END-TO-END DYSARTHRIC SPEECH RECOGNITION USING MULTIPLE DATABASES

Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University, Japan

ABSTRACT

We present in this paper an end-to-end automatic speech recognition (ASR) system for a person with an articulation disorder resulting from athetoid cerebral palsy. In the case of a person with this type of articulation disorder, the speech style is quite different from that of a physically unimpaired person, and the amount of their speech data available to train the model is limited because their burden is large due to strain on the speech muscles. Therefore, the performance of ASR systems for people with an articulation disorder degrades significantly. In this paper, we propose an end-to-end ASR framework trained by not only the speech data of a Japanese person with an articulation disorder but also the speech data of a physically unimpaired Japanese person and a non-Japanese person with an articulation disorder to relieve the lack of training data of a target speaker. An end-to-end ASR model encapsulates an acoustic and language model jointly. In our proposed model, an acoustic model portion is shared between persons with dysarthria, and a language model portion is assigned to each language regardless of dysarthria. Experimental results show the merit of our proposed approach of using multiple databases for speech recognition.

Index Terms— Speech recognition, multilingual, assistive technology, end-to-end model, dysarthria

1. INTRODUCTION

In this study, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. In the case of a person with this type of articulation disorder, his/her movements are sometimes more unstable than usual [1]. That means their utterances (especially their consonants) are often unstable or unclear due to the athetoid symptoms. Athetoid symptoms also restrict the movement of their arms and legs. Most people suffering from athetoid cerebral palsy cannot communicate by sign language or writing, so there is great need for voice systems for them.

Automatic speech recognition (ASR) has been widely popularized with services such as the personal assistant on the smartphone. However, there has been very little benefit for orally-challenged people, such as those with speech impediments. Among the reasons for this are differences in speech style and limited speech data. In the case of people with an articulation disorder, because their speech style is quite different from that of physically unimpaired persons, a speaker-independent ASR system trained using physically unimpaired people is almost useless. Also, the amount of speech data obtained from a person with an articulation disorder that is available to train the model is limited because their burden is large due to strain on their speech muscles. Therefore, a "limited data"-driven approach is required.

Recent advances in deep learning for ASR have introduced remarkable progress [2, 3, 4], when there is a large amount of training data that can be used. However, collecting a large amount data from a person with an articulation disorder is significantly difficult due to their athetoid symptoms. Voice conversion (VC) is an approach that can be used to tackle this problem. Aihara *et al.* [5] have proposed a partial least square-based VC method using the phoneme-discriminative feature that converts dysarthric voice into non-dysarthric speech. Jiao *et al.* [6] have proposed a data-augmentation method based on convolutional generative adversarial network-based VC. As another approach, in this work, we adopt data augmentation using multilingual datasets. To be specific, we utilize not only the speech data of a target person with an articulation disorder but also the speech data of a physically unimpaired person who speaks the corresponding language and the speech data of non-Japanese people with articulation disorders.

Previous works on ASR have proposed end-to-end learning models combining acoustic and language models within a sequence to sequence framework [7, 8, 9]. Unlike traditional hidden Markov model-ASR systems, these models learn all the components of the ASR system jointly. Therefore, it is easy to develop ASR systems for new applications and configurations. In this work, we investigate an end-to-end ASR model based on the Listen, Attend and Spell (LAS) model [10] for the dysarthric speech. In multilingual speech recognition tasks, Toshniwal *et al.* [11] have jointly trained a single LAS model across data from 9 Indian languages, and shown improvement over monolingual models. This suggests that a LAS model has the capability to provide richer internal representation across several languages. In addition, we assume that a language model can be shared between speakers with or without dysarthria. From these considerations, in this

paper, we propose an end-to-end speech recognition framework that consists of a dysarthria-specific acoustic model portion, a healthy person-specific acoustic model portion, an English language model portion, a Japanese language model portion. A dysarthria-specific acoustic model portion is trained using speech data of a Japanese speaker and foreign speakers with an articulation disorder where the latter is included in publicly-available speech corpora. A Japanese language model portion is trained using Japanese speakers with and without an articulation disorder. In the phone prediction step, the speech uttered by a target speaker passes through these components to the output. Therefore, we can obtain a well-trained model for dysarthric speech recognition even if the amount of speech data of a target speaker is small. We show the effectiveness of our proposed approach through a phone recognition task.

2. RELATED WORKS

Previously, we have published some research on Japanese people with articulation disorders, which was collected in our own way. In [12], we proposed robust feature extraction based on principal component analysis, which has more stable utterance data, instead of discrete cosine transform. In [13], we used multiple acoustic frames as an acoustic dynamic feature to improve the recognition rate of a person with dysarthria, particularly for speech recognition using dynamic features only. In [14], we proposed a convolutional neural network-based feature extraction to deal with the small local fluctuations of the speech uttered by a person with an articulation disorder.

For clinical speech applications, some public databases are available [15, 16, 17]. Some researchers have worked on developing an ASR system using these databases [18, 19]. However, the speakers included in these databases are English speakers, and there is no publicly-available database for Japanese speakers. Thus, creating an ASR for Japanese speakers with articulation disorders is a very challenging task. The knowledge transferability across languages allows a multilingual ASR system to improve performance and allows the amount of training data for each language to decrease [20]. In this paper, we introduce the multilingual training of an end-to-end dysarthric ASR system.

3. LISTEN, ATTEND AND SPELL MODEL

A LAS model [10] consists of two modules: a listener and a speller, which are trained jointly. The goal of this model is to produce the probability of a grapheme sequence from the previous graphemes and a sequence of acoustic feature as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_s P(\boldsymbol{y}_s|\boldsymbol{x}, \boldsymbol{y}_{<s}), \tag{1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_t, ..., \boldsymbol{x}_T)$ and $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_s, ..., \boldsymbol{y}_S)$ are a sequence of acoustic features and graphemes, respectively. $\boldsymbol{x}_t$, $\boldsymbol{y}_s$, T and S are the input acoustic feature frame, the posterior distribution of the output grapheme, the number of the input acoustic features, and the output graphemes, respectively. A listener is an encoder-recurrent neural network (RNN) that transforms an input sequence $\boldsymbol{x}$ of acoustic features into a high level representation $\boldsymbol{h} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_U)$, where $\boldsymbol{h}_u$ and $U \leq T$ are the encoder output feature, and the number of the encoder output sequence, respectively. A speller is a decoder RNN that consumes $\boldsymbol{h}$ and produces a probability distribution over graphemes sequence $\boldsymbol{y}$.

The listener is a stacked pyramid bidirectional long short term memory (pBLSTM) on top of the bottom BLSTM layer. The pyramid structure reduces the computational complexity and the convergence time, and allows the speller to extract the relevant information from a smaller number of time steps. The listener is considered the acoustic model in an ASR system. The listener operation is written as follows:

$$\boldsymbol{h} = \mathrm{Listen}(\boldsymbol{x}; \theta_{\mathrm{Lis}}), \tag{2}$$

where θ_{Lis} denotes the parameters of a listener.

The speller is an attention-based LSTM transducer which is a stacked unidirectional RNN. At every time step, the speller produces a probability distribution over the next graphemes conditioned on all the graphemes seen previously. The attention mechanism allows the speller to generate the next output over graphemes encapsulating the information in the acoustic signal. The speller is considered the language model in an ASR system. The speller operation is written as follows:

$$P(\boldsymbol{y}|\boldsymbol{x}) = P(\boldsymbol{y}|\boldsymbol{h}; \theta_{Spl}) = \prod_s P(\boldsymbol{y}_s|\boldsymbol{h}, \boldsymbol{y}_{<s}; \theta_{Spl}) \tag{3}$$

$$= \mathrm{Spell}(\boldsymbol{h}; \theta_{\mathrm{Spl}}), \tag{4}$$

where θ_{Spl} denotes the parameters of a speller.

The model is trained to optimize the discriminative loss as follows:

$$\mathcal{L}_{LAS}(\mathcal{D}, \theta_{Lis}, \theta_{Spl}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}}[-\log(P(\boldsymbol{y}|\boldsymbol{x}))], \tag{5}$$

where $\mathcal{D}$ denotes the joint distribution over input sequence $\boldsymbol{x}$ and label sequence $\boldsymbol{y}$.

4. PROPOSED METHOD

4.1. Architecture

Fig. 1 shows the overview of our proposed method. Our proposed ASR system is based on the LAS model, which consists of two listeners and two spellers.

First, we configure a dysarthria-specific listener "D-Listen" and a controlled speaker (a physically unimpaired

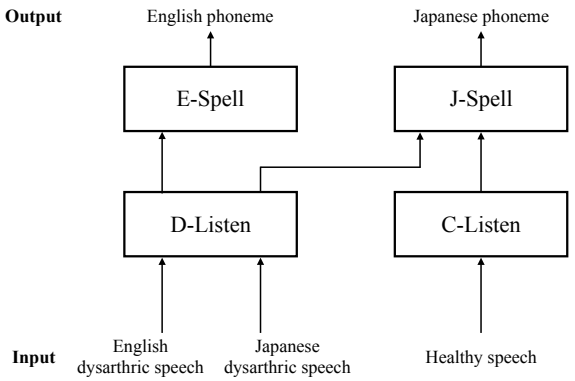

Fig. 1. Overview of our proposed method. Japanese dysarthric speech is uttered by one target speaker in this paper.

person)-specific listener "C-Listen". D-Listen is shared between persons with articulation disorders regardless of their languages. Even though the amount of target dysarthric speaker data is small, we are able to obtain the well-trained listener module specialized in dysarthria using a large amount of speech data of foreign people with articulation disorders. We expect that this multilingual mechanism will help the listener module to capture a better high-level representation. C-Listen is fed the input features from only a physically unimpaired person.

Next, we configure an English speller "E-Spell" and a Japanese speller "J-Spell". E-Spell and J-Spell produce the phone sequence of the English word and the Japanese word, respectively. J-Spell is fed the input features from D-Listen and C-Listen to output a probability distribution over phone sequence. Because the language model can be shared between speakers with or without dysarthria, we are able to obtain a well-trained speller using a large amount of speech data of a physically unimpaired person.

4.2. Training

In this section, we describe the loss function used to optimize our proposed model. For a "Japanese person with articulation disorder" dataset, let $\mathcal{D}_{JD}$ be the joint distribution over the input sequence and the corresponding label sequence. $\mathcal{D}_{ED}$ and $\mathcal{D}_{JC}$ are analogously defined for the "English persons with an articulation disorder" dataset and a controlled Japanese person dataset, respectively. Our proposed model is optimized while adjusting the parameters to minimize the loss function as follows:

$$\mathcal{L}_{LAS}(\mathcal{D}_{ED}, \theta_{D\text{-}Lis}, \theta_{E\text{-}Spl}) + \mathcal{L}_{LAS}(\mathcal{D}_{JD}, \theta_{D\text{-}Lis}, \theta_{J\text{-}Spl}) + \mathcal{L}_{LAS}(\mathcal{D}_{JC}, \theta_{C\text{-}Lis}, \theta_{J\text{-}Spl}), \quad (6)$$

where $\theta_{D\text{-}Lis}$, $\theta_{C\text{-}Lis}$, $\theta_{E\text{-}Spl}$ and $\theta_{J\text{-}Spl}$ denote the parameters of D-Listen, C-Listen, E-Spell and J-Spell, respectively. In this model, all components are simultaneously learned.

5. EXPERIMENTAL RESULTS

5.1. Experimental setup

Our proposed approach was evaluated on a phone recognition task for two Japanese males with an articulation disorder. We recorded 216 words included in the ATR Japanese speech database [21], repeating each word five times for "Dysarthric speaker1". "Dysarthric speaker2" was not able to utter six words out of the 216 words due to his athetoid symptoms, so we recorded only the 210 words that he was able to utter. In our experiments, the first utterances of each word were used for evaluation, and the other utterances (e.g., 864 words for Dysarthric speaker1) were used to train models). The Japanese physically unimpaired person is one male speaker whose speech is stored in the ATR Japanese speech database. We used 5,240 words for training and another 216 words for evaluation, which were the same words as the dysarthric dataset. For the "foreign people with articulation disorders" speech dataset, we used the TORGO database [17], which includes three females and four males. We selected 2,726 words across all speakers from this database. We used 95% of each dataset as the training set (e.g., 821 words for Dysarthric speaker1), and the remaining were used as the validation set. We used 39-dimensional mel-frequency cepstrum coefficient (MFCC) features (13-order MFCCs, their delta, and acceleration) as the input feature computed every 10ms over a 25ms window.

For the baseline system, we trained two models based on the conventional LAS model. The first model was trained using only the data of a physically unimpaired Japanese person, and the second model was trained using both a physically unimpaired Japanese person's data and the data of a Japanese person with an articulation disorder. For the listener configuration, we used 2 layers of 512 pBLSTM nodes (256 nodes per direction) on top of a BLSTM that operates on the input. For the speller configuration, we used a one-layer LSTM with 512 nodes. In this work, we used the phone sequence as the output sequence. The output dimensionality is 59 and 56 with the start-of-sequence and end-of-sequence token for English and Japanese spellers, respectively. The network is optimized using an Adam optimizer [22] with a batch size of 64, label-smoothing, and with early-stopping using validation set. The learning rate is set to 1e-4.

5.2. Results and discussion

First, we confirmed the performance of a model trained on the speech data of a physically unimpaired person only. Table 1 shows the results of the character error rate (CER). In this table, Top-1 indicates the top-1 error rate, and Top-3 indicates the top-3 error rate with beam width 3. As expected, the CER value for persons with articulation disorders is quite

Table 1. CER (%) of a model trained on speech data of a physically unimpaired person only.

Test speaker	Top-1	Top-3
Controlled speaker	12.2	9.2
Dysarthric speaker1	72.2	69.2
Dysarthric speaker2	76.3	75.3

Table 2. CER (%) of a model trained on joint speech data of a physically unimpaired person and a person with an articulation disorder.

Test speaker	Top-1	Top-3
Controlled speaker	11.2	7.2
Dysarthric speaker1	27.2	22.2
Controlled speaker	11.2	9.2
Dysarthric speaker2	32.1	27.0

a bit higher than that of a physically unimpaired person. In fact, it is shown that the model trained using the speech data of a physically unimpaired person is almost useless.

Next, we evaluated the performance of models trained jointly on speech data of both a physically unimpaired person and a person with an articulation disorder. As shown in Table 2, this joint training method achieved lower CER values than those in Table 1. This is because the speech data of an evaluation speaker was used to train the model.

Table 3 shows results of our proposed method. Our proposed method achieved 14.7% and 6.5% relative improvement over the joint training method for dysarthric speaker1 and dysarthric speaker2 respectively, using the same amount of data to train the Japanese speller. This means that the dysarthria-specific listener learned better representation using the speech of English and Japanese speakers with dysarthria. As shown in previous work [11], multilingual training boosts the performance in a LAS framework even though a speaker is the person with an articulation disorder.

Table 4 shows some of the examples where our proposed model fails to predict the correct word from the speech spoken by dysarthric speaker1. From this table, we can see that the consonant tends to be lacking. The model correctly predicted the vowel, but not the consonant. Due to the athetoid symptoms, a person with an articulation disorder has difficulty uttering consonants, so their speech is difficult to understand for listeners in the real world. The model learned this characteristic properly. These output sequences show the tendency of phone errors, which enable us to apply these results to error correction [23] in order to recognize the correct word.

Table 3. CER (%) of our proposed method.

Target speaker	Top-1	Top-3
Dysarthric speaker1	23.2	19.2
Dysarthric speaker2	30.0	24.9

Table 4. Example of ASR results for dysarthric speaker1 predicted from our proposed method. "pau" indicates the pause.

Ground Truth	pau u r a y a- m a sh ii pau
Predicted	pau u m a a a a s ii pau
Ground Truth	pau d a ky ou pau
Predicted	pau d a k ou pau

Finally, we measured the performance of our framework as a function of the number of training words using the speech data of dysarthric speaker1. As shown in Fig 2, we observed that the CER value decreases as the number of training words increases. This is because the speech data of a physically unimpaired person dominates the capability of a Japanese speller. In our proposed method, the Japanese speller is fed input features from both the dysarthria-specific listener and the controlled person-specific listener. Therefore, the Japanese speller must deal with inputs features from two different distributions. This may be another reason for the degradation of performance.

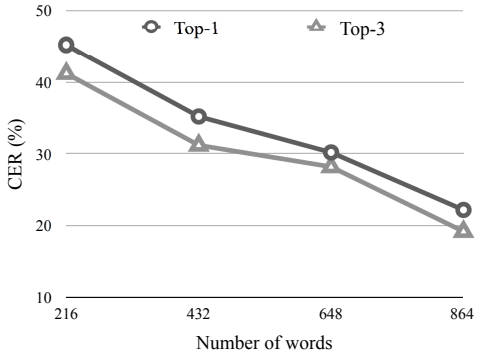

Fig. 2. The correlation between error rates and the number of training words.

6. CONCLUSION

In this paper, we investigated an end-to-end ASR system for a Japanese person with an articulation disorder based on a LAS model. Due to their athetoid symptoms, the amount of speech data obtained from such speakers is very small. To tackle this problem, we used the additional speech data from a physically unimpaired person and foreign people with articulation disorders. In our phone recognition experiments, we showed the effectiveness of our proposed approach. In future work, we will research how to relieve the domain confusion in the Japanese speller.

7. REFERENCES

[1] Frederic L. Darley, A.E. Aronson, and J.R. Brown, *Motor Speech Disorders*, Audio seminars in speech pathology. Saunders, 1975.

[2] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS workshop on deep learning for speech recognition and related applications*, 2009.

[3] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *ICASSP*, 2011, pp. 4688–4691.

[4] Tara N. Sainath, Abdel rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for LVCSR," in *ICASSP*, 2013, pp. 8614–8618.

[5] Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, "Phoneme-discriminative features for dysarthric speech conversion," in *INTERSPEECH*, 2017, pp. 3374–3378.

[6] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *ICASSP*, 2018, pp. 6009–6013.

[7] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets," in *International Conference on Machine Learning*, 2006.

[8] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *CoRR*, vol. abs/1412.1602, 2014.

[9] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," in *NIPS*, 2014.

[10] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[11] Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro J. Moreno, Eugene Weinstein, and Kanishka Rao, "Multilingual speech recognition with a single end-to-end model," in *ICASSP*, 2018, pp. 4904–4908.

[12] Hironori Matsumasa, Tetsuya Takiguchi, Yasuo Ariki, Ichao Li, and Toshitaka Nakabayashi, "Integration of metamodel and acoustic model for speech recognition," in *INTERSPEECH*, 2008, pp. 2234–2237.

[13] Chikoto Miyamoto, Yuto Komai, Tetsuya Takiguchi, Yasuo Ariki, and Ichao Li, "Multimodal speech recognition of a person with articulation disorders using aam and maf," in *MMSP*, 2010, pp. 517–520.

[14] Yuki Takashima, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition," in *EUSIPCO*, 2015, pp. 1411–1415.

[15] Xavier Menéndez-Pidal, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzio, and H. Timothy Bunnell, "The nemours database of dysarthric speech," in *International Conference on Spoken Language Processing*, 1996.

[16] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008, pp. 1741–1744.

[17] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[18] Cristina Espaa-Bonet and Jos A. R. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *IberSPEECH*, 2016, pp. 97–107.

[19] Neethu Mariam Joy, S. Umesh, and Basil Abraham, "On improving acoustic models for TORGO dysarthric speech database," in *INTERSPEECH*, 2017, pp. 2695–2699.

[20] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[21] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane, "Automatic speech recognition errors detection and correction: A review," in *ICNLSP*, 2015, vol. 128, pp. 32–37.