INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Dysarthric Speech Recognition Based on Deep Metric Learning

Yuki Takashima[1], *Ryoichi Takashima*[2], *Tetsuya Takiguchi*[2], *Yasuo Ariki*[2]

[1]Hitachi, Ltd. Research & Development Group, Japan
[1]Graduate School of System Informatics, Kobe University, Japan
yuki.takashima.ot@hitachi.com, rtakashima@port.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

We present in this paper an automatic speech recognition (ASR) system for a person with an articulation disorder resulting from athetoid cerebral palsy. Because their utterances are often unstable or unclear, speech recognition systems have difficulty recognizing the speech of those with this disorder. For example, their speech styles often fluctuate greatly even when they are repeating the same sentences. For this reason, their speech tends to have great variation even within recognition classes. To alleviate this intra-class variation problem, we propose an ASR system based on deep metric learning. This system learns an embedded representation that is characterized by a small distance between input utterances of the same class, while the distance of the input utterances of different classes is large. Therefore, our method makes it easy for the ASR system to distinguish dysarthric speech. Experimental results show that our proposed approach using deep metric learning improves the word-recognition accuracy consistently. Moreover, we also evaluate the combination of our proposed method and transfer learning from unimpaired speech to alleviate the low-resource problem associated with impaired speech.

Index Terms: assistive technology, dysarthria, metric learning, speech recognition

1. Introduction

In this study, we focus 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 utterances are often unstable or unclear owing to their athetoid symptoms. These symptoms also restrict the movement of their 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].

Automatic speech recognition (ASR) has gained wide use in items such as personal assistants on smartphones. In addition, remarkable progress has been made with respect to recent developments in deep learning for ASR [3, 4, 5] in fields with access to a large amount of training data. However, most dysarthric speech cannot be recognized correctly because these ASR systems are trained on "typical" speech. For persons with articulation disorders, it is difficult to collect a sufficient amount of speech data to train the model. Moreover, their speech style is quite different from that of physically unimpaired persons owing to their athetoid symptoms, which makes it difficult to recognize their speech. Therefore, special consideration is needed to construct an ASR system that works well for dysarthric speech.

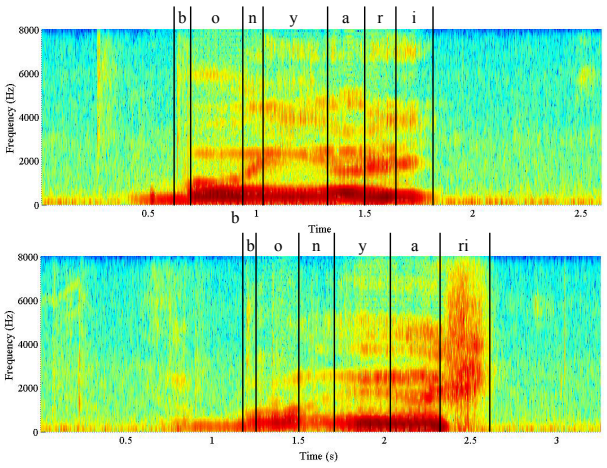


Figure 1: *Examples of spectrogram uttered for /b o n y a r i/ of a person with an articulation disorder.*

In this paper, we focus on the intra-class variation problem associated with dysarthric speech recognition. Unlike the speech of physically unimpaired persons, the speech of a person with an articulation disorder has great variations, even when he/she utters the exact same content. This is because sudden breaths and lost consonants, for example, are unexpectedly caused by athetoid symptoms. When dealing with speech recognition, this phenomenon causes a large variation in the speech features within the same class. Fig. 1 depicts two spectrograms of a person with an articulation disorder for the Japanese word "bonyari" ("hazy" in English). We can see that there are significant differences between the two spectrograms although those examples represent the same word. To alleviate this intra-class variation problem, we employ deep metric learning [6, 7, 8], which makes a neural network-based model learn the similarity or distance between samples from the observed data. This technique maps the inputs of the same class into similar embeddings, while mapping those of a different class into dissimilar embeddings. In this paper, we propose a deep metric learning-based ASR system that takes into consideration intra-class variation. Our method learns not only the boundary that classifies the dysarthric speech, but also the discriminative embedding inside the network. As the metric learning method in this work, we employ the additive angular margin loss (ArcFace) function [8] that has shown significant performance for face recognition tasks. ArcFace has the advantages of intra-class compactness and inter-class discrepancy, and can be applied to general discriminative tasks.

In experiments, we show the effectiveness of our proposed approach through a word-recognition task. We also investigate whether or not our approach can be combined with trans-

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-2267

fer learning to further enhance performance. This is because of the problem of the limited data availability associated with dysarthric speech. To solve this problem, we employ transfer learning from the speech data of physically unimpaired persons [9].

The rest of this paper is organized as follows: In Section 2, related works are described. In Section 3, our proposed method is explained. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Related works

In this work, we focus on a Japanese person with an articulation disorder. Several databases are publicly available for clinical speech applications [10, 11, 12]. There has been some work done on developing an ASR system using these databases [13, 14, 15]. Christensen, *et al.* [16] have shown that, through various adaptation experiments, adapting an ASR system on typical speech to the domain of disordered speech improves performance. In [17], an articulatory-based representation has been proposed to obtain robustness for the inter- and intra-speaker variability in disordered speech using deep neural networks. To establish a sufficiently large training dataset for multi-speaker ASR, speech data synthesis based on a vocal tract model was used. In another clinical speech domain, some researchers have analyzed for aphasic speech with ASR using a large-scale aphasic speech corpus [18, 19]. These researches used databases that include mainly English speakers. However, 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.

Metric learning learns a metric function from training data to calculate the similarity or distance between samples. There are two main approaches being taken in this research: approaches that use an intermediate layer in a multi-class classifier as an embedding [20] and those that learn an embedding directly [7]. These methods have shown successful improvement for face recognition. Among them, in this work, we employ approaches based on an angular margin penalty [8, 21, 22]. These approaches train a classification network where the weights in the last fully-connected layer represent the basis of each class. During training, a penalty is added to the correct class output. The model then tries to estimate the correct output class that overcomes the penalty. This mechanism leads to an improvement in the discriminative power of the trained model and useful embedding for various tasks [23, 24]. In our task, we expect that the trained model has adequate robustness for unstable dysarthric speech.

There is the problem of limited data availability associated with dysarthric speech. To construct an ASR system that works well for dysarthric speech, it is necessary to compensate for the low availability of training data. There are two main approaches being researched: voice conversion and transfer learning. Voice conversion (VC) is a technique for converting the specific information in speech while maintaining the other information in the utterance. VC has been applied to increase the intelligibility of dysarthric speech [25, 26] or to generate pseudo-dysarthric speech for data augmentation [27, 28]. However, in such a conversion-based approach, the performance of ASR depends on the quality of the generated speech. Transfer learning is a technique [29] that seeks to apply the knowledge learned in one or more domains or tasks to another domain or task. Some studies have shown that using the data of physically unimpaired persons boosts recognition performance [9, 30]. This approach is advantageous because we can use real speech that is not degraded to train the model. Therefore, we also combine our proposed method with transfer learning to further improvements.

3. Proposed method

In this work, we employ the additive angular margin loss (ArcFace) function [8] as our deep metric learning method. This method has the advantages of achieving remarkable performance and being easy to implement. Moreover, as ArcFace does not require any additional parameters, it is favored in low-resource scenarios, such as ASR for dysarthric speech.

3.1. ArcFace

ArcFace is defined in the last fully-connected layer. Given an embedding feature $f_i = f(x_i) \in \mathbb{R}^K$ and the weight $w_j \in \mathbb{R}^K$ in the last fully-connected layer, we normalize these vectors as $f' = \frac{f}{||f||}$ and $w' = \frac{w}{||w||}$. Here, K is the number of dimensions of the embedding feature. The cosine similarity between the embedding feature and the weight is equal the dot product between f' and w' as follows:

$$\cos\theta_{ij} = \frac{w_j^T f_i}{||w_j||\,||f_i||} = w_j'^T f_i', \tag{1}$$

where i and j denote the i-th sample in a batch and the j-th column vector of the weight matrix $W = [w_1, ..., w_J] \in \mathbb{R}^{K\times J}$, respectively. J indicates the number of classes. θ_{ij} is the angle between the weight w_j and the feature f_i. ArcFace adds an additive angular margin penalty m between f_i and w_{l_i} to simultaneously enhance the intra-class compactness and inter-class discrepancy. Here, l_i is the class to which f_i belongs. The ArcFace loss function is defined as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{il_i}+m)}}{e^{s\cos(\theta_{il_i}+m)}+\sum_{j=1,j\neq l_i}^{J}e^{s\cos\theta_{ij}}}, \tag{2}$$

where N and s are the batch size and a re-scale parameter, respectively. θ_{ij} can be calculated by the arc-cosine function. Finally, the ArcFace loss can be written as a cross entropy loss function. As ArcFace does not require any other loss functions or constraints, the training is extremely stable. (For more details, see [8].)

3.2. Application to dysarthric speech recognition

In this paper, we apply deep metric learning for dysarthric ASR in order to obtain a robust model for the variation of dysarthric speech. Deep metric learning provides a smaller intra-class variation of the embedding feature and thus improves the discriminative power of the model.

Our model consists of two parts: an encoder and a classifier, as depicted in Fig. 2. The encoder is a stacked pyramid bidirectional long short-term memory (pBLSTM) [31]. The pyramid structure reduces the computational complexity and the convergence time, and allows the subsequent module to extract the relevant information from a smaller number of time steps. The encoder transforms an input sequence $\boldsymbol{x} = (x_1, ..., x_t, ..., x_T)$ of acoustic features into a high-level representation $\boldsymbol{h} = (h_1, ..., h_u, ..., h_U)$, where x_t, h_u, T and $U \leq T$ are the input acoustic feature frame, the encoder output feature, the number of the input acoustic features, and the number of the encoder output features, respectively.

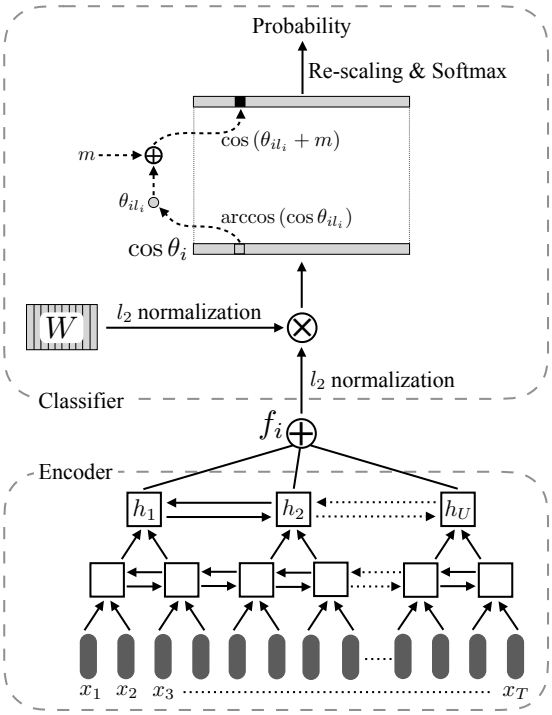


Figure 2: *Overview of our proposed method.*

We get the embedding by summing the encoded features over the time axis as $f = \sum_{u=1}^{U} h_u$. The classifier is a fully-connected layer to estimate the corresponding word label. Then, the ArcFace loss function is calculated from the embedding feature f using Eq. (2). During testing, we use this model to generate the probability of a word as setting the margin m to zero.

4. Experiments

4.1. Conditions

Our proposed approach was evaluated on a word-recognition task carried out on five Japanese-speaking males with athetoid-type cerebral palsy. For each subject, we recorded 216 phonetically-balanced words that are listed in the ATR Japanese speech database [32]. Each word was continuously and repeatedly uttered three or five times by the subject. (The number of repetitions varies depending on the symptoms of each subject, as shown in Table 1.) In our experiments, the first utterances of every word were used for evaluation, and the other utterances (e.g., 1,080 - 216 = 864 utterances for JM3) were used to train a model [1]. When we trained modules, we used only the single speaker's speech. In other words, 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 10ms over a 25ms window. The margin m and the re-scaling parameter s were set to 0.5 and 30, respectively.

[1] When the dysarthric person with athetoid-type cerebral palsy utters the same word repeatedly, the first utterance, in particular, tends to be unstable and difficult to recognize [33]. This is because the athetoid symptoms often occur when he/she is starting to speak. Since the ASR system should recognize the speech correctly without any retries, we target the first utterance to evaluate the ASR performance.

Table 1: *Dataset statistics of Japanese people with the articulation disorder.*

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

Table 2: *Word recognition accuracy [%] for each speaker using the speaker-independent GMM-HMM.*

JM1	JM2	JM3	JM4	JM5
10.8	4.3	7.0	1.4	1.9

For the baseline system, we trained the monophone-hidden Markov models (54 phonemes) with 3 states and 6 mixtures of Gaussians ('GMM-HMM'). To evaluate the effect of ArcFace, we also trained the model using the traditional softmax loss. Considering the network configuration, we used 2 layers of 512 pBLSTM nodes (256 nodes per direction) and the final fully-connected layer (216 nodes). The output dimensionality was 216. The network was optimized using an Adam optimizer [34]. The batch size was 1, and the learning rate was set to 1e-4. When training models from the random initialization, the number of epochs was 50.

4.2. Preliminary experiment

Our recorded dataset described in section 4.1 does not include a pathologist's assessment of the speaker's symptom severity. Therefore, instead of an assessment by a pathologist, to show how the speaking styles of the evaluated five subjects differ from those of physically unimpaired (PU) people, we evaluate the ASR performance of each subject on a speaker-independent GMM-HMM-based ASR system trained using the speech of PU persons. The training data for this system consists of 216 words from 9 PU speakers (five males and four females) recorded in the ATR database (216 $\times$ 9 = 1,944 utterances in total). Table 2 shows the word-recognition accuracy of each subject. In our evaluation of recognizing the utterances of a physically unimpaired male, the word-recognition accuracy was 99.0%. However, as shown in table 2, we found that this model can hardly recognize the speech of the subjects in the experiment. This result indicates that the speaking styles of those subjects are quite different from those of PU people, and, therefore, it supports our motivation to employ a speaker-dependent ASR system for the dysarthric person.

4.3. Results and discussion

We confirmed the performance of the model trained on the speech data of the target dysarthric subject. Fig. 3 shows the word recognition accuracy corresponding to each method. In this figure, a higher value means a better result. In our preliminary experiment, in the case of a physically unimpaired person, softmax loss attains an accuracy of 98.61%. Our proposed method using ArcFace achieved relative improvements of 33.8% and 55.9% on average compared with the traditional GMM-HMM-based model and the model trained on the softmax loss, respectively. This is because that the ArcFace loss improves the discriminative power of the ASR model while keeping the amount of parameters. We visualized examples

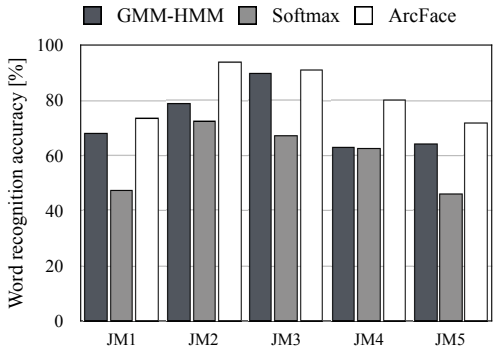
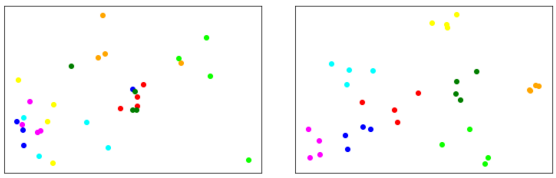


Figure 3: *Word recognition accuracy for each method.*



(a) Softmax (b) ArcFace

Figure 4: *Visualization of embedding features trained from the data of speaker JM3 using principal component analysis.The difference in color corresponds to the difference in class (word). We depicted samples of only ten classes in whole of 216 classes for simplicity. Each class has four samples.*

of learned embeddings as shown in Fig. 4. In the case of the softmax loss, embeddings are chaotically distributed without making any clusters. This is because there is no discriminative constraint. In contrast, we can see that the embeddings on ArcFace are clearly separated for each class. This figure shows that the ArcFace loss makes the embedding more discriminative. These results show that metric learning is especially effective for dysarthric ASR.

4.4. Transfer learning from the speech data of physically unimpaired persons

As shown in Fig. 3, we also found that the DNN with softmax loss showed lower accuracies than GMM-HMM. In general, DNN-based models require larger training data than GMM-based models. Therefore, it can be considered that the training data is insufficient to achieve the full potential of DNN-based models. However, in the real situation, it is difficult to collect sufficient training data from the target dysarthric speaker. To investigate this issue, we evaluate the DNN-based models by combining them with transfer learning approach [29] from the speech data of PU persons. In our previous work [35], we have shown that the speech data of PU persons can be of help in training the model efficiently. In this work, we pre-train a model using the speech data of PU persons, and then, for each dysarthric subject, we fine-tune it using the subject's speech.

In the experiment, the speech data of five PU males and five PU females were used to train the initial model for transfer learning. Their speech is stored in the ATR Japanese speech database. For each speaker, we used the same 216 words as those uttered by dysarthric subjects. When we pre-trained modules using the speech of PU persons, we used all the speakers' speech. The number of epochs of fine-tuning was 10.

The results of transfer learning from the speech data of PU persons are shown in Fig. 5. The model trained by the ArcFace loss outperforms the model trained by the softmax loss. These results indicate that the knowledge transferred from the speech data of PU persons is informative for dysarthric ASR.

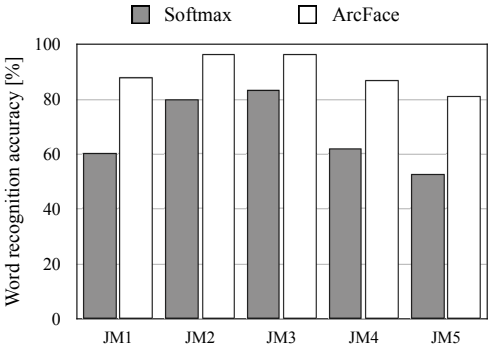


Figure 5: *Word recognition accuracy [%] for each method using transfer learning.*

Table 3: *Average word recognition accuracy [%] for each method with or without the final fully-connected layer update.*

Method	Random init.	Transfer learning	
		w/ update	w/o update
Softmax	63.28	67.80	68.94
ArcFace	82.07	89.79	87.89

In a previous work [36] on dysarthric ASR, fine-tuning while fixing the final layer achieved superior performance compared to fine-tuning the entire network. Table 3 shows the word recognition accuracy using transfer learning with or without the final layer update. In the case of the softmax loss, fine-tuning without updating the final layer improved the accuracy compared to updating all layers. However, fine-tuning on ArcFace without updating the final layer showed slightly lower performance. We guess that this is because it is difficult to make the margin between embeddings due to the fixing of the weight in the final layer.

5. Conclusions

In this paper, we propose the deep metric learning-based ASR system for people with an articulation disorder. Compared to the traditional softmax loss function, our proposed method makes it possible to alleviate the intra-class variation problem of dysarthric speech, and to improve the word-recognition performance. Additionally, we also evaluated the combination of our proposed method with transfer learning. Experimental results show that transfer learning using additional speech data from a physically unimpaired person, further improves performance.

In this study, we constructed a speaker-dependent system. In future work, we will work on a speaker-independent system for dysarthric speech and study how it can be put to practical use. Moreover, our future work includes the evaluation of our proposed method on non-Japanese dysarthric corpora [10, 11, 12], in addition to the investigation of continuous speech recognition. Deep metric learning could be applied to few-shot learning [37] where a classifier generalizes to new classes using only a small number of examples of each new class. In the future, we will further investigate the potential of our proposed method, focusing on an unknown word.

6. References

[1] J. R. Duffy, *Motor Speech disorders: Substrates, differential diagnosis, and management*, 3rd ed. Elsevier, 2013.

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

[3] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[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 Processing*, 2011, pp. 4688–4691.

[5] T. N. Sainath, A. rahman Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2013, pp. 8614–8618.

[6] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[9] H. Christensen, M. B. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Proc. ISCA Interspeech*, 2013, pp. 3642–3645.

[10] 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. Processing*, 1996.

[11] 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. ISCA Interspeech*, 2008, pp. 1741–1744.

[12] F. Rudzicz, A. K. Namasivayam, and T. 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.

[13] N. M. Joy, S. Umesh, and B. Abraham, "On improving acoustic models for TORGO dysarthric speech database," in *Proc. ISCA Interspeech*, 2017, pp. 2695–2699.

[14] 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, 2017.

[15] 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, 2018.

[16] H. Christensen, S. Cunningham, C. Fox, P. D. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. ISCA Interspeech*, 2012, pp. 1776–1779.

[17] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *ITG Symposium on Speech Communication*. VDE / IEEE, 2018, pp. 1–5.

[18] D. Le and E. M. Provost, "Improving automatic recognition of aphasic speech with aphasiabank," in *Proc. ISCA Interspeech*, 2016, pp. 2681–2685.

[19] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, pp. 1–12, 2018.

[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[21] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2017, pp. 6738–6746.

[22] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. IEEE conf. Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[23] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2019, pp. 3652–3656.

[24] X. Lu, P. Shen, S. Li, Y. Tsao, and H. Kawai, "Class-wise centroid distance metric learning for acoustic event detection," in *Proc. ISCA Interspeech*, 2019, pp. 3614–3618.

[25] R. Aihara, T. Takiguchi, and Y. Ariki, "Phoneme-discriminative features for dysarthric speech conversion," in *Proc. ISCA Interspeech*, 2017, pp. 3374–3378.

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

[27] 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 Processing*, 2018, pp. 6009–6013.

[28] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. ISCA Interspeech*, 2018, pp. 471–475.

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

[30] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, "Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition," *IEEE Access*, vol. 7, pp. 164 320–164 326, 2019.

[31] 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 Processing*, 2016, pp. 4960–4964.

[32] 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 Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[33] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "Integration of metamodel and acoustic model for speech recognition," in *Proc. ISCA Interspeech*, 2008, pp. 2234–2237.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representations*, 2015.

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

[36] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, and Y. Matias, "Personalizing asr for dysarthric and accented speech with limited data," in *Proc. ISCA Interspeech*, 2019, pp. 784–788.

[37] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4077–4087.