Unsupervised Domain Adaptation for Dysarthric Speech Detection via Domain Adversarial Training and Mutual Information Minimization

Disong Wang[1], Liqun Deng[2], Yu Ting Yeung[2], Xiao Chen[2], Xunying Liu[1], Helen Meng[1]*

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Huawei Noah's Ark Lab, China

{dswang,xyliu,hmmeng}@se.cuhk.edu.hk, {dengliqun.deng,yeung.yu.ting,chen.xiao2}@huawei.com

Abstract

Dysarthric speech detection (DSD) systems aim to detect characteristics of the neuromotor disorder from speech. Such systems are particularly susceptible to domain mismatch where the training and testing data come from the source and target domains respectively, but the two domains may differ in terms of speech stimuli, disease etiology, etc. It is hard to acquire labelled data in the target domain, due to high costs of annotating sizeable datasets. This paper makes a first attempt to formulate cross-domain DSD as an unsupervised domain adaptation (UDA) problem. We use labelled source-domain data and unlabelled target-domain data, and propose a multi-task learning strategy, including dysarthria presence classification (DPC), domain adversarial training (DAT) and mutual information minimization (MIM), which aim to learn dysarthria-discriminative and domain-invariant biomarker embeddings. Specifically, DPC helps biomarker embeddings capture critical indicators of dysarthria; DAT forces biomarker embeddings to be indistinguishable in source and target domains; and MIM further reduces the correlation between biomarker embeddings and domain-related cues. By treating the UASPEECH and TORGO corpora respectively as the source and target domains, experiments show that the incorporation of UDA attains absolute increases of 22.2% and 20.0% respectively in utterance-level weighted average recall and speaker-level accuracy.

Index Terms: dysarthric speech detection, unsupervised domain adaptation, adversarial training, mutual information

1. Introduction

Dysarthria encapsulates various speech disorders caused by a set of neurodegenerative conditions and diseases, such as cerebral palsy, Parkinson's disease or amyotrophic lateral sclerosis, which lead to poor control of muscles including lips, tongue, jaw, velum and throat [1]. Therefore, patients with dysarthria often produce harsh and breathy speech with unstable prosody and imprecise articulation. To facilitate the clinical diagnosis and treatment of neurological diseases, early onset detection of dysarthric speech may serve as a promising tool.

Current research on dysarthric speech detection (DSD) mainly focuses on building models trained and validated on the data from the same domain, where high DSD accuracy can be achieved. However, DSD models are less robust under domain mismatch conditions, i.e., DSD models trained by the data from the source domain will suffer from marked performance degradation when they are exposed to data from an unseen target domain with different statistical distributions. The difference may be the types of stimuli, phonetic context, vocal quality, disease etiology, recording environments and devices, etc. Training on

labelled data from the target domain will improve DSD accuracy. However, such data is often too costly to acquire [2]. Therefore, leveraging available, labelled source data along with unlabelled target data to create a DSD model that generalizes well to target domain is desirable, and can be treated as an *unsupervised* domain adaptation (UDA) problem [3], as no supervision information is available in the target domain.

To alleviate the domain mismatch issues, domain adversarial training (DAT) and mutual information minimization (MIM) are proposed to extract domain-invariant biomarker embeddings that are used to identify the dysarthria for accurate DSD. The UDA framework consists of three learning tasks: The primary task employs a biomarker encoder to extract biomarker embeddings for dysarthria presence classification (DPC). The second task applies DAT to force biomarker embeddings to be indistinguishable in the source and target domains by deceiving a domain discriminator, so that the biomarker embeddings from source and target domains have similar distributions. The last task strives to minimize the mutual information between the biomarker embeddings and the counterpart domain embeddings that are extracted by a domain encoder, which further removes domain cues in biomarker embeddings. The proposed UDA framework facilitates the learning of biomarker embeddings that are invariant across domains while capturing critical information used for dysarthria detection.

This work paves the way towards the under-explored problem of cross-domain DSD that is widely encountered in practice. The main contribution lies in the novel approach by combining DPC, DAT and MIM for cross-domain DSD that is formulated as an UDA problem for the first time. Extensive experiments have been conducted to verify the effectiveness of proposed methods by using different kinds of neural networks.

2. Related work

Diagnosis of speech symptoms is commonly used in the identification of dysarthria. Traditional approaches are involved with clinicians or speech-language pathologists conducting a series of subjective listening tests, e.g., Frenchay Dysarthria Assessment [4], which may be affected by subjectivity among assessors. This motivates researchers to turn to objective evaluation of dysarthria based on building statistical DSD models, which are economical with potentials for remote patient rehabilitation monitoring [5]. To develop an efficient DSD system, previous work mostly design handcrafted acoustic features as biomarkers that capture dysarthric patterns, including prosodic, spectral, phonological and glottal features [6–10]. Besides, automatic feature extraction from raw speech via a learnable frontend is proposed in [11]. Though significant progress has been achieved, effectiveness of previous methods requires further verification under cross-domain conditions. A few previ-
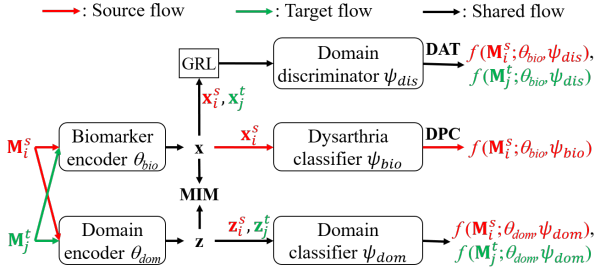

Figure 1: *Diagram of the proposed UDA framework for DSD based on multi-task learning, which includes dysarthria presence classification (DPC), domain adversarial training (DAT) and mutual information minimization (MIM). 'GRL' denotes the gradient reversal layer.*

ous efforts investigate this problem by carefully designing and selecting domain-robust features for cross-language [12] and cross-dataset [13] scenarios. In contrast, this paper focuses on automatically extracting dysarthria-discriminative and domain-invariant biomarkers from simple acoustic features, e.g., mel-spectrograms, which is formulated as an UDA problem.

UDA has been explored in many speech tasks including automatic speech recognition [14, 15], speech emotion recognition [16] and speaker recognition [17], where DAT [18] is widely used to remove domain variations and project the data of different domains into the same subspace. There is still much room to apply DAT for the DSD task. Besides, inspired by information theory [19], MIM is proposed to reduce the dependency between biomarker embeddings and domain-related information. The combination of DAT and MIM forces biomarker embeddings to achieve robustness for detection of dysarthria.

3. Proposed approach

Assuming that there are I and J utterances with corresponding mel-spectrograms $\{\mathbf{M}_i^s\}$ and $\{\mathbf{M}_j^t\}$ in the source domain and target domain, respectively. Each source mel-spectrogram $\mathbf{M}_i^s$ is associated with a binary label y_i^s denoting whether the corresponding speech is dysarthric, while no such label is provided in the target domain. Given the data $\{\mathbf{M}_i^s, y_i^s\}(i = 1, 2, ..., I)$ and $\{M_j^t\}(j = 1, 2, ..., J)$, the goal is to build a DSD system that generalizes well to the target domain. To achieve robustness in a DSD system, we propose a multi-task learning based UDA framework as shown in Figure 1, which consists of three learning tasks: dysarthria presence classification, domain adversarial training and mutual information minimization.

3.1. Dysarthria presence classification (DPC)

This primary task performs binary classification for the presence or absence of dysarthria by using the labelled source data. Specifically, a biomarker encoder θ_{bio} takes in mel-spectrogram $\mathbf{M}_i^s$ to derive a single vector $\mathbf{x}_i^s$, which is denoted as biomarker embedding and fed into the dysarthria classifier ψ_{bio}, to give the dysarthria presence posterior $f(\mathbf{M}_i^s; \theta_{bio}, \psi_{bio})$. During training, the biomarker encoder and dysarthria classifier are optimized to minimize the cross-entropy loss as follows:

$$L_{bio} = -\frac{1}{I}\sum\nolimits_{i=1}^{I}\left[y_i^s \log f\left(\mathbf{M}_i^s; \theta_{bio}, \psi_{bio}\right) + \\ (1 - y_i^s)\log\left(1 - f\left(\mathbf{M}_i^s; \theta_{bio}, \psi_{bio}\right)\right)\right] \quad (1)$$

3.2. Domain adversarial training (DAT)

The second task applies DAT to render biomarker embeddings indistinguishable in the source and target domains, by introduc-

Algorithm 1. Training process for UDA based DSD system

Input: source data $\{\mathbf{M}_i^s, y_i^s\}$, target data $\{\mathbf{M}_j^t\}$, learning rate α, β and γ

1. **for** each training iteration **do**
2.   freeze θ_{bio}, ψ_{bio}, θ_{dom}, ψ_{dom} and ϕ, compute discrimination loss (2) using $\{\mathbf{M}_i^s\}$ and $\{\mathbf{M}_j^t\}$, then update ψ_{dis}:
3. $\psi_{dis} \leftarrow \psi_{dis} - \alpha\nabla_{\psi_{dis}}L_{dis}$
4.   freeze θ_{bio}, ψ_{bio}, θ_{dom}, ψ_{dom} and ψ_{dis}, compute log-likelihood (5) using $\{\mathbf{M}_i^s\}$ and $\{\mathbf{M}_j^t\}$, then update ϕ:
5. $\phi \leftarrow \phi + \beta\nabla_{\phi}L_{lld}$
6.   freeze ψ_{dis} and ϕ, compute DSD loss (6) using $\{\mathbf{M}_i^s, y_i^s\}$ and $\{\mathbf{M}_j^t\}$, then update θ_{bio}, ψ_{bio}, θ_{dom} and ψ_{dom}:
        $\theta \leftarrow \theta - \gamma\nabla_{\theta}L_{dsd}, \quad \theta \in \{\theta_{bio}, \psi_{bio}, \theta_{dom}, \psi_{dom}\}$
7. **end for**
8. **return** θ_{bio}, ψ_{bio}

ing a gradient reversal layer (GRL) and a domain discriminator ψ_{dis}, as shown in Figure 1. During training, parameters of the domain discriminator and biomarker encoder are updated alternatively. On the one hand, by freezing θ_{bio}, the domain discriminator is trained to determine whether the input biomarker embeddings are from the source or target domains by minimizing the discrimination loss [18]:

$$L_{dis} = \frac{1}{I}\sum\nolimits_{i=1}^{I}\log f\left(\mathbf{M}_i^s; \theta_{bio}, \psi_{dis}\right) + \\ \frac{1}{J}\sum\nolimits_{j=1}^{J}\log\left(1 - f\left(\mathbf{M}_j^t; \theta_{bio}, \psi_{dis}\right)\right) \quad (2)$$

On the other hand, by freezing ψ_{dis}, the biomarker encoder is trained to maximize the above discrimination loss to deceive the discriminator, which is realized via GRL that passes the data during forward propagation and inverts the sign of the gradient during backward propagation. The alternative processes in training force the domain discriminator and biomarker encoder to compete against each other in an adversarial manner [20], which encourages the distribution of biomarker embeddings across domains to be similar, so that the dysarthria-related cues learned from the source domain in the DPC task remain effective in the target domain.

3.3. Mutual information minimization

The last task strives to reduce the dependency between the biomarker embeddings and domain-related information via MIM. To extract domain-related information, a domain encoder θ_{dom} and a domain classifier ψ_{dom} are utilized. The domain encoder takes in $\mathbf{M}_i^s$ and $\mathbf{M}_j^t$ to extract domain emebddings $\mathbf{z}_i^s$ and $\mathbf{z}_j^t$ respectively, which are used for domain prediction via the domain classifier. Therefore, θ_{dom} and ψ_{dom} are jointly trained by minimizing the domain classification loss that is similar with (2) as:

$$L_{dom} = \frac{1}{I}\sum\nolimits_{i=1}^{I}\log f\left(\mathbf{M}_i^s; \theta_{dom}, \psi_{dom}\right) + \\ \frac{1}{J}\sum\nolimits_{j=1}^{J}\log\left(1 - f\left(\mathbf{M}_j^t; \theta_{dom}, \psi_{dom}\right)\right) \quad (3)$$

The embeddings $\mathbf{z}_i^s$ and $\mathbf{z}_j^t$ are domain-dependent and can be used to represent domain-related information.

Then the mutual information between the biomarker embeddings $\mathbf{x}$ ($\mathbf{x}_i^s$ or $\mathbf{x}_j^t$) and domain embeddings $\mathbf{z}$ ($\mathbf{z}_i^s$ or $\mathbf{z}_j^t$) is used to measure the dependency as Kullback-Leibler (KL) divergence between their joint and marginal distribution: $\mathbf{H}(\mathbf{x}, \mathbf{z}) = D_{KL}(p(\mathbf{x}, \mathbf{z}); p(\mathbf{x})p(\mathbf{z}))$. As the computation of

Table 1: *Within-domain DSD results (%) for different methods and corpora, where training and testing are performed in the target domain with labelled data, DSD systems are trained with 10 rounds, and mean ± standard-variance are reported.*

Methods	UASPEECH			TORGO		
	WAR-utterance	UAR-utterance	ACC-speaker	WAR-utterance	UAR-utterance	ACC-speaker
RCNN [10]	85.71 ± 1.43	85.34 ± 1.50	93.57 ± 2.67	52.93 ± 3.78	54.25 ± 2.40	62.86 ± 2.86
RNN-A [11]	86.78 ± 1.56	86.77 ± 1.64	94.29 ± 3.64	58.15 ± 1.83	58.02 ± 1.28	70.00 ± 5.35
CBRNN-A (proposed)	**87.87 ± 1.53**	**87.89 ± 1.56**	**95.71 ± 1.43**	**63.18 ± 1.14**	**62.76 ± 1.93**	**78.57 ± 5.82**

mutual information is challenging for high-dimensional continuous variables with unknown probability distributions, variational contrastive log-ratio upper bound (vCLUB) [21] is used to calculate the mutual information loss as:

$$L_{mi} = E_{p(\mathbf{x},\mathbf{z})}[\log q_\phi(\mathbf{x}|\mathbf{z})] - E_{p(\mathbf{x})}E_{p(\mathbf{z})}[\log q_\phi(\mathbf{x}|\mathbf{z})] \quad (4)$$

where $q_\phi(\mathbf{x}|\mathbf{z})$ is the variational approximation of the ground-truth posterior of $\mathbf{x}$ given $\mathbf{z}$ and can be parameterized by a network ϕ. During training, θ_{bio} and θ_{dom} are optimized to minimize L_{mi} (4), while the variational approximation network ϕ is optimized to maximize the log-likelihood:

$$L_{lld} = \log q_\phi(\mathbf{x}|\mathbf{z}) \quad (5)$$

3.4. Integrating the learning tasks

By combining the three learning tasks, the total DSD training loss used for updating θ_{bio}, ψ_{bio}, θ_{dom} and ψ_{dom} is:

$$L_{dsd} = \lambda_1 L_{bio} - \lambda_2 L_{dis} + \lambda_3 L_{dom} + \lambda_4 L_{mi} \quad (6)$$

where λ_k (k=1, 2, 3, 4) are positive constant weights. The final training process is summarized in Algorithm 1, where the well-trained biomarker encoder θ_{bio} and dysarthria classifier ψ_{bio} are retained to perform the detection of dysarthria.

4. Experiments

4.1. Experimental setup

To verify the effectiveness of proposed methods, the UASPEECH [22] and TORGO [23] corpora are used for experimentation. UASPEECH contains 15 dysarthric speakers (11 males and 4 females) with cerebral palsy, and 13 healthy speakers (9 males and 4 females). Each speaker has three blocks of utterances with isolated words, where the speech stimuli in each block contain 100 uncommon words and 155 repetitive words (i.e., 10 digits, 26 alphabets, 19 computer commands and 100 common words), which are recorded by 7-channel microphone arrays, we select the data of M6-channel for experiments. TORGO contains 7 dysarthric speakers (4 males and 3 females) with cerebral palsy or amyotrophic lateral sclerosis and 7 healthy speakers (4 males and 3 females). Different from UASPEECH, the speech stimuli include not only words, but also non-words and sentences. Words are mainly chosen from the word intelligibility section of the Frenchay Dysarthria Assessment [4] and Yorkston-Beukelman Assessment [24]. Non-words involve 5–10 repetitions of /iy-p-ah/, /ah-p-iy/, and /p-ah-t-ah-k-ah/ along with high-pitch and low-pitch vowels. Sentences are formed by Grandfather passage from the Nemours database [25], 162 sentences from the sentence intelligibility section of Yorkston-Beukelman Assessment, 460 sentences of MOCHA database [26] and spontaneously elicited descriptive texts. Due to discrepancies of speech stimuli types, phonetic context, articulation patterns, recording environments and devices, UASPEECH and TORGO can be treated as *two different domains* with distinct data distributions.

All speech signals are sampled at 16kHz, 80-band mel-spectrogram is calculated with hanning window of 25ms and hop length of 10ms. Utterance-level z-score normalization for mel-spectrograms is performed before feeding them into the DSD system. The biomarker encoder and domain encoder adopt the same architecture that contains Convolution Banks and Recurrent Neural Network [27] with Attention [28] (CBRNN-A). There are 8 convolution banks with kernel size varying from 1 to 8, one-layer long-shot term memory (LSTM) with 128 units is employed for Recurrent Neural Network, attention module includes two linear layers (100 and 1 unit) with a softmax layer to obtain a vector that is used to weight the linear combination of LSTM outputs to form the final biomarker embedding and domain embedding. Both dysarthria and domain classifiers contain a linear layer with the sigmoid function, and the domain discriminator contains two linear layers with hidden size of 128 and 1. The variational approximation $q_\phi(\mathbf{x}|\mathbf{z})$ in (4) is parameterized by the Gaussian distribution as $q_\phi(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mu(\mathbf{z}), diag(\sigma^2(\mathbf{z})))$ with mean $\mu(\mathbf{z})$ and variance $\sigma^2(\mathbf{z})$ inferred by two-way linear layers with a hidden size of 256. All networks are trained by the Adam optimizer [29] for 8 epochs with learning rate α, β and γ set to 1e-4, 1e-4 and 1e-3 respectively, and the weights λ_1, λ_2, λ_3 and λ_4 in loss (6) are set to 1, 1e-1, 1 and 1e-4 respectively.

We compare CBRNN-A with other networks that also use mel-spectrograms to detect dysarthria, including Recurrent Convolutional Neural Network (RCNN) [10] and Recurrent Neural Network with Attention (RNN-A) [11]. We adopt the leave-one-subject-out cross validation scheme, i.e., all speakers are used for training except the one that is left out for testing. Three evaluation metrics are used: (1) Utterance-level weighted average recall (WAR), denoting the ratio of utterances that are correctly classified; (2) Utterance-level unweighted average recall (UAR), denoting the accuracy per class averaged by total number of classes; (3) Speaker-level accuracy (ACC), denoting the ratio of speakers for which more than 50% of the individual's utterances are classified correctly.

4.2. Experimental results and analysis

4.2.1. Within-domain DSD performance

We first evaluate within-domain DSD performance, where the training and testing are both performed in a target domain, assuming that labelled data is provided, i.e., the ideal condition. The results are shown in Table 1, we can see that the proposed CBRNN-A outperforms RCNN and RNN-A with higher utterance-level WAR, UAR and speaker-level ACC for both UASPEECH and TORGO corpora. This shows the effectiveness of CBRNN-A by using multiple convolution banks with varied kernel size to capture articulation patterns at different scales for accurate DSD.

4.2.2. Cross-domain DSD performance

Next, we consider two domain mismatch conditions: 'UASPEECH → TORGO' and 'TORGO → UASPEECH', where the former treats UASPEECH as the source domain and

Table 2: *Cross-domain DSD results (%) for different methods under different domain mismatch conditions.*

| Methods	UASPEECH → TORGO			TORGO → UASPEECH		
	WAR-utterance	UAR-utterance	ACC-speaker	WAR-utterance	UAR-utterance	ACC-speaker
RCNN [10]	32.73 ± 0.26	49.72 ± 0.21	50.00 ± 0.00	59.58 ± 1.55	59.66 ± 1.44	64.43 ± 4.84
+DAT & MIM	50.55 ± 8.75	58.06 ± 1.23	58.57 ± 9.48	64.41 ± 2.88	65.35 ± 2.82	67.14 ± 4.25
RNN-A [11]	34.82 ± 1.65	51.24 ± 0.94	50.00 ± 0.00	64.55 ± 1.82	64.48 ± 1.73	72.29 ± 3.93
+DAT & MIM	52.58 ± 4.78	57.46 ± 4.20	65.71 ± 5.35	67.20 ± 1.34	67.87 ± 1.29	75.00 ± 3.19
CBRNN-A (proposed)	35.21 ± 2.93	51.32 ± 1.27	50.00 ± 0.00	63.14 ± 1.95	62.71 ± 2.10	70.71 ± 5.95
+DAT	53.68 ± 7.08	57.84 ± 6.38	62.86 ± 5.43	66.00 ± 1.89	66.47 ± 1.98	76.43 ± 1.75
+MIM	43.80 ± 3.53	54.27 ± 3.63	51.43 ± 2.86	64.96 ± 3.08	65.15 ± 3.08	71.43 ± 5.31
+DAT & MIM	$\mathbf{57.42 \pm 4.74}$	$\mathbf{60.70 \pm 5.22}$	$\mathbf{70.00 \pm 5.29}$	$\mathbf{68.44 \pm 3.23}$	$\mathbf{68.89 \pm 3.26}$	$\mathbf{79.29 \pm 4.17}$

Table 3: *Utterance-level WAR (%) for words, non-words and sentences under the mismatched condition of 'UASPEECH → TORGO'.*

Methods	Words	Non-words	Sentences
CBRNN-A	32.43 ± 4.10	42.86 ± 1.27	29.88 ± 2.79
+DAT	50.73 ± 5.60	48.19 ± 3.20	58.73 ± 1.52
+MIM	40.89 ± 3.17	45.84 ± 2.15	44.12 ± 5.41
+DAT & MIM	$\mathbf{55.10 \pm 5.44}$	$\mathbf{49.87 \pm 1.43}$	$\mathbf{64.05 \pm 1.29}$

TORGO as the target domain, and vice versa for the latter. Results are shown in Table 2. First, for 'UASPEECH → TORGO', the performance of all DSD systems without DAT and MIM drops significantly. As TORGO has data imbalance where the ratio between healthy and dysarthric utterances is around 2:1, and healthy utterances are often incorrectly classified, less than 50% WAR-utterance and only 50% ACC-speaker are achieved. This shows the susceptibility of DSD systems to domain mismatch issues. Second, detection accuracy can be improved by using DAT or MIM, and the combination of DAT and MIM can greatly boost DSD performance for different kinds of networks, where the proposed CBRNN-A outperforms RCNN and RNN-A when DAT and MIM are used, showing the effectiveness of proposed methods for learning dysarthria-discriminative and domain-invariant biomarker embeddings for robust dysarthria detection. Third, compared with 'TORGO → UASPEECH', larger improvements can be achieved under 'UASPEECH → TORGO' condition by using DAT and MIM, e.g., the absolute values of WAR-utterance and ACC-speaker are increased with 22.2% and 20.0% respectively by using CBRNN-A. As UASPEECH contains utterances with limited words, while TORGO contains richer words and unseen speech stimuli types including non-words and sentences, the DSD systems trained on UASPEECH generalize poorly to TORGO. This can be verified by the utternace-level WAR results for words, non-words and sentences as shown in Table 3, CBRNN-A performs worst for sentences, followed by words and non-words. With the proposed DAT and MIM, 34.2%, 22.7% and 7.0% absolute increase in WAR can be achieved on average for sentences, words and non-words respectively. As TORGO contains richer speech stimuli, DSD systems trained on TORGO have better generalization capability, which can be further enhanced by the proposed DAT and MIM approaches.

4.2.3. Visualization of biomarker embeddings

To acquire an intuition regarding how the UDA framework extracts effective biomarker embeddings, we consider 'UASPEECH → TORGO' condition. Principal component analysis (PCA) is performed on biomarker embeddings extracted by the DSD systems trained without and with DAT & MIM. The first and second components of PCA results are illustrated in Figure 2, where different colors denote different domains, different shapes denote the presence or absence of

(a) w/o DAT & MIM (b) w/ DAT & MIM
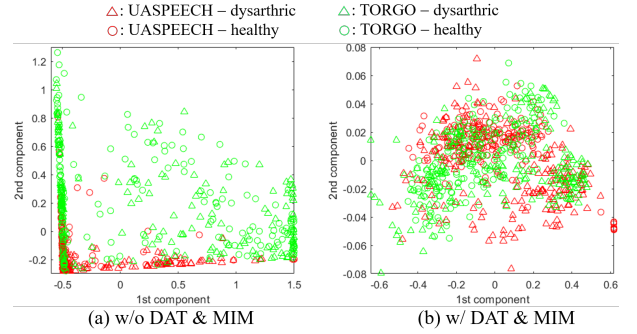
Figure 2: *Visualization of the 1st and 2nd components of the biomarker embeddings extracted from utterances of UASPEECH and TORGO corpora based on DSD systems trained w/o DAT & MIM (left) and w/ DAT & MIM (right) under the mismatched condition of 'UASPEECH → TORGO'.*

dysarthria. We observe that the biomarker embeddings from UASPEECH and TORGO tend to be separate when DAT & MIM is not used, while biomarker embeddings are mixed when DAT & MIM is used, indicating that without additional regularizations, the biomarker embeddings contain domain cues, but these can be effectively removed by proposed DAT and MIM. Besides, it can be seen that dysarthric and healthy biomarker embeddings of DSD systems with DAT & MIM are more dysarthria-discriminative with more obvious cluster formation than those of DSD systems without DAT & MIM, which further proves the superiority of proposed methods to achieve higher detection accuracy across domains.

5. Conclusions

This paper studies an under-explored field of DSD, i.e., cross-domain DSD where the DSD system is trained and tested on different domains with distinct data distributions. We propose a multi-task learning strategy, where the primary task performs DPC using labelled source data, while DAT and MIM tasks leverage large amounts of additional, unlabelled target-domain data that can be easily acquired to align the domain distributions. The proposed approach can obtain domain-invariant biomarker embeddings that contain critical indicators of dysarthria presence for accurate and robust detection. This is verified by extensive experiments with different kinds of network architectures. Our future study will focus on applying and improving the proposed UDA methods for more challenging domain mismatch conditions, e.g., cross-language condition.

6. Acknowledgements

7. References

[1] Y. Yunusova, G. Weismer, J. R. Westbury, and M. J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 596–611, 2008.

[2] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[3] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[4] P. Enderby, "Frenchay dysarthria assessment," *British Journal of Disorders of Communication*, vol. 15, no. 3, pp. 165–173, 1980.

[5] K. Gurugubelli and A. K. Vuppala, "Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6410–6414.

[6] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G. Cecchi, "Detection of amyotrophic lateral sclerosis (als) via acoustic analysis," *bioRxiv*, p. 383414, 2018.

[7] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences." in *Interspeech*, 2018, pp. 3403–3407.

[8] I. Kodrasi and H. Bourlard, "Super-gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6400–6404.

[9] A. Mayle, Z. Mou, R. C. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks." in *Interspeech*, 2019, pp. 4514–4518.

[10] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable deep learning model for the detection and reconstruction of dysarthric speech," *Interspeech*, pp. 3890–3894, 2019.

[11] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5831–5835.

[12] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, 2016.

[13] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence." in *Interspeech*, 2017, pp. 3127–3131.

[14] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[15] D. Woszczyk, S. Petridis, and D. Millard, "Domain adversarial neural networks for dysarthric speech recognition," *Interspeech*, pp. 3875–3879, 2020.

[16] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[17] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[18] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[19] B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel, "Mutual information analysis," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2008, pp. 426–442.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[21] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1779–1788.

[22] 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 *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[23] 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.

[24] K. M. Yorkston, D. R. Beukelman, and C. Traynor, *Assessment of intelligibility of dysarthric speech*. Pro-ed Austin, TX, 1984.

[25] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1962–1965.

[26] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *In Proceedings 5 th Seminar of Speech Production*. Citeseer, 2000.

[27] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[28] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2526–2530.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.