An Investigation of End-to-End Speech Recognition Using Model Adaptation for Dysarthric Speakers

Yuya Sawa
Graduate School of System Informatics
Kobe University
Kobe, Japan
yuyasawa@stu.kobe-u.ac.jp

Ryoichi Takashima
Graduate School of System Informatics
Kobe University
Kobe, Japan
rtakashima@port.kobe-u.ac.jp

Tetsuya Takiguchi
Graduate School of System Informatics
Kobe University
Kobe, Japan
takigu@kobe-u.ac.jp

***Abstract*—In this paper, we present an end-to-end automatic speech recognition (ASR) system for dysarthric speech. Because the speaking style of a person suffering from an articulation disorder is quite different from that of a physically unimpaired person, speech recognition systems for such persons need to be constructed in such a way that they specialize in meeting the needs of such dysarthric people. However, the amount of training data that can be collected from dysarthric people is limited because of their large burden. Therefore, it is a challenge to effectively train an ASR model for dysarthric people. In this paper, we introduce a model adaptation approach to train a more accurate model with limited training data, which adapts an ASR model trained by non-dysarthric speech samples for dysarthric speech recognition. From our experiments on an ASR task with two dysarthric subjects, the model adaptation approach with non-dysarthric speech showed better performance than training from scratch.**

***Index Terms*—Speech recognition, dysarthria, model adaptation**

I. Introduction

Recently, the accuracy of automatic speech recognition (ASR) systems has been improved with the development of deep learning technology. ASR systems are expected to be used for handicapped people because these systems have the merit of hands-free operation. However, it is difficult to use ASR systems for the people suffering from speech disorders. One of the causes of speech disorders is cerebral palsy, which results from damage to the central nervous system. Athetoid cerebral palsy, which is the focus of this paper, causes involuntary movements of the muscles when the person is in motion. These involuntary movements influence the movements of the face and tongue, and for this reason, the utterances of people with athetoid cerebral palsy are often unclear or unstable. Because athetoid symptoms also restrict their limb movements, they are unable to use alternative means of communication, such as sign language, writing, typing, and so on. Therefore, there is a great need for a reliable ASR system for those suffering from dysarthria.

II. Model adaptation to dysarthric speakers

In this paper, we attempt to construct an end-to-end ASR model for a dysarthric person. In the conventional DNN-HMM hybrid model for a dysarthric speaker, a problem arises because it is difficult to obtain the alignment information required to make labels during training because of the unstable speech associated with athetoid cerebral palsy. Therefore, we developed an end-to-end speech recognition model that does not require the alignment information. However, the amount of speech data obtained from dysarthric people is limited because their burden is large due to strain on their speech muscles. To construct the ASR model with a small amount of data, we use a model adaptation approach, where a source model is adapted to a target domain. In this work, the source model is pretrained on a large amount of non-dysarthric speech data in advance, and then it is fine-tuned on a small amount of dysarthric speech data. In this way, we can reuse an existing ASR model trained on a large set of training data of non-dysarthric speakers.

A number of studies of model adaptation for DNN-based acoustic models has been conducted. A previous work [1] demonstrated that a model adaptation approach with a DNN-HMM hybrid ASR model improves the accuracy of dysarthric speech recognition. In our study, we confirm that the model adaptation is also effective for training an end-to-end model.

III. Experiments

A. Experiment setup

We recorded the speech of two dysarthric subjects (DYS1 and DYS2) having athetoid cerebral palsy. Each dysarthric subject read 503 sentences included in the ATR Japanese speech database [2]. For comparison, we also evaluated the recognition accuracy of a non-dysarthric speaker (MHT) recorded in the ATR database. We conducted the speech recognition experiments for each speaker independently. For each speaker, we divided 503 sentences into 50 utterances for validation, 50 for evaluation, and the rest for training a speaker-dependent model. When we applied the model adaptation, we pretrained the ASR model by using about 240-hours of non-dysarthric speech recorded in the CSJ dataset [3], and then, we fine-tuned the model by using the above training data mentioned above.

For the ASR model, we trained a hybrid CTC/attention [4] model by using an ESPnet toolkit [5]. The input features consisted of 80-order mel-filterbank features and 3-order pitch-based features. The output label was consisted of 39 phonemes plus the unknown symbol <unk>, the start of sequence <sos>, and the end of sequence <eos>. The shared encoder

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

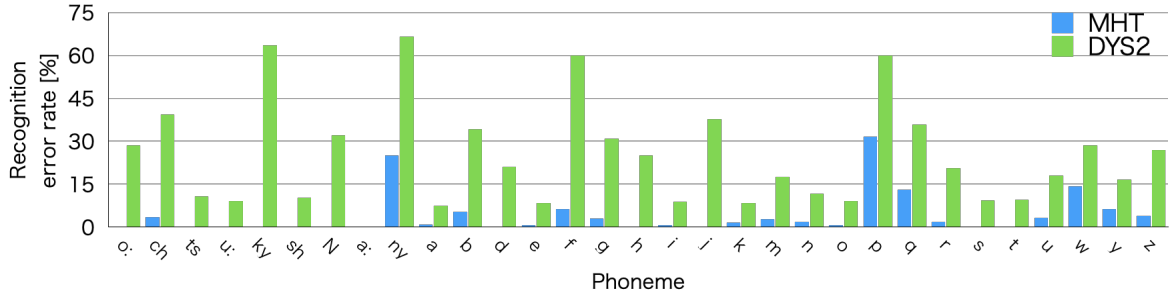

Fig. 1. Recognition error rates per phoneme

of the hybrid CTC/attention model consisted of four layers of pyramid bi-directional long short-term memory (pBLSTM) [6] having 320 cells for each layer. The decoder consisted of an unidirectional LSTM layer with 300 cells and a softmax output layer for phoneme entries. The location-aware attention [7] was used as the attention mechanism. The weight for the CTC-loss function was set to 0.5 during the multi-task training. We used the AdaDelta [8] method for optimizing networks. The weight for the output probability of CTC was also set to 0.5 during the decoding.

B. Results

Table I shows the phoneme error rates (PERs) of a model trained by CSJ dataset. Comparing the PERs of the dysarthric speakers (DYS1 and DYS2) and the non-dysarthric speaker (MHT), the PERs of the dysarthric speakers were significantly higher than those of the non-dysarthric speaker. These results indicate that the speaking styles of dysarthric people are quite different from those of non-dysarthric people.

TABLE I
PERs [%] OF A MODEL TRAINED ON NON-DYSARTHRIC SPEECH DATA

Speaker	PER
DYS1	47.6
DYS2	75.2
MHT	3.3

Table II shows the PERs of the model trained from scratch and that adapted from the non-dysarthria model. As shown in this table, a model adaptation approach decreased the PERs by 53.8% (DYS1) and 54.4% (DYS2) relatively. These results show that, although there is a significant difference between the speaking styles of dysarthric speakers and those of non-dysarthric speakers, the knowledge trained from non-dysarthric speeches is still helpful when training a model using limited dysarthric speech samples.

Fig. 1 shows the recognition error rates per phoneme that occurred five or more times in the dataset of DYS2 and MHT. For the dysarthric speaker, consonants like 'ky' and 'f' tended to be miss-recognized. These results indicate that consonants, which need to be strongly blown, are difficult to recognize. By analyzing such error tendencies for each dysarthric speaker, it is expected that the phonemes that are difficult to pronounce will become apparent.

TABLE II
PERs [%] OF TWO APROACHES TO TRAIN THE SPEAKER-DEPENDENT DYSARTHRIA MODEL

Speaker	Training from scratch	Adapted from non-dysarthria model
DYS1	28.8	13.3
DYS2	29.4	13.4

IV. Conclusion

In this paper, we investigated dysarthric speech recognition using a hybrid CTC/Attention model. The use of this model adaptation approach, which adapts an ASR model trained by non-dysarthric speech samples to dysarthric speech, decreased the error rates. The analysis of PERs also shows it is hard for dysarthric speakers to pronounce certain consonants. In future work, we will construct and evaluate a character-level or word-level end-to-end model, which does not need a pronunciation dictionary. We will also evaluate its performance with more dysarthric speakers.

References

[1] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acous- tic model adaptation for dysarthric speech recognition," in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6104–6108.
[2] "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol. 9, no. 4, pp. 357–363, 1990.
[3] K. Maekawa, "Corpus of spontaneous Japanese : its design and evaluation," Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR), pp. 7–12, 2003.
[4] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.
[5] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in Interspeech, 2018, pp. 2207–2211.
[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4960–4964.
[7] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 577–585.
[8] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," CoRR, vol. abs/1212.5701, 2012.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE