IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 21, 2019, accepted October 30, 2019, date of publication November 6, 2019, date of current version November 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2951856

Knowledge Transferability Between the Speech Data of Persons With Dysarthria Speaking Different Languages for Dysarthric Speech Recognition

YUKI TAKASHIMA, RYOICHI TAKASHIMA, (Member, IEEE), TETSUYA TAKIGUCHI, (Member, IEEE), AND YASUO ARIKI, (Member, IEEE)
Graduate School of System Informatics, Kobe University, Kobe 6578501, Japan

Corresponding author: Yuki Takashima (takashima@kobe-u.ac.jp)

This work was supported in part by JSPS KAKENHI under Grant JP17J04380.

ABSTRACT In this paper, we present an end-to-end speech recognition system for Japanese persons with articulation disorders resulting from athetoid cerebral palsy. Because their utterance is often unstable or unclear, speech recognition systems struggle to recognize their speech. Recent deep learning-based approaches have exhibited promising performance. However, these approaches require a large amount of training data, and it is difficult to collect sufficient data from such dysarthric people. This paper proposes a transfer learning method that transfers two types of knowledge corresponding to the different datasets: the language-dependent (phonetic and linguistic) characteristic of unimpaired speech and the language-independent characteristic of dysarthric speech. The former is obtained from Japanese non-dysarthric speech data, and the latter is obtained from non-Japanese dysarthric speech data. In the proposed method, we pre-train a model using Japanese non-dysarthric speech and non-Japanese dysarthric speech, and thereafter, we fine-tune the model using the target Japanese dysarthric speech. To handle the speech data of the two different languages in one model, we employ language-specific decoder modules. Experimental results indicate that our proposed approach can significantly improve speech recognition performance compared with other approaches that do not use additional speech data.

INDEX TERMS Assistive technology, deep learning, dysarthria, end-to-end model, knowledge transfer, multilingual, speech processing, speech recognition.

I. INTRODUCTION

In this study, we focused on the problem of speech recognition for persons with articulation disorders caused by the athetoid type of cerebral palsy. Cerebral palsy is usually caused by damage to the central nervous system, and consequently, it causes movement disorders. Movements of a person with this type of the articulation disorder can sometimes be more unstable than usual [1]. For this reason, their utterance (especially consonants) is often unstable or unclear owing to athetoid symptoms. These symptoms also restrict the movement of arms and legs. Most persons suffering from athetoid cerebral palsy are unable to communicate using sign language or writing and therefore have a critical need for voice-driven assistive systems [2]. Fig. 1 depicts the spectrograms of a physically unimpaired person and of a person with an articulation disorder for the Japanese word ''uchiawase'' (''meeting'' in English). As shown in this figure, the high-frequency spectral power of the person with an articulation disorder is weaker in comparison to that of the physically unimpaired person, demonstrating the unclear speech of persons with an articulation disorder, which makes their speech difficult to understand.

Automatic speech recognition (ASR) has been widely spread within services such as personal assistants on smartphones. In addition, remarkable progress has been made with respect to recent developments in deep learning for ASR [3]–[5] in fields with availability to a large amount of training data. However, there has been no significant beneficial use of ASR achieved for persons with speech disorders as a result of differences in various speech styles and the limited amount of the training speech data available. In the case of persons

The associate editor coordinating the review of this manuscript and approving it for publication was Gerard-Andre Capolino.

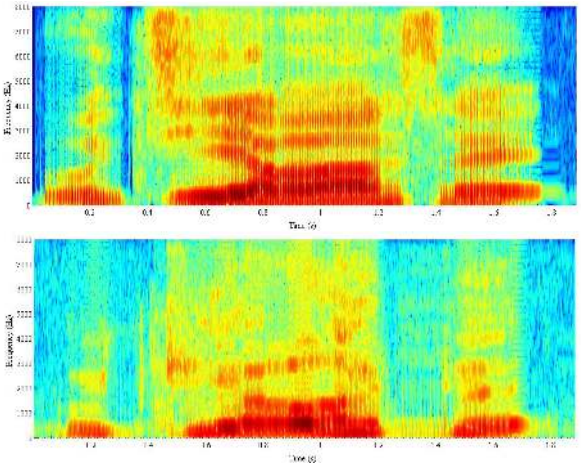

FIGURE 1. Example of spectrogram uttered for /u ch i a w a s e/ of a physically unimpaired person (top) and of a person with an articulation disorder (bottom). These spectrograms are stretched using dynamic time warping to be more easily observed.

with articulation disorders, owing to their speech style differing significantly from that of physically unimpaired persons, a speaker-independent ASR system trained using the data of physically unimpaired persons is almost useless. However, it is difficult to collect sufficient speech data from persons with articulation disorders to train the model. Therefore, there is a need for an approach specifically tailored to overcome the low data availability of impaired speech.

To solve the problem of limited data availability, we employ transfer learning [6], which seeks to apply the knowledge learned in one or more domains or tasks to another domain or task. To be specific, we use three different sources of data: speech data from a target speaker with an articulation disorder, speech data from physically unimpaired persons in the same target language, and speech data of persons with articulation disorders in other languages. In our previous work [7], a data-augmentation method based on an end-to-end ASR model was proposed. The model comprises a dysarthria-specific encoder, a physically unimpaired person-specific encoder, an English decoder, and a Japanese decoder. This method has exhibited the ability to efficiently integrate the knowledge gathered from the speech data of physically unimpaired Japanese-speaking persons, as well as Japanese-speaking and English-speaking persons with articulation disorders. However, it was difficult to train the model because it optimized all modules simultaneously.

In this paper, we investigate the knowledge transfer of the language-dependent characteristic corresponding to the speech of physically unimpaired persons as well as the language-independent characteristic of the dysarthric speech. We refer to a physically unimpaired Japanese-speaking person, an English-speaking person with an articulation disorder, and a Japanese-speaking person with an articulation disorder as JU, ED, and JD, respectively. As is widely known, a large amount of speech data of physically unimpaired persons is publicly available. We assume that knowledge of Japanese language-specific characteristics from JU speech data can be transferred and applied to JD speech. Moreover, several non-Japanese speech databases of persons with articulation disorders have been published [8]–[12]. The speech impairment and abnormalities caused by dysarthria, such as imprecise consonants and distorted vowels, are common among speakers of different languages. Consequently, we assume that the characteristics of dysarthric speech are independent of language. According to these assumptions, we propose a method to transfer the knowledge learned from these rich speech data sources to the problem of Japanese dysarthric speech recognition, for which training data is insufficient. Utilization of the speech data from physically unimpaired persons who speak the same language as the target speaker has been considered in some literatures [13]–[15]. However, these studies only considered the linguistic characteristics within the language. Following the assumption that the characteristics of dysarthric speech are language-independent, our work utilizes data from persons with articulation disorders who speak different languages in addition to using data from persons who speak the same language as the target speaker. Compared to the random initialization without using additional data, our proposed approach provides a considerable reduction in phoneme error rate.

The rest of this paper is organized as follows. In Section II, we introduce discussion on related works. In Section III, we provide a summary of the Listen, Attend and Spell (LAS) model concepts. In Section IV, our proposed method is explained in detail. In Section V, the experimental data are evaluated, and the final section is devoted to the conclusions of this research.

II. RELATED WORKS

Previously, we have published several research works on speech recognition for JDs using speech data which was collected using our own methods. Instead of using discrete cosine transform, we proposed robust feature extraction based on principal component analysis [16], which provides more stable utterance data. In [17], multiple acoustic frames as an acoustic dynamic feature to improve the speech recognition rate of a person with dysarthria, particularly for speech recognition using dynamic features only. We proposed feature extraction based on a convolutional neural network to process small local fluctuations of speech uttered by a person with an articulation disorder [18]. These methods handle feature extraction and do not account for the limited amount of speech data specific to persons with articulation disorders.

Voice conversion (VC) is an approach that can be used to mitigate the problem of limited data availability. Aihara *et al.* [19] have proposed a VC method based on partial least square using a phoneme-discriminative feature that converts a dysarthric voice into non-dysarthric speech. Jiao *et al.* [13] have proposed a data-augmentation method based on convolutional generative adversarial network (GAN) [20]. In [21], speech synthesis-based data augmentation was proposed. The dysarthric-like speech was

generated using temporal and speed modifications applied to speech of physically unimpaired persons, and this generated speech dataset was used thereafter to train an ASR based on a deep neural network. Vachhani *et al.* [14] have proposed a feature enhancement method based on an autoencoder. This method implies training the autoencoder with the use of a speech signal obtained from a physically unimpaired control speaker, after which this speech data was used to convert dysarthric speech into an improved feature representation.

Several public databases are available for clinical speech applications [8]–[10]. Several researchers have worked on developing an ASR system using these databases [22]–[24]. However, speakers included in these databases are English speakers, and there is no publicly available database with speech data obtained from Japanese speakers. Therefore, establishing an ASR for Japanese speakers with articulation disorders is very challenging.

Knowledge transferability across different languages allows performance improvements for a multilingual ASR system and is especially efficient for low-resource languages. Considering multilingual speech recognition tasks, Toshniwal *et al.* [25] have jointly trained a single ASR model across a dataset composed of data corresponding to nine Indian languages; this approach has shown improvements over monolingual models. This suggests that an ASR model can provide richer internal representation across several languages. To train a model, we proposed an end-to-end speech recognition framework [7] that uses the speech data of JUs, EDs, and a JD. This approach suggested a means to transfer knowledge across a language for speech affected by dysarthria; however, the difficulty related to model training remained. In contrast, the new method proposed in this paper allows training a model easily on both the dysarthria-specific and Japanese-specific acoustic characteristics.

III. LISTEN, ATTEND AND SPELL MODEL

In this work, we employ an end-to-end ASR model for dysarthric speech recognition. Previous works on ASR have proposed various end-to-end learning models combining acoustic and language models within a sequence-to-sequence framework [26], [27]. Unlike ASR systems based on traditional hidden Markov models, these models learn all components of the ASR system jointly. Therefore, it enables the development of ASR systems for new applications and configurations. In this work, we investigate an end-to-end ASR model based on the LAS model [28] for dysarthric speech.

The LAS model [28] consists of a listener module and a speller module which are trained jointly. The goal of this model is to generate the probability of a grapheme sequence based on information from the previous graphemes and a sequence of acoustic features. The model can be defined as follows:

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

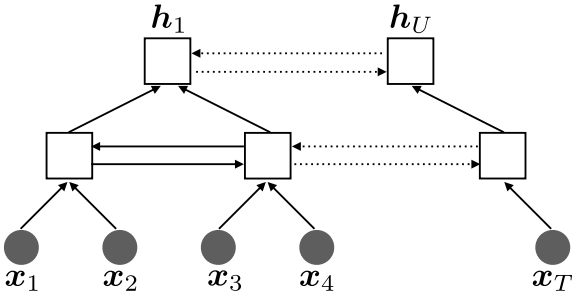

FIGURE 2. Network structure of the listener. Each layer has a pyramid structure that takes every two consecutive frames of the output from the previous layer as input.

where $\boldsymbol{x} = (x_1, \ldots, x_t, \ldots, x_T)$ and $\boldsymbol{y} = (y_1, \ldots, y_s, \ldots, y_S)$ denote sequences of acoustic features and graphemes, respectively. Here, x_t, y_s, T, and S denote 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 $\boldsymbol{h} = (h_1, \ldots, h_u, \ldots, h_U)$, where 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 grapheme sequence $\boldsymbol{y}$.

The listener is organized as a stacked pyramid bidirectional long short-term memory (pBLSTM) as shown in Fig. 2. The pyramid structure allows for reducing the computational complexity and the convergence time and provides the speller with the ability to extract the relevant information within a smaller number of time steps. The listener is considered as the acoustic model in an ASR system. Its operation is defined as follows:

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

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

The speller is an attention-based long short-term memory (LSTM) transducer, which is organized as a stacked unidirectional RNN. At each time step, the speller produces a probability distribution over the subsequent graphemes conditioned on all the graphemes obtained previously. The attention mechanism allows the speller to generate the next output over graphemes encapsulating information within the acoustic signal. The speller is considered as a language model in an ASR system. The speller operation is written as follows:

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

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

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

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

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

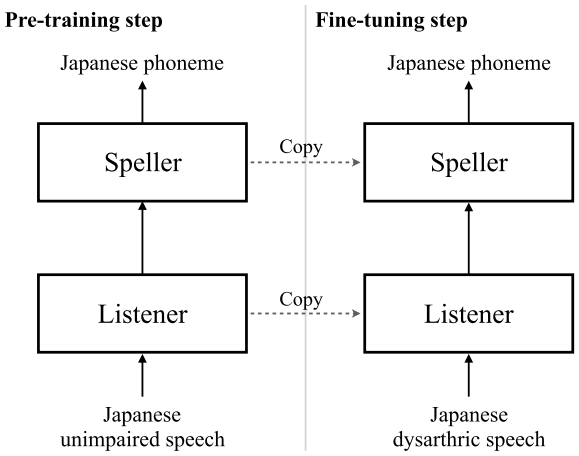

FIGURE 3. Training scheme of a single LAS model with pre-training using the speech of JUs.

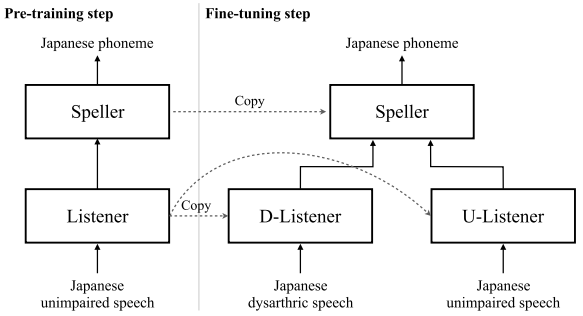

FIGURE 4. Training scheme of a two-encoder LAS model with pre-training using the speech of JUs. The speech data of physically unimpaired persons are used to explicitly separate dysarthria characteristics during the fine-tuning step.

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

IV. PROPOSED METHOD

In this section, we explain two knowledge transfer methods, using speech data obtained from JUs and EDs, which could be used for speech recognition for a JD. Our proposed ASR system is based on the LAS model. Considering the JD dataset, let $\mathcal{D}_{JD}$ be the joint distribution over the input sequence and the corresponding label sequence. $\mathcal{D}_{ED}$ and $\mathcal{D}_{JU}$ are analogously defined for the EDs and JUs datasets, respectively. In this work, the speller produces a phoneme of the corresponding language.

A. TRANSFER LEARNING USING THE SPEECH DATA OBTAINED FROM THE PHYSICALLY UNIMPAIRED PERSONS

In this section, we explain the intra-language knowledge transfer from the speech data obtained from physically unimpaired persons to a speech representation for a person with an articulation disorder. Fig. 3 depicts the overview of this pre-training scheme. First, we pre-train a single LAS model using the speech data obtained from physically unimpaired persons while adjusting the parameters to minimize the loss function as follows:

$$\hat{\theta}_{Lis}, \hat{\theta}_{Spl} = \underset{\theta_{Lis}, \theta_{Spl}}{arg\,min}\ \mathcal{L}(\mathcal{D}_{JU}, \theta_{Lis}, \theta_{Spl}), \tag{6}$$

where $\hat{\theta}_{Lis}$ and $\hat{\theta}_{Spl}$ are the optimized parameters of the listener and speller, respectively. It is assumed that these pre-trained parameters have appropriate representation for producing Japanese phonemes. Then, obtained parameters $\hat{\theta}_{Lis}$ and $\hat{\theta}_{Spl}$ are fine-tuned using the speech data taken from a person with an articulation disorder, optimized as follows:

$$\underset{\hat{\theta}_{Lis}, \hat{\theta}_{Spl}}{arg\,min}\ \mathcal{L}(\mathcal{D}_{JD}, \hat{\theta}_{Lis}, \hat{\theta}_{Spl}). \tag{7}$$

At the prediction step, we use the well-trained listener and speller parameters to produce Japanese phonemes.

Inspired by our previous work [7], we constructed another model that consists of two listeners and one speller for fine-tuning as shown in Fig. 4. One is a dysarthria-specific listener called "D-Listener", and the other is a physically unimpaired speaker-specific listener called "U-Listener". This model is also trained for optimizing parameters as follows:

$$\mathcal{L}(\mathcal{D}_{JD}, \theta_{D\text{-}Lis}, \hat{\theta}_{Spl}) + \mathcal{L}(\mathcal{D}_{JU}, \theta_{U\text{-}Lis}, \hat{\theta}_{Spl}), \tag{8}$$

where $\theta_{D\text{-}Lis}$ and $\theta_{U\text{-}Lis}$ denote parameters of D-Listener and U-Listener and are initialized to $\hat{\theta}_{Lis}$ before training. In this fine-tuning, the speller is initialized to $\hat{\theta}_{Spl}$ and is shared between a person with an articulation disorder and the physically unimpaired persons. The acoustic characteristic of dysarthric speech considerably differs from that of the speech of a physically unimpaired person owing to the athetoid symptoms. Therefore, we use the dysarthria-specific listener for the speech data of JD.

B. TRANSFER LEARNING USING MULTILINGUAL SPEECH DATA OBTAINED FROM A PERSON WITH DYSARTHRIA

In this section, we explain the knowledge transfer from the speech data obtained from both JUs and EDs to a speech representation for JD. First, we configure a shared listener called "S-Listener", a Japanese speller called "J-Speller", and an English speller called "E-Speller", as shown in Fig. 5. The S-Listener is shared between EDs and JUs. This mechanism is similar to multitask learning [29]. Multitask learning simultaneously executes multiple tasks for one input data. However, our proposed approach differs in that each input data from different domains is processed to estimate the phonemes of only the corresponding language. E-Speller and J-Speller produce phoneme sequences in English and Japanese, respectively. This model is optimized by adjusting the parameters to minimize the loss function as follows:

$$\mathcal{L}(\mathcal{D}_{ED}, \theta_{S\text{-}Lis}, \theta_{E\text{-}Spl}) + \mathcal{L}(\mathcal{D}_{JU}, \theta_{S\text{-}Lis}, \theta_{J\text{-}Spl}), \tag{9}$$

where $\theta_{S\text{-}Lis}$, $\theta_{E\text{-}Spl}$, and $\theta_{J\text{-}Spl}$ denote parameters of S-Listener, E-Speller, and J-Speller, respectively.

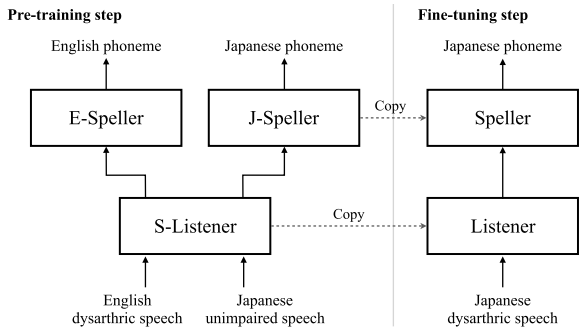

FIGURE 5. Training scheme of a single LAS model with pre-training using the speech of JUs and the speech of EDs. In the pre-training step, the speller is switched according to the input language.

TABLE 1. Dataset statistics gathered for JDs.

Speaker	# words	# phonemes	# repetitions	# utterances
JM1	204	1,639	3	612
JM2	210	1,687	5	1,050
JM3	216	1,731	5	1,080
JM4	215	1,721	3	645
JM5	213	1,705	3	639

All components are learned jointly. We expect that this cross-lingual mechanism will help the listener module capture a better high-level representation containing both the dysarthria-specific characteristic and the expression capability of Japanese. Then, the obtained parameters $\theta_{S\text{-}Lis}$ and $\theta_{J\text{-}Spl}$ are jointly fine-tuned using speech data of JD with an articulation disorder, as defined in (7).

V. EXPERIMENTAL EVALUATION

In this work, we conducted experiments on the speaker-dependent system for each target speaker with dysarthria.

A. CONDITIONS

Our proposed approach was evaluated on a phoneme recognition task suggested to five Japanese-speaking males with articulation disorders. We repeatedly recorded 216 words included in the ATR Japanese speech database [30] for each speaker as shown in Table 1. The number of repetitions differed for each speaker owing to the athetoid symptoms. In our experiments, the first utterances of each word were used for evaluation, and the other utterances (e.g., 864 words for JM3) were used to train models. In the experiment, JUs were five male and five female speakers, whose speech is stored in the ATR Japanese speech database. We used the same 216 words (1,731 phonemes) for each speaker as in the dysarthric dataset. Considering the speech dataset corresponding to EDs, we used the TORGO database [10], which includes three female and five male persons. This database has missing data owing to a clipping error; therefore, we selected usable word speech as shown in Table 2. When we pre-trained (in all proposed methods) or fine-tuned (only in the two-encoder LAS) modules using JU speech and ED speech, we used all speakers' speech. When we fine-tuned modules

TABLE 2. Dataset statistics gathered from the TORGO database for EDs.

Speaker	# phonemes	# utterances
F01	1,014	188
F03	5,290	777
F04	3,531	489
M01	4,217	556
M02	4,086	568
M03	3,860	594
M04	3,519	499
M05	3,346	443

using Japanese dysarthric speech for a speaker-dependent model, we used only the target speaker's speech. In this manner, for each Japanese dysarthric subject, we trained the speaker-dependent model and evaluated the model independently. We used 39-dimensional mel-frequency cepstral coefficient (MFCC) features (13-order MFCCs, their delta, and acceleration) as the input features, computed every 10 ms over a 25 ms window.

Considering the listener configuration, we used 2 layers of 512 pBLSTM nodes (256 nodes per direction). For the speller configuration, we used a one-layer LSTM with 512 nodes. In this work, we used the phoneme sequence as the output sequence. The numbers of phonemes for English and Japanese were 57 and 54, respectively. The network was optimized using an Adam optimizer [31] with label smoothing [32]. We constructed one batch with sub-batches of 64 samples for each domain. The number of epochs was 500, and the learning rate was set to 1e-4.

For the baseline system, we trained two models based on the conventional single LAS model. The first model was trained using only speech data of JUs ('rand init'), and the second model was trained using the data of both JUs and JD ('multi').

B. RESULTS

Table 3 lists the phoneme error rates (PERs) corresponding to each method. In this table, a lower PER indicates a better result. Here, 'trans. 1' and 'trans. 2' are the transfer learning methods using the speech data of only JUs, and 'trans. 3' is the transfer learning method using the speech data of both JUs and EDs.

Multi-condition learning using the speech data of JUs ('multi') achieved a slightly lower PER score than the speaker-dependent model with the random initialization ('rand init'). However, for speakers JM2 and JM5, the performance deteriorated. This result indicates the need for proper initialization and integration of the different data domains.

It is possible to observe the effects of knowledge transfer from the speech data of other domains. Pre-training using the speech data of JUs ('trans. 1') achieved 21.2% and 16.3% average relative improvements compared with 'rand init' and 'multi', respectively. Moreover, fine-tuning with two encoders ('trans. 2') slightly outperformed 'trans. 1'.

Compared with the random initialization, our proposed transfer learning method using speech obtained from JUs

TABLE 3. Phoneme error rates [%] estimated for each method. Jpn and Eng denote Japanese and English respectively. All systems are based on the target speaker-dependent models.

	Pre-training		Fine-tuning		Speaker					
	Architecture	Database	Architecture	Database	JM1	JM2	JM3	JM4	JM5	mean
rand init	-	-	single LAS	JD	48.70	19.29	21.56	53.75	49.16	38.49
multi	-	-	single LAS	JD & JU	40.37	22.53	18.81	49.67	49.72	36.22
trans. 1	single LAS	JU	single LAS	JD	35.34	18.40	11.06	41.63	45.24	30.33
trans. 2	single LAS	JU	two-encoder LAS	JD & JU	35.09	17.17	10.42	40.32	43.74	29.35
trans. 3	two-decoder LAS	JU & ED	single LAS	JD	26.37	15.95	9.66	33.86	42.60	25.69

and English dysarthric speech ('trans. 3') achieved a 33.3% relative improvement. Furthermore, we obtained a significant improvement of 15.3% relative PER compared to pre-training excluding speech data obtained from EDs.

C. DISCUSSION

We assume that the ability to recognize Japanese dysarthric speech can be transferred from the combination of dysarthric speeches in other languages with physically unimpaired Japanese speech. The proposed approach achieved significantly better performance than the random initialization. These results indicate that the language transferability as provided in [25] is effective even in the case of dysarthric speech. Additionally, we obtained significant improvements even if the speaker had small amounts of speech data. For example, in the case of speaker JM1, our proposed approach achieved a 45.9% relative improvement compared with the random initialization. As the speech data of speakers with articulation disorders is quite limited, this effect is deemed to be crucial for the purposes of the present research.

VI. CONCLUSION

In this paper, we proposed a novel knowledge transfer approach for dysarthric speech recognition that uses speech data obtained both from physically unimpaired persons and from persons with dysarthria speaking in a different language. The amount of speech data obtained from speakers with articulation disorders is quite limited owing to the athetoid symptoms. To solve this problem, we used additional speech data obtained from physically unimpaired persons speaking in a different language. We demonstrated the effectiveness of the proposed approaches through the phoneme recognition task.

Our future research will further investigate the usage of speech data published in the publicly available databases obtained from persons with dysarthria speaking in other languages. This research will then also investigate increasing the amount of the speech data of JUs and persons with dysarthria who do not speak Japanese. Moreover, we will apply our proposed approach to the conventional deep neural network-hidden Markov model hybrid ASR system to compare the results of the proposed methods against the previously proposed frameworks [22]. It should be noted that, in this work, we use the phoneme sequence to train the model. However, dysarthric speech may contain undesirable errors with respect to an expected phoneme sequence owing to the athetoid symptoms. In such cases, the given training phoneme sequence would be unreliable. To solve this problem, we will investigate approaches to apply unsupervised domain adaptation. Domain adaptation using deep learning has been researched and has shown the remarkable progress. Therefore, we expect the application of a domain adaptation technique to dysarthric ASR to improve performance.

REFERENCES

[1] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 3rd ed. New York, NY, USA: Elsevier, 2013.

[2] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 70–78.

[3] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learn. Speech Recognit. Related Appl.*, Vancouver, BC, Canada, 2009, p. 39.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 4688–4691.

[5] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 8614–8618.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[7] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brighton, U K., May 2019, pp. 6395–6399.

[8] X. Menéndez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, USA, Oct. 1996, pp. 1962–1965.

[9] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. INTERSPEECH*, Brisbane, QLD, Australia, 2008, pp. 1741–1744.

[10] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, 2012.

[11] D.-L. Choi, B.-W. Kim, Y.-J. Lee, Y. Um, and M. Chung, "Design and creation of dysarthric speech database for development of QoLT software technology," in *Proc. Int. Conf. Speech Database Assessments (Oriental COCOSDA)*, Oct. 2011, pp. 47–50.

[12] K.-H. Wong, Y. T. Yeung, E. H. Y. Chan, P. C. M. Wong, G.-A. Levow, and H. Meng, "Development of a Cantonese dysarthric speech corpus," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 329–333.

[13] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 6009–6013.

[14] B. Vachhani, C. Bhat, B. Das, and S. K. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1854–1858.

[15] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brighton, U.K., May 2019, pp. 5836–5840.
[16] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "Integration of metamodel and acoustic model for speech recognition," in *Proc. INTERSPEECH*, Brisbane, QLD, Australia, 2008, pp. 2234–2237.
[17] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Saint Malo, France, Oct. 2010, pp. 517–520.
[18] Y. Takashima, T. Nakashika, T. Takiguchi, and Y. Ariki, "Feature extraction using pre-trained convolutive bottleneck nets for dysarthric speech recognition," in *Proc. EUSIPCO*, Nice, France, Aug./Sep. 2015, pp. 1411–1415.
[19] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-discriminative features for dysarthric speech conversion," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3374–3378.
[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2672–2680.
[21] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 471–475.
[22] N. M. Joy, S. Umesh, and B. Abraham, "On improving acoustic models for TORGO dysarthric speech database," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2695–2699.
[23] S. Chandrakala and N. Rajeswari, "Representation learning based speech assistive system for persons with dysarthria," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1510–1517, Sep. 2017.
[24] N. M. Joy and S. Umesh, "Improving acoustic models in TORGO dysarthric speech database," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 637–645, Mar. 2018.
[25] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. J. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 4904–4908.
[26] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 369–376.
[27] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 577–585.
[28] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 4960–4964.
[29] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[30] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, 1990.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

YUKI TAKASHIMA received the B.E. and M.E. degrees in computer science from Kobe University, in 2015 and 2017, respectively. He is currently a Research Fellow (DC1) with the Japan Society for the Promotion of Science. His research interests include speech and image recognition, and statistical signal processing. He is a member of ASJ.

RYOICHI TAKASHIMA received the B.E., M.E., and Dr.Eng. degrees in computer science from Kobe University, in 2008, 2010, and 2013, respectively. From 2013 to 2018, he was a Researcher with Hitachi Ltd., Tokyo, Japan, and from 2016 to 2018, he had been on loan to the National Institute of Information and Communication Technology (NICT), Kyoto, Japan. He is currently an Associate Professor with Kobe University. His research interests include machine learning and signal processing. He is a member of ASJ.

TETSUYA TAKIGUCHI received the M.Eng. and Dr.Eng. degrees in information science. He was a Researcher with the Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University, where he has been a Professor, since 2016. From May 2008 to September 2008, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington. From March 2010 to September 2010, he was a Visiting Scholar with the Institute for Learning and Brain Sciences, University of Washington. From April 2013 to October 2013, he was a Visiting Scholar with the Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He is a member of IEICE, IPSJ, and ASJ.

YASUO ARIKI received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, in 1974, 1976, and 1979, respectively. He was an Assistant Professor with Kyoto University, from 1980 to 1990, and stayed at Edinburgh University as a visiting academic, from 1987 to 1990. From 1990 to 1992, he was an Associate Professor and from 1992 to 2003, a Professor with Ryukoku University. From 2003 to 2016, he was a Professor with Kobe University. He is mainly engaged in speech and image recognition and interested in dialogue systems. He is a member of IEICE, IPSJ, JSAI, and ASJ.