled training data was provided to participants, and once the human listening tests were completed, we evaluated the prediction accuracy of our teams. The French TTS made up two tracks of our challenge (one for speaker-dependent TTS and one for speaker-adaptive TTS, following the track design of the Blizzard Challenge), the singing voice conversion was the third track, and we also had a fourth track for noisy and enhanced Chinese-language speech for which we did provide some labeled data to participants to develop their systems, since evaluation of speech enhancement is a very closely-related task to the evaluation of synthesized speech. Ten teams from academia and industry participated. One surprising result was the good prediction accuracy for the singing voice conversion track, despite none of the teams using any singing data to develop their predictors. This indicates that the domain gap between singing and spoken synthesis was not as large as we had expected. Another interesting result was that none of the teams had consistent prediction accuracy across all of the tracks, and teams that made good predictions for all tracks indicated that they developed different systems for different tracks. This indicates that general-purpose MOS prediction is still an open research question, motivating future editions of the challenge. The ten teams' prediction correlations with the real listening test results, along with predictions from two baseline systems, are shown in Figure 2.

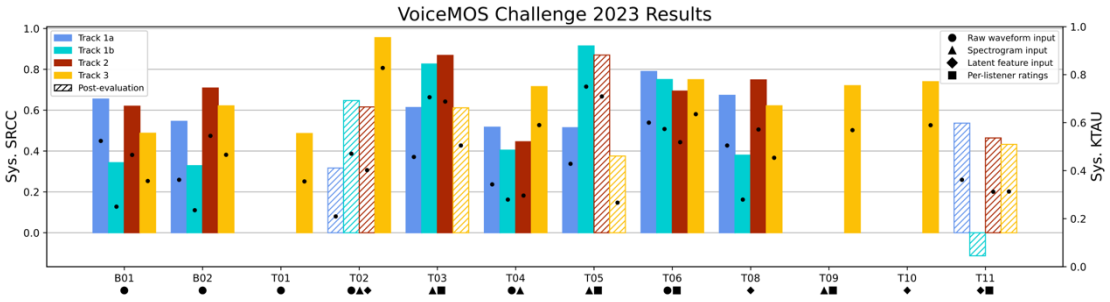

Figure 2: Results for the VoiceMOS Challenge 2023. Bars indicate system-level Spearman rank correlation coefficients (SRCC), and dots indicate Kendall Tau correlation (KTAU). Hatched bars indicate that the results were received during the post-evaluation phase after the challenge officially ended. Types of input that each team used are marked under their team ID.

[1] Cheng-I Jeff Lai, Erica Cooper, Yang Zhang, Shiyu Chang, Kaizhi Qian, Yi-Lun Liao, Yung-Sung Chuang, Alexander H. Liu, Junichi Yamagishi, David Cox, James Glass, "On the Interplay Between Sparsity, Naturalness, Intelligibility, and Prosody in Speech Synthesis," IEEE ICASSP 2022.

[2] Orian Sharoni, Roee Shenberg, Erica Cooper, "SASPEECH: A Hebrew Single Speaker Dataset for Text to Speech and Voice Conversion," ISCA Interspeech 2023.

[3] Cheng Gong, Xin Wang, Erica Cooper, Dan Wells, Longbiao Wang, Jianwu Dang, Korin Richmond, Junichi Yamagishi, "ZMM-TTS: Zero-shot Multilingual and Multispeaker Speech Synthesis Conditioned on Self-supervised Discrete Speech Representations," https://arxiv.org/abs/2312.14398 (under review).

[4] Erica Cooper, Wen-Chin Huang, Tomoki Toda, Junichi Yamagishi, "Generalization Ability of MOS Prediction Networks," IEEE ICASSP 2022.

[5] Aditya Ravuri, Erica Cooper, Junichi Yamagishi, "Uncertainty as a Predictor: Leveraging Self-Supervised Learning for Zero-Shot MOS Prediction," IEEE ICASSP 2024 Workshop on Self-Supervision in Audio, Speech and Beyond.

[6] Hemant Yadav, Erica Cooper, Junichi Yamagishi, Sunayana Sitaram, Rajiv Ratn Shah, "Partial Rank Similarity Minimization Method for Quality MOS Prediction of Unseen Speech Synthesis Systems in Zero-Shot and Semi-Supervised Setting," IEEE ASRU 2023.

[7] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi, "The VoiceMOS Challenge 2022," ISCA Interspeech 2022.

[8] Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi, "The VoiceMOS Challenge 2023: Zero-Shot Subjective Speech Quality Prediction for Multiple Domains," IEEE ASRU 2023.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Cooper Erica、Huang Wen-Chin、Tsao Yu、Wang Hsin-Min、Toda Tomoki、Yamagishi Junichi	４．巻 advpub
２．論文標題 A review on subjective and objective evaluation of synthetic speech	５．発行年 2024年
３．雑誌名 Acoustical Science and Technology	６．最初と最後の頁 1-26
掲載論文のDOI（デジタルオブジェクト識別子） 10.1250/ast.e24.12	査読の有無 無
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12件（うち招待講演　3件／うち国際学会　9件）

１．発表者名 Erica Cooper, Wen-Chin Huang, Tomoki Toda, Junichi Yamagishi
２．発表標題 Generalization Ability of MOS Prediction Networks
３．学会等名 ICASSP 2022（国際学会）
４．発表年 2022年

１．発表者名 Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, Tomoki Toda
２．発表標題 LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech
３．学会等名 ICASSP 2022（国際学会）
４．発表年 2022年

１．発表者名 Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi
２．発表標題 The VoiceMOS Challenge 2022
３．学会等名 Interspeech 2022（国際学会）
４．発表年 2022年

１．発表者名
Erica Cooper

２．発表標題
The VoiceMOS Challenge: Data-Driven Mean Opinion Score Prediction for Synthesized Speech

３．学会等名
2022 Autumn Meeting of the Acoustical Society of Japan（招待講演）

４．発表年
2022年

１．発表者名
Erica Cooper

２．発表標題
Objective Evaluation in TTS

３．学会等名
KTH Seminar on Speech Synthesis Evaluation, KTH Royal Institute of Technology, Department of Speech, Music, and Hearing（招待講演）

４．発表年
2022年

１．発表者名
Erica Cooper, Wen-Chin Huang

２．発表標題
The VoiceMOS Challenge 2022

３．学会等名
Special Interest Group on Spoken Language Processing, Information Processing Society of Japan（招待講演）

４．発表年
2022年

１．発表者名
Cheng-I Jeff Lai, Erica Cooper, Yang Zhang, Shiyu Chang, Kaizhi Qian, Yi-Lun Liao, Yung-Sung Chuang, Alexander H. Liu, Junichi Yamagishi, David Cox, James Glass

２．発表標題
On the Interplay Between Sparsity, Naturalness, Intelligibility, and Prosody in Speech Synthesis

３．学会等名
ICASSP 2022（国際学会）

４．発表年
2022年

１．発表者名
Orian Sharoni, Roee Shenberg, Erica Cooper

２．発表標題
SASPEECH: A Hebrew Single Speaker Dataset for Text to Speech and Voice Conversion

３．学会等名
Interspeech 2023（国際学会）

４．発表年
2023年

１．発表者名
Erica Cooper, Junichi Yamagishi

２．発表標題
Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech

３．学会等名
Interspeech 2023（国際学会）

４．発表年
2023年

１．発表者名
Hemant Yadav, Erica Cooper, Junichi Yamagishi, Sunayana Sitaram, Rajiv Ratn Shah

２．発表標題
Partial Rank Similarity Minimization Method for Quality MOS Prediction of Unseen Speech Synthesis Systems in Zero-Shot and Semi-supervised Setting

３．学会等名
ASRU 2023（国際学会）

４．発表年
2023年

１．発表者名
Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi

２．発表標題
The VoiceMOS Challenge 2023: Zero-shot Subjective Speech Quality Prediction for Multiple Domains

３．学会等名
ASRU 2023（国際学会）

４．発表年
2023年

１．発表者名
Aditya Ravuri, Erica Cooper, Junichi Yamagishi

２．発表標題
Uncertainty as a Predictor: Leveraging Self-Supervised Learning for Zero-Shot MOS Prediction

３．学会等名
IEEE ICASSP 2024 Workshop on Self-Supervision in Audio, Speech and Beyond（国際学会）

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

SASPEECH: Hebrew speech and transcripts for TTS: https://openslr.org/134/
Listening test data for "Range-Equalizing Bias": https://zenodo.org/records/10005796
Implementation of Partial Rank Similarity: https://github.com/nii-yamagishilab/partial_rank_similarity
VoiceMOS Challenge 2023 Homepage: https://voicemos-challenge-2023.github.io
The VoiceMOS Challenge 2022 Homepage: https://voicemos-challenge-2022.github.io
Open-source code for SSL-based MOS predictor: https://github.com/nii-yamagishilab/mos-finetune-ssl
TTS Pruning: https://people.csail.mit.edu/clai24/prune-tts/

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
研究分担者	Ｋｒｕｅｎｇｋｒａｉ　Ｃａｎａｓａｉ (Kruengkrai Canasai) (10895907)	国立情報学研究所・コンテンツ科学研究系・特任助教 (62615)	削除:2023年10月18日

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