Data Augmentation for Dysarthric Speech Recognition Based on Text-to-Speech Synthesis

Matsuzaka, Yuki
Takashima, Ryouichi
Sasaki, Chiho
Takiguchi, Tetsuya

(Citation)
2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech):399-400

(Issue Date)
2022-04-14

(Resource Type)
conference paper

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/0100481911

Data Augmentation for Dysarthric Speech Recognition Based on Text-to-Speech Synthesis

Yuki Matsuzaka
Faculty of Engineering
Kobe University
Kobe, Japan
matsuzaka@stu.kobe-u.ac.jp

Ryoichi Takashima
Graduate School of System Informatics
Kobe University
Kobe, Japan

Chiho Sasaki
Kumamoto Health Science University
Kumamoto, Japan

Tetsuya Takiguchi
Graduate School of System Informatics
Kobe University
Kobe, Japan

***Abstract*—In the field of automatic speech recognition (ASR) for people with dysarthria, it is problematic that not enough training speech data can be collected from people with dysarthria. To solve this problem, we propose a method of data augmentation using text-to-speech (TTS) synthesis. In the proposed data augmentation method, a deep neural network (DNN)-based TTS model is trained by utilizing speech data recorded from a speaker with dysarthria, and the trained TTS model is then used to generate the speaker's speech data for training the ASR model for the speaker. The results of a speech recognition experiment on a person having spinal muscular atrophy (SMA) showed that the speech recognition error rate was improved by using the proposed data augmentation.**

***Index Terms*—speech recognition, data augmentation, dysarthria, speaking disorder, speech synthesis**

I. Introduction

Dysarthria is the inability to pronounce words correctly due to abnormalities in one's speech organs or nerves, even though the person understands the language. There are various causes or diseases that can result in dysarthria, such as cleft lip and palate, cerebral palsy, amyotrophic lateral sclerosis, and so on. In this study, we will focus on dysarthria caused by spinal muscular atrophy (SMA).

Many people with SMA have difficulty moving their bodies, which makes their daily lives difficult. Speech assistants that utilize automatic speech recognition (ASR), such as AI speakers, have shown promise as support systems for people with physical disabilities. However, the speech of dysarthric people differs greatly from that of normal people, making it difficult for existing ASR models trained using normal speech to recognize them. Therefore, in order to recognize the speech of dysarthric people accurately, it is necessary to train the ASR model using their own speech. However, it is difficult to collect enough training data because recording the speech of dysarthric people places a heavy burden on them.

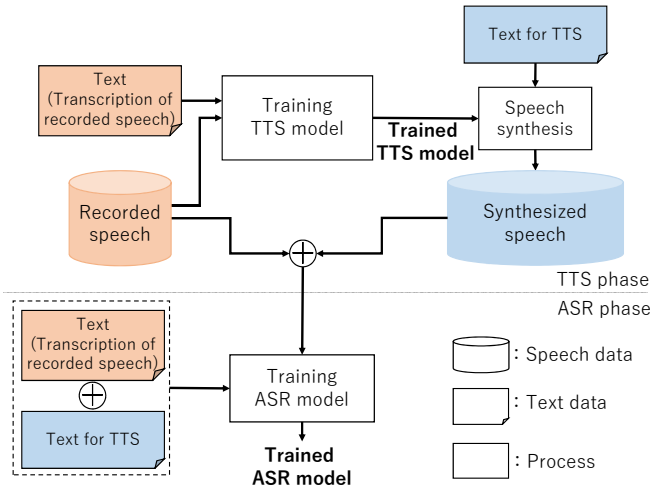

Fig. 1. Proposed procedure for training an ASR model using synthesized dysarthric speech.

One possible way to solve the problem of insufficient training data, model adaptation approaches have been studied, where the ASR model is trained using a large amount of normal speech data and then fine-tuned using a small amount of dysarthric speech [1]. Data augmentation approach, which artificially generates the training data, has also been studied [2]. In this study, we propose a data augmentation method by using speech generated by text-to-speech (TTS) synthesis. In the proposed method, a TTS model is trained using the recorded speech of a dysarthric speaker. And then, by using the trained TTS model, we generate the training data of the dysarthric speaker. The effectiveness of the proposed method is evaluated using speech recognition experiments on SMA patients.

II. Data Augmentation using TTS

The proposed method to train an ASR model is shown in Fig. 1. The proposed method consists of two phases: the first

is to generate training speech data using TTS (TTS phase), and the next is to train the ASR model using the generated speech and the recorded speech (ASR phase). First, the TTS model is trained using the recorded speech of the dysarthric person and its transcribed text. Next, the speech of the dysarthric person is synthesized by inputting the text into the trained TTS model. By adding the generated speech to the recorded speech, the training data can be augmented and the ASR model is trained.

In order to train the ASR model to be robust to differences in speech content, we synthesize speech for train the ASR model using text that has different content from the transcribed text of the recorded speech.

III. DNN-BASED PARAMETRIC TTS MODEL

In this study, we used the DNN-based parametric TTS model [3] as our TTS model. In recent years, end-to-end TTS models, such as Tacotron2 [4], have been studied. However, end-to-end TTS models require a large amount of training data. On the other hand, although the sound quality of a DNN-based parametric speech synthesis model is lower than end-to-end models, it can be trained with a relatively small amount of speech data, which is why this model was adopted for this study.

IV. EXPERIMENTS

A. Experimental conditions

We recorded the speech of one subject with SMA. The subject uttered each of the 216 words in the ATR dataset [5] five times ($216 \times 5 = 1{,}080$ utterances in total). Of these, 150 words $\times$ 5 utterances were used as training data, 30 words $\times$ 5 utterances as development data, and the remaining 36 words $\times$ 5 utterances as test data. The TTS model was trained using 180 words $\times$ 5 utterances from the training data and the development data.

750 words of text in the ATR dataset (different from the 216 words mentioned above) and the trained TTS model were used to generate dysarthric speech. The generated 750-word speech data was used as training data for the ASR model, in addition to the data from the 150-word recorded speech $\times$ 5 utterances. When training the ASR model, pre-training was also conducted using the speech of a normal person. We used basic5000 (5,000 sentences) from the JSUT corpus [6] as the speech of a healthy person.

The duration model of the TTS model consists of fully-connected layers. The input consists of 325-dimensional phoneme-level linguistic features, and the output is the duration of each phoneme. The acoustic model of the TTS model also consists of fully-connected layers, with the input being frame-level linguistic features in 329 dimensions. The output is a mel-cepstrum (40 dimensions), log F0 (1 dimension), aperiodic parameter (1 dimension) and their first and second derivatives, and the voice/unvoice flag (1 dimension), for a total of 127 acoustic features. WORLD [7], [8] was used to extract the acoustic features and convert the acoustic features to a speech waveform.

TABLE I
PER [%] OF EACH CONDITION FOR TRAINING ASR MODEL.

Training data set	PER[%]
recoded data (w/o pre-training)	60.03
recoded data (w/ pre-training)	49.58
recoded data + synthetic data	46.22

CTC [9] was used as the ASR model. The middle layer of the model contains a five-layer bidirectional GRU, with log mel-filterbank features (40 dimensions) as input and phonemes as output.

B. Experimental results

We compared the performance of the ASR model when only recorded speech was used as training data and when synthetic speech was also added as training data. In addition, we compared the performance with and without pre-training of healthy speech when using recorded speech only. Table I shows the phoneme error rate (PER) for each condition.

The results show that the PER was greatly improved when pre-training was carried out using the speech of healthy subjects. It was also shown that using not only recorded speech but also synthesized speech as training data further improved the PER.

V. CONCLUSION

In this study, we proposed a method of data augmentation using TTS to solve the problem of insufficient training data in speech recognition for dysarthric people. Based on the experimental results, we confirmed that the performance of speech recognition is improved by using synthesized speech as training data. However, the sound quality of the synthesized speech is inferior to that of recorded speech. If the quality of the synthesized speech can be improved in the future, the speech recognition performance is expected to be further improved. In the future, we will also conduct evaluation experiments using sentences.

REFERENCES

[1] R. Takashima, T. Takiguchi and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," Proc. ICASSP, pp. 6104-6108, 2020.

[2] K. Fujiwara et al., "Data augmentation based on frequency warping for recognition of cleft palate speech," Proc. APSIPA, 2021.

[3] H. Zen, A. Senior, M Schuster, "Statistical parametric speech synthesis using deep neural networks," Proc. ICASSP, pp.7962-7966, 2013.

[4] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," Proc. ICASSP, pp. 4779-4783, 2018.

[5] A. Kurematsu *et al.*, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, 9, pp.357-363, 1990.

[6] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint, 1711.00354, 2017.

[7] M. Morise, F. Yokomori, K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE transactions on information and systems, E99-D (7), 1877-1884, 2016.

[8] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," Speech Communication, 84, 57-65, 2016.

[9] A. Graves *et al.*, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," ICML, pp.369-376, 2006.