EXPLOITING CROSS DOMAIN ACOUSTIC-TO-ARTICULATORY INVERTED FEATURES FOR DISORDERED SPEECH RECOGNITION

Shujie Hu*, Shansong Liu*, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shoukang Hu, Mingyu Cui, Xunying Liu, Helen Meng

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong, China

ABSTRACT

Articulatory features are inherently invariant to acoustic signal distortion and have been successfully incorporated into automatic speech recognition (ASR) systems for normal speech. Their practical application to disordered speech recognition is often limited by the difficulty in collecting such specialist data from impaired speakers. This paper presents a cross-domain acoustic-to-articulatory (A2A) inversion approach that utilizes the parallel acoustic-articulatory data of the 15-hour TORGO corpus in model training before being cross-domain adapted to the 102.7-hour UASpeech corpus and to produce articulatory features. Mixture density networks based neural A2A inversion models were used. A cross-domain feature adaptation network was also used to reduce the acoustic mismatch between the TORGO and UASpeech data. On both tasks, incorporating the A2A generated articulatory features consistently outperformed the baseline hybrid DNN/TDNN, CTC and Conformer based end-to-end systems constructed using acoustic features only. The best multi-modal system incorporating video modality and the cross-domain articulatory features as well as data augmentation and learning hidden unit contributions (LHUC) speaker adaptation produced the lowest published word error rate (WER) of 24.82% on the 16 dysarthric speakers of the benchmark UASpeech task.

Index Terms— Articulatory Inversion, Dysarthric Speech, Speech Recognition, Domain Adaptation

1. INTRODUCTION

Speech disorders such as dysarthria are often associated with neuro-motor conditions [1], including cerebral palsy [2], amyotrophic lateral sclerosis [3], Parkinson disease [4], stroke or traumatic brain injuries [5], leading to weakness or paralysis of muscles controlling articulation [6] and reduced intelligibility of speech for human listeners. They affect millions of people around the world. People with speech impairment often experience co-occurring physical disabilities and mobility issue. Their difficulty in using keyboard, mouse and touch screen based user interfaces makes speech controlled assistive technologies more natural alternatives [7], even though speech quality is degraded. Despite the rapid progress of automatic speech recognition (ASR) technologies in the past few decades, recognition of disordered speech is still a very challenging task to date due to severe mismatch against normal speech, difficulty in large scale data collection for system development and high level of variability among speakers [8–13].

Human speech production is a process that involves the coordinated movements of various articulators such as the tongue, lips, teeth and palate. Articulatory movement features are inherently invariant to extrinsic acoustic distortion, for example, due to environmental noise. They have been successfully applied to both normal speech [14–19] and pathological speech [20–25] recognition tasks.

The practical and wider use of articulatory features in ASR systems for both normal and disordered speech task domains is often limited by the difficulty in collecting sufficient quantities of such specialist data that are essential for current deep learning technologies. In practice, recording detailed articulatory movements and vocal tract shape normally requires the use of intrusive electromagnetic articulography (EMA) [26] technologies or magnetic resonance imaging (MRI) [27]. In the context of articulatory recordings from impaired speakers, the requirement of specialist facilities is further compounded with their underlying neuro-motor conditions, mobility issues and fatigue when speaking, leading to increasing difficulty in articulatory data collection.

An alternative approach to obtain articulatory information is to estimate it from the more accessible acoustic speech signals using data driven artificial neural network based acoustic-to-articulatory (A2A) inversion [25, 28–31] techniques based on, for example, multilayer perceptron (MLP) [28] and mixture density network (MDN) [30, 31]. As the A2A inversion model training only requires a part of training materials to contain parallel acoustic-articulatory data, the resulting inversion model can be used to produce articulatory features when only audio recordings are available. A wider and more practical application of articulatory feature representation in ASR systems thus becomes possible. Prior researches on A2A inversion were conducted predominantly on normal speech task domains [28–31]. In contrast, very limited researches were carried out on A2A inversion for disordered speech recognition [23, 25].

In order to address the issues mentioned above, this paper presents a cross-domain A2A inversion approach that utilizes the parallel acoustic-articulatory data of 15-hour TORGO corpus [32] in model training before being cross-domain adapted to the 102.7-hour UASpeech corpus [33] to produce articulatory features. Mixture density networks based deep neural network A2A inversion models were used. A cross-domain adaptation network was used to reduce the acoustic mismatch between the TORGO and UASpeech data. On both tasks, incorporating the generated articulatory features consistently outperformed the baseline hybrid DNN/TDNN [34], CTC [35] or Conformer [36] based end-to-end systems constructed using acoustic features only.

The main contributions are summarized below. To the best of our knowledge, this work is the first use of real acoustic-articulatory parallel recordings trained A2A inversion models for articulatory features generation targeting disordered speech recognition. In contrast, related previous works either used: a) synthesized normal

---

*Equal contribution. Xurong Xie is currently with Institute of Software, Chinese Academy of Sciences.

speech acoustic-articulatory features trained A2A inversion models before being applied to dysarthric speech [23], while the large mismatch between normal and impaired speech encountered during inversion model training and articulatory feature generation stages was not taken into account; or b) only considered the cross-domain or cross-corpus A2A inversion [25] while the quality of generated articulatory features was not assessed using the back-end disordered speech recognition systems. In addition, the lowest published WER of 24.82% on the benchmark UASpeech task in comparison against recent researches [8–13, 37–39] was obtained using the proposed cross-domain acoustic-to-articulatory inversion approach.

The rest of the paper is organized as follows. The baseline ASR systems and their incorporation of articulatory features are presented in Section 2. Section 3 presents the cross-domain A2A inversion systems. Experimental results are shown in Section 4. The conclusions are drawn and future works are discussed in Section 5.

2. ARTICULATORY FEATURE BASED DISORDERED SPEECH RECOGNITION

This section describes the time delay neural network (TDNN) [34] based ASR and acoustic-articulatory feature based speech recognition (AASR) system architecture on the TORGO dataset [32] which provides parallel acoustic-articulatory data.
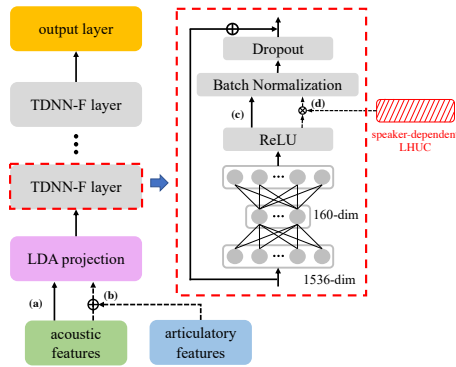

Fig. 1. The factored TDNN based ASR and AASR system architecture for the TORGO task. The AASR system uses the acoustic-articulatory concatenation via connection (b). The connection (d) is used for LHUC-SAT training for both ASR and AASR systems.

Both of the baseline ASR and AASR systems share the same main structure based on a 7-layer factorized TDNN (TDNN-F) model with a semi-orthogonal constraint and they were trained using the sequence discriminative lattice-free MMI (LF-MMI) criterion (see Figure 1). Linear discriminant analysis (LDA) based affine projection was also applied to the input acoustic only for the ASR system, or concatenated acoustic-articulatory features for the AASR system. The following TDNN-F hidden layers positioned after LDA projection are shown in the red dotted box in Figure 1. Each TDNN-F layer contains a set of neural operations performed in sequence immediately after the factorized hidden layers including: rectified linear unit (ReLU) activation, batch normalization and dropout modules. Each hidden layer's inputs prior to context splicing were scaled and added to its outputs by a skip connection.

To model the large variability among disordered speakers, learning hidden unit contributions (LHUC) [40] based speaker adaptive training (SAT) was used (right part of Figure 1). Speaker-level LHUC scaling factors (in red) are applied to the ReLU activation outputs via connection (d). Supervised estimation of LHUC factors is performed for each speaker during the training stage. During test

adaptation, unsupervised LHUC adaptation is used to re-estimate the LHUC scaling factors based on the speaker specific data.

3. ACOUSTIC-TO-ARTICULATORY INVERSION

3.1. In-domain A2A Inversion

Data augmentation techniques play a vital role to address the data sparsity problem in current disordered speech recognition systems [37, 38]. Spectral-temporal perturbation of the limited audio data collected from impaired speakers is normally used to inject more diversity into the augmented data to improve the resulting ASR system generalization on the same task, for example, the TORGO corpus. The construction of AASR systems using such augmented acoustic data requires an in-domain acoustic-to-articulatory (A2A) inversion process to produce the desired articulatory features for the expanded audio data. One of the commonly adopted neural network based A2A inversion methods is based on mixture density networks (MDNs) [30, 31]. This is also considered in this paper. Instead of directly generating articulatory features, MDNs model the Gaussian mixture model density distribution parameters that characterise the articulatory movements. The MDN loss function is defined as

$$\mathcal{L}_{MDN} = -\sum_t \ln \sum_m^M \mathcal{S}_m(\mathbf{y}_t^{\boldsymbol{\lambda}}) \mathcal{N}(\mathbf{a}_t; \boldsymbol{\mu}_{t,m}, \boldsymbol{\sigma}_{t,m}^2) \qquad (1)$$

where M is the number of mixture components, $\mathbf{a}_t$ denotes articulatory feature vector at the t-th frame, $\mathcal{S}$ and $\mathcal{N}$ denote the Softmax activation and Gaussian distribution respectively, $\mathbf{y}_t^{\boldsymbol{\lambda}}$ represents the MDN network output fed into the Softmax activation to produce the mixture component weights $\mathcal{S}_m(\mathbf{y}_t^{\boldsymbol{\lambda}})$ at time t. The t-th frame mixture component mean and variance parameters are predicted using the respective MDN outputs as $\boldsymbol{\mu}_{t,m} = \mathbf{y}_{t,m}^{\boldsymbol{\mu}}$, and $\boldsymbol{\sigma}_{t,m}^2 = \exp^2(\mathbf{y}_{t,m}^{\boldsymbol{\sigma}})$. The articulatory movements directly produced by MDNs usually contain artefacts. These can be further smoothed using the maximum likelihood parameter generation (MLPG) algorithm [41] performed on the articulatory trajectories augmented with their differentials Δ and $\Delta\Delta$.

In this paper, a multi-task learning (MTL) approach was also adopted to construct the A2A inversion system illustrated in the right part of Figure 2, where the acoustic features were fed into the inversion model using the connection (a). Two groups of tasks including: a) an interpolation between the MDN error loss of Eqn. (1), MSE and Pearson correlation computed against the ground truth articulatory features; and b) an auxiliary monophone classification task based on phonetic alignments obtained from the HTK toolkit [42] are combined in the following MTL cost function.

$$\mathcal{L} = \underbrace{\omega_1 \mathcal{L}_{MDN} + \omega_2 \mathcal{L}_{MSE} + \omega_3 \mathcal{L}_{Pearson}}_{\text{regression}} + \underbrace{\omega_4 \mathcal{L}_{CE}}_{\text{classification}} \qquad (2)$$

where $\mathcal{L}_{MDN}$ is calculated by Eqn. (1), $\mathcal{L}_{MSE}$ and $\mathcal{L}_{Pearson}$ are the MSE and negative Pearson correlation coefficient calculated using the inverted and target articulatory features respectively. $\mathcal{L}_{CE}$ is the monophone level cross entropy (CE) loss. In this paper, the task weight parameters $\omega_1, \omega_2, \omega_3, \omega_4$ were set as the same weight 0.25[1].

3.2. Cross-domain A2A Inversion

Due to the acoustic domain mismatch, a direct cross-domain application of the A2A inversion model (described in Section 3.1) trained on the TORGO acoustic-articulatory parallel data to the UASpeech acoustic data is problematic, as was shown in the previous research on cross-domain audio-visual inversion [12]. To this end, the large

---

[1]Compared with using equal weights, alternative weighting by removing the CE cost, increasing/decreasing the CE cost weight or removing one of the other three costs all led to WER increase in the resulting AASR systems.

acoustic domain mismatch between the two data sets can be minimized using multi-level adaptive networks (MLAN) [12, 39, 43] before A2A inversion can be performed. An example MLAN model is shown in the left part of Figure 2 (circled in red dotted line).
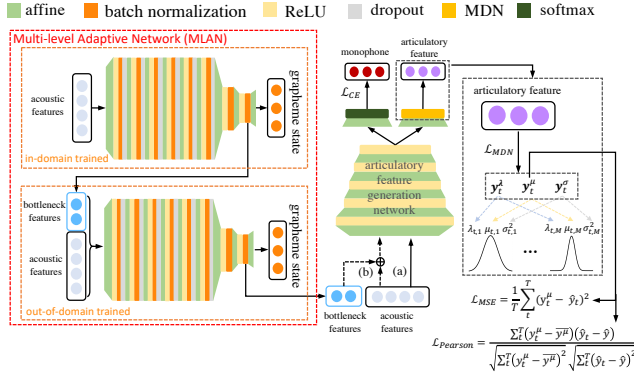

Fig. 2. Cross-domain articulatory feature generation system architecture. The left part shows the MLAN cross domain feature adaptation network, while on the right is the multi-task trained MDN based A2A inversion model. The inputs fed into the A2A inversion model component are based on either: a) splicing windowed frames of original acoustic features using connection (a) for in-domain A2A inversion performed on TORGO acoustic data w/o data augmentation applied; or b) the MLAN adapted UASpeech bottleneck features for cross-domain A2A inversion of the UASpeech acoustic data.

An example MLAN network consisting of two DNN components is shown in the left portion of Figure 2. Each component DNN contains a bottleneck layer positioned immediately before the output layer. The MLAN training process includes the following steps: 1) the first-level DNN was trained with the audio data from the in-domain UASpeech corpus; 2) the resulting in-domain dysarthric speech trained DNN was then used to produce bottleneck features for the out-of-domain data of the TORGO audio; 3) the second-level DNN was trained using the out-of-domain TORGO audio data concatenated with the bottleneck features computed from the previous step. When feedforwarding the UASpeech data into the resulting MLAN network, the combined effect produced by these two cascaded component DNNs is such that the final bottleneck features produced at the second-level DNN component will exhibit smaller mismatch against the bottleneck features obtained by feedforwarding the TORGO data into the MLAN network. These cross-domain adapted bottleneck features are used in A2A inversion model training and articulatory feature generation (in the right part of Figure 2 using connection (b)) for the UASpeech audio data.

4. EXPERIMENTS

4.1. Experiments on the TORGO dataset

Task Description and Experimental Setup : The TORGO dataset is a disordered speech corpus with acoustic-articulatory parallel recordings and contains 8 dysarthric and 7 control speakers. In this paper, 13.49-hour audio data was used, in which the total duration of short sentence based utterances is 5.81 hours while that of single word based utterances is 7.68 hours. The total number of speech utterances is 16394. A speaker level data partitioning similar to that used in the UASpeech [33] corpus is used. All 7 control speakers' data together with two thirds of the 8 dysarthric speakers' data were merged into the training set (11.7 hours) while the remaining dysarthric speech served as the test data (1.79 hours). The percentage of spoken content overlap between the training and test data is

50%. Excessive silence at the sentence start and end was removed using a HTK toolkit [42] trained GMM-HMM system. After silence stripping, the training set contains 6.46 hours of data while the test set has 1.02 hours of speech. After a combination of disordered speaker independent and dependent speed perturbation [37] based data augmentation, the total amount of training data was expanded by a factor of 6 times and increased to 34.11 hours in total.

In our experiments, the 7-layer LF-MMI based TDNN-F model as shown in Figure 1 was implemented using the Kaldi toolkit [44] while Conformer based end-to-end systems[2] were implemented using the Espnet toolkit [45]. A 3-frame context window was used in both ASR and AASR hybrid LF-MMI trained TDNN systems. 40-dimension Mel-scale filter banks (FBKs) were used as the input acoustic features, and the articulatory features were represented by the measured articulatory trajectories. 6 trajectory variables (TV) were selected as articulatory features, i.e. tongue tip (TT), tongue middle (TM), tongue back (TB), upper lip (UL), lower lip (LL) and lower incisor (LI). The X, Y and Z coordinates which capture the spatial movement of the measured articulators were used to construct 18-dimension articulatory feature vectors before being concatenated with FBK features. A 3-gram LM trained by all the TORGO transcripts with a vocabulary size of 1578 words was used in decoding.

Table 1. Comparison of the WER results produced by various ASR and AASR systems on the 8 TORGO dysarthric speakers test set. The dysarthric speakers are grouped by their intelligibility levels, i.e. "Severe", "Moderate" and "Mild". "Aug." and "arti." are the abbreviations of augmentation and articulatory respectively. AA fusion includes different AA modality fusion: A) input feature concatenation; AND B) score fusion. † and ‡ denote a statistically significant improvement compared with baseline and augmented ASR systems respectively (Sys. 1 and Sys. 6).

Sys.	Model	arti. source	Data Aug.	LHUC SAT	AA Fusion	WER % Severe	Moderate	Mild	Average
1		✗	✗	✗	✗	16.22	10.31	3.87	11.62
2	TDNN	original	✗	✗	input	15.00	9.59	4.18	10.93
3			✗	✗	score A+AA	13.98	9.39	3.79	10.25†
4		inversion	✗	✗	input	15.61	10.10	3.79	11.24
5			✗	✗	score A+AA	14.67	9.80	3.41	10.59†
6		✗	✓	✗	✗	12.80	8.78	3.64	9.47
7	TDNN	original	✓	✗	input	13.25	7.24	3.02	9.21
8			✓	✗	score A+AA	12.68	7.76	2.86	8.98
9		inversion	✓	✗	input	12.72	7.45	3.41	9.09
10			✓	✗	score A+AA	12.28	7.96	2.86	8.81‡
11	TDNN		✓	✓	✗	12.52	8.27	3.25	9.11
12		inversion	✓	✓	input	12.52	7.14	3.17	8.85
13			✓	✓	score A+AA	12.03	7.35	2.94	8.58
14	Conformer	✗	✓	✗	✗	22.28	7.35	4.72	14.39
15		inversion	✓	✗	input	21.79	7.55	3.79	13.93

Results : The performance[3] of various ASR and acoustic-articulatory feature based AASR systems on the 8 TORGO dysarthric speakers test set is shown in Table 1. Several trends can be found: **1)** the incorporation of the original articulatory features provided in the TORGO data consistently outperform the acoustic only TDNN ASR systems with or without data augmentation (Sys. 2 & 3 vs. Sys. 1; Sys. 7 & 8 vs. Sys. 6); among these, statistically significant WERs reduction of 1.37% absolute (11.79% relative) and 1.03% absolute (8.86% relative) were obtained over the baseline ASR system using the AASR system Sys. 3 (original articulatory features) and Sys. 5 (inverted articulatory features) respectively. The RMSE of the MDN based A2A inversion system for inverted features in Sys. 5 is 0.808mm. **2)** further score fusion between the ASR and

---

[2]8 encoder plus 4 decoder layers, feed-forward layer dim = 1024, attention heads = 4, dim of attention heads = 256, interpolated CTC+AED cost.

[3]A matched pairs sentence-segment word error based statistical significance test was performed at a significance level $\alpha = 0.05$

AASR systems consistently outperforms the AASR systems using the acoustic-articulatory feature concatenation at the input (Sys. 3 vs Sys. 2; Sys. 8 vs. Sys. 7). **3)** The AASR systems trained using A2A inverted features produced performance comparable to those using the original articulatory features (Sys. 9 & 10 vs. Sys. 7 & 8) after data augmentation. **4)** Consistent WER reductions over the baseline ASR systems were obtained using the AASR systems constructed using the A2A inverted articulatory features before and after LHUC-SAT speaker adaptation (Sys. 10 vs. Sys. 6; Sys. 13 vs. Sys. 11). **5)** Similar trends can also be found on the Conformer based systems (Sys. 15 vs. Sys. 14). No score fusion was used on Conformer due to no performance gain compared to TDNN systems.

4.2. Experiments on the UASpeech dataset

Task Description and Experimental Setup : The UASpeech corpus is the largest publicly available disordered speech corpus that is designed as an isolated word recognition task [33] consisting of 16 dysarthric and 13 control speakers. The speech materials contain 155 common words and 300 uncommon words. The entire corpus is further divided into 3 subset blocks per speaker, with each block containing all 155 common words and one third of the uncommon words. The data from Block 1 (B1) and Block 3 (B3) of all the 29 speakers are used as the training set (69.1 hours of audio, 99195 utterances in total), while the data of Block 2 (B2) collected from all the 16 dysarthric speakers serves as the evaluation data set (22.6 hours of audio, 26520 utterances in total). After removing excessive silence at the start and end of speech audio segments, a combined total of 30.6 hours of audio data from Block 1 and 3 were used as the training set, while 9 hours of speech from Block 2 was used for performance evaluation. After speaker independent and dependent speed perturbation based data augmentation [37, 39], the total amount of training data was increased to 130.1 hours in total.

Table 2. WERs of baseline ASR and AASR systems using the cross-domain inverted articulatory features on the UASpeech test set of 16 dysarthric speakers grouped by intelligibility levels: "Very low", "Low", "Mild" and "High". AA fusion includes different AA modality fusion: A) input feature concatenation; B) hidden layer fusion; AND C) score fusion. Optional further incorporation of video modality is used. † denotes statistical significant differences obtained against the baseline ASR systems (Sys. 1, 7, 12 & 14)

Sys.	model	AA Fusion	Data Aug.	LHUC SAT	MLAN	Very Low	Low	Mild	High	Average
1	Hybrid DNN	✗	✗	✗	✗	69.82	32.61	24.53	10.40	31.45
2		7th hidden	✗		✗	69.41	33.01	24.52	10.35	31.46
3					✓	67.82†	31.25†	22.88†	9.77†	30.15†
4	Hybrid DNN	✗	✓	✗	✗	66.84	28.70	20.39	9.37	28.67
5		7th hidden			✗	66.43	30.26	22.13	9.63	29.41
6					✓	66.27	28.42	19.78	9.40	**28.37**
7		✗	✓	✓	✗	64.22	27.87	18.01	7.60	26.85
8		7th hidden			✗	62.40	27.81	18.64	8.69	26.93
9					✓	62.08†	26.23†	17.33	8.47	**26.13†**
10		score A+AA	✓	✓	✓	61.22†	25.51†	16.15†	7.64	**25.26†**
11	+visual	score A+AA	✓	✓	✓	60.14	25.11	15.92	7.46	24.82
12	CTC	✗	✓	✗	✓	78.25	53.18	46.56	34.09	50.79
13		input				74.33†	52.53	47.13	34.49	**50.04†**
14	Con-former	✗	✓		✓	68.22	49.38	47.51	41.77	50.44
15		input				67.42	48.53†	47.20	41.95	**50.05**

Results : Domain adaptation is essential in transferring articulatory features from TORGO corpus to UASpeech dysarthric corpus (Sys. 3 vs. Sys. 2; Sys. 6 vs. Sys. 5; Sys. 9 vs. Sys. 8). After incorporating the corss-domain inverted articulatory features obtained on the UASpeech audio data, the resulting AASR systems shown in Table 2 consistently outperform the comparable baseline ASR systems using acoustic feature only by a significant margin in WER (Sys. 3 vs. Sys. 1; Sys. 6 vs. Sys. 4; Sys. 9 vs. Sys. 7) before and after data augmentation and LHUC-SAT based speaker adaptation applied to a state-of-the-art hybrid DNN system [39], a CTC end-to-end system (Sys. 13 vs. Sys. 12) and a Conformer end-to-end system (Sys. 15

vs. Sys. 14). In particular, significant WER reductions of 1.64%-2.14% were obtained on the "Very low" and "Low" subgroups after data augmentation and LHUC-SAT (Sys. 9 vs. Sys. 7).

Further ablation studies were conducted to investigate alternative forms of acoustic-articulatory modality fusion. Table 2 shows the performance of 130.1-hour augmented training data based LHUC-SAT speaker adapted baseline acoustic feature only ASR system (Sys. 7), and two acoustic-articulatory features based AASR systems constructed using either a) 7-th hidden layer feature fusion (Sys. 9), or b) score interpolation (Sys. 10, equal weighting to ASR and AASR systems) based acoustic-articulatory modality fusion, with an optional further incorporation of video modality (Sys. 11).

Score fusion between the ASR and AASR systems produced a statistically significant WER reduction of 1.59% over the comparable ASR system (Sys. 10 vs. Sys. 7) and 0.87% over the AASR system using hidden layer feature fusion (Sys. 10 vs. Sys. 9), although with increased system complexity. The lowest WER of 24.82% was obtained by further incorporating visual features [39] (Sys. 11). To the best of our knowledge, this is the lowest WER published so far on the UASpeech test set of 16 dysarthric speakers reported in the literature. Performance of this system against previously published systems on the same task are shown in Table 3.

Table 3. A comparison between published systems on UASpeech and our system. "DA" stands for data augmentation.

systems	WER%
Sheffield-2013 Cross domain augmentation [8]	37.50
Sheffield-2015 Speaker adaptive training [9]	34.80
CUHK-2018 DNN System Combination [10]	30.60
Sheffield-2020 Fine-tuning CNN-TDNN speaker adaptation [11]	30.76
CUHK-2020 Cross-domain AVSR [12]	26.84
CUHK-2020 DNN + DA + LHUC SAT [37]	26.37
CUHK-2021 DNN + GAN based DA + LHUC SAT [38]	25.89
CUHK-2021 NAS DNN + DA + LHUC SAT + AV fusion [39]	25.21
DNN + DA + LHUC SAT + AAV fusion (ours)	**24.82**

5. CONCLUSION

This paper presents a cross-domain acoustic-to-articulatory (A2A) inversion approach that utilizes small amounts of parallel acoustic-articulatory data of the 15-hour TORGO corpus in model training before being cross-domain adapted to a larger 102.7-hour UASpeech dysarthric corpus to produce articulatory features for ASR system construction incorporating articulatory features. Experimental results on both tasks suggest that incorporating the A2A generated articulatory features consistently outperformed the baseline hybrid DNN/TDNN, CTC and Conformer based end-to-end systems constructed using acoustic features only, while producing the lowest published WER of 24.82% on the 16 dysarthric speakers of the benchmark UASpeech task. The proposed cross-domain A2A inversion method allows a more practical and wider use of articulatory features in ASR systems targeting disordered speech.

6. ACKNOWLEDGEMENT

7. REFERENCES

[1] Edwin Maas et al., "Principles of motor learning in treatment of motor speech disorders," *ASHA*, 2008.

[2] Tara L Whitehill et al., "Speech errors in Cantonese speaking adults with cerebral palsy," *Clinical linguistics & phonetics*, 2000.

[3] Tanja Makkonen et al., "Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms," *IJLCD*, 2018.

[4] Sophie Scott et al., "Evidence for an apparent sensory speech disorder in Parkinson's disease.," *JNNP*, 1984.

[5] Peter Jerntorp and Goran Berglund, "Stroke registry in Malmö, Sweden.," *STROKE*, 1992.

[6] Thomas J Hixon et al., "Restricted motility of the speech articulators in cerebral palsy," *JSHD*, 1964.

[7] Victoria Young et al., "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *ASSIST TECHNOL*, 2010.

[8] Heidi Christensen et al., "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech.," in *INTERSPEECH*, 2013.

[9] Siddharth Sehgal et al., "Model adaptation and adaptive training for the recognition of dysarthric speech," in *SLPAT*, 2015.

[10] Jianwei Yu et al., "Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus.," in *INTERSPEECH*, 2018.

[11] Feifei Xiong et al., "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *ICASSP*. IEEE, 2020.

[12] Shansong Liu et al., "Exploiting Cross-Domain Visual Feature Generation for Disordered Speech Recognition.," in *INTERSPEECH*, 2020.

[13] Mengzhe Geng et al., "Spectro-Temporal Deep Features for Disordered Speech Assessment and Recognition," in *INTERSPEECH*, 2021.

[14] Igor Zlokarnik, "Adding articulatory features to acoustic features for automatic speech recognition," *JASA*, 1995.

[15] Alan Wrench et al., "Continuous speech recognition using articulatory data," in *ICSLP*, 2000.

[16] Katrin Kirchhoff et al., "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, 2002.

[17] Prasanta Kumar Ghosh et al., "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *JASA*, 2011.

[18] Vikramjit Mitra et al., "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, 2017.

[19] Vikramjit Mitra et al., "Articulatory information for noise robust speech recognition," *IEEE TASLP*, 2010.

[20] Yunbin Deng et al., "Disordered speech recognition using acoustic and sEMG signals," in *INTERSPEECH*, 2009.

[21] Frank Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE TASLP*, 2010.

[22] Jose A Gonzalez et al., "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE TASLP*, 2017.

[23] Feifei Xiong et al., "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *ITG-Symposium*, 2018.

[24] Emre Yılmaz et al., "Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech," *Computer Speech & Language*, 2019.

[25] Sarthak Kumar Maharana et al., "Acoustic-to-Articulatory Inversion for Dysarthric Speech by Using Cross-Corpus Acoustic-Articulatory Data," in *ICASSP*. IEEE, 2021.

[26] Olov Engwall, "Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA," in *ICSLP*, 2000.

[27] Shrikanth Narayanan et al., "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *JASA*, 2014.

[28] George Papcun et al., "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *JASA*, 1992.

[29] Prasanta Kumar Ghosh et al., "A generalized smoothness criterion for acoustic-to-articulatory inversion," *JASA*, 2010.

[30] Benigno Uria et al., "Deep architectures for articulatory inversion," in *INTERSPEECH*, 2012.

[31] Xurong Xie et al., "Investigation of stacked deep neural networks and mixture density networks for acoustic-to-articulatory inversion," in *ISCSLP*. IEEE, 2018.

[32] Frank Rudzicz et al., "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, 2012.

[33] Heejin Kim et al., "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.

[34] Vijayaditya Peddinti et al., "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[35] Alex Graves et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[36] Anmol Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH*, 2020.

[37] Mengzhe Geng et al., "Investigation of Data Augmentation Techniques for Disordered Speech Recognition.," in *INTERSPEECH*, 2020.

[38] Zengrui Jin et al., "Adversarial Data Augmentation for Disordered Speech Recognition," in *INTERSPEECH*, 2021.

[39] Shansong Liu et al., "Recent Progress in the CUHK Dysarthric Speech Recognition System," *IEEE TASLP*, 2021.

[40] Pawel Swietojanski et al., "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE TASLP*, 2016.

[41] Keiichi Tokuda et al., "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*. IEEE, 2000.

[42] Steve Young et al., "The HTK book," *Cambridge university engineering department*, 2002.

[43] Peter Bell et al., "Transcription of multi-genre media archives using out-of-domain data," in *SLT*. IEEE, 2012, pp. 324–329.

[44] Daniel Povey et al., "The Kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.

[45] Shinji Watanabe et al., "Espnet: End-to-end speech processing toolkit," *arXiv:1804.00015*, 2018.