ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413625

ACOUSTIC-TO-ARTICULATORY INVERSION FOR DYSARTHRIC SPEECH BY USING CROSS-CORPUS ACOUSTIC-ARTICULATORY DATA

Sarthak Kumar Maharana[1], *Aravind Illa*[1], *Renuka Mannem*[1], *Yamini Belur*[2], *Preetie Shetty*[2], *Veeramani Preethish Kumar*[2], *Seena Vengalil*[2], *Kiran Polavarapu*[2], *Nalini Atchayaram*[2], *Prasanta Kumar Ghosh*[1]

[1]Department of Electrical Engineering, Indian Institute of Science, Bengaluru 560012, India
[2]Department of Speech Pathology and Audiology, NIMHANS, Bengaluru 560029, India

ABSTRACT

In this work, we focus on estimating articulatory movements from acoustic features, known as acoustic-to-articulatory inversion (AAI), for dysarthric patients with amyotrophic lateral sclerosis (ALS). Unlike healthy subjects, there are two potential challenges involved in AAI on dysarthric speech. Due to speech impairment, the pronunciation of dysarthric patients is unclear and inaccurate, which could impact the AAI performance. In addition, acoustic-articulatory data from dysarthric patients is limited due to the difficulty in the recording. These challenges motivate us to utilize cross-corpus acoustic-articulatory data. In this study, we propose an AAI model by conditioning speaker information using x-vectors at the input, and multi-target articulatory trajectory outputs for each corpus separately. Results reveal that the proposed AAI model shows relative improvements of the Pearson correlation coefficient (CC) by $\sim$13.16% and $\sim$16.45% over a randomly initialized baseline AAI model trained with only dysarthric corpus in seen and unseen conditions, respectively. In the seen conditions, the proposed AAI model outperforms the three baseline AAI models, that utilize the cross-corpus, by $\sim$3.49%, $\sim$6.46%, and $\sim$4.03% in terms of CC.

Index Terms— Amyotrophic lateral sclerosis, acoustic-to-articulatory inversion, transfer learning, x-vectors, BLSTM.

1. INTRODUCTION

Amyotrophic Lateral Sclerosis, abbreviated as ALS, is a nervous system disease with a progressive increase in severity. The disease affects the brain and the spinal cord, with an eventual decline in the electrical signals sent by the brain. This slows down the muscular responses [1]. The major effects of ALS also include an inability to speak, lifting of hands, and doing other basic motor actions. Since a subject suffering from ALS finds it difficult to speak, this leads to a poor pronunciation alongside a mumbling and an unintelligible speech, with the condition being termed as dysarthria [2]. Dysarthria adversely affects the articulators [3] including the lips, jaw, tongue, and velum, particularly with an increase in the dysarthria severity level [3, 4]. The symptoms of the disease have no impact on the comprehension and intellectual aspects of the natural language of a patient [5].

Dysarthria comes with wide variability in the impediment of articulation, which varies from patient to patient. Conventionally, to informally assess the degradation of articulation, speech-language pathologists (SLPs) resort to different speech stimuli viz. reading a passage, reading a word, spontaneous speech or rehearsed speech [6]. To make a critical assessment of the poor articulation of a dysarthric patient, it is important to analyze the real-time articulatory movements, collected using Electromagnetic Articulography (EMA). The presence of sensor coils and wires prohibits a long recording session for each patient, typically leading to a small amount of articulatory recordings. For example, we, for this work, could collect roughly 2 to 5 minutes of acoustic-articulatory data per patient. The scarcity of data motivates us to develop techniques for estimating articulatory movements from acoustic recordings, or acoustic-to-articulatory inversion (AAI) [7], for dysarthric patients in such a low-resource condition.

Deep learning approaches, especially Bidirectional Long-Short Term Memory (BLSTM) as recurrent neural networks, have been shown to achieve the state-of-the-art performance for AAI in low-resource data conditions [8]. However, it demands a significant amount of acoustic-articulatory data from a subject. As the amount of data from ALS patients is comparatively less, to train such an AAI model, in this work, we conduct experiments to study the usage of a cross-corpus acoustic-articulatory data, comprising data of healthy control subjects only, for training an AAI model for dysarthric speech by deploying transfer learning and joint-training techniques. The objectives of our work are to : 1) study the AAI model's performance on the dysarthric speech when the model is trained in a corpus dependent manner using a matched low-resource dysarthric corpus or using a mismatched cross-corpus with rich acoustic-articulatory data, 2) investigate the benefit of utilizing cross-corpus data using transfer learning techniques and by joint-training for the articulatory predictions of healthy control and dysarthric subjects, 3) assess the variations in the performance of AAI for different speech stimuli used for informal assessment by SLPs, 4) do articulatory specific analysis on AAI performance and analysis of the frequency characteristics of the ground truth and predicted articulatory trajectories for healthy controls and patients. The findings from this investigation could benefit pathological speech assessment applications like, developing speech-based assistive tools for SLPs to monitor the decline in articulatory movements directly from the speech acoustics of dysarthric patients.

2. DATASETS

Experiments were carried out with acoustic-articulatory data which was recorded using Electromagnetic Articulograph (EMA), AG501 [9]. Using EMA recording setup, synchronous speech acoustics and articulatory movements were captured. Speech was recorded using t.bone EM9600 shotgun [10], unidirectional electret condenser microphone at 44.1 kHz sampling rate. Articulatory movements were captured using EMA sensors at a 250 Hz sampling rate. We recorded the movements of six articulators, namely, upper lip (UL), lower lip (LL), jaw (JAW), tongue tip (TT), tongue body (TB), and tongue dorsum (TD). For head movement correction, two sensors were attached behind the ears [11]. The sensors on the articulators were glued, following the guidelines provided in [12]. Articulatory movements were considered in the horizontal (X) and vertical (Y) directions in the midsagittal plane from 4 articulators for this work. TB and TD sensors were ignored since these measurements were not available for all patients (to minimize discomfort we avoided connecting these sensors to several patients). This results in 8-dimensional articulatory features, which are indicated by UL_x, UL_y, LL_x, LL_y, JAW_x, JAW_y, TT_x, TT_y. The acoustic-articulatory data used in this work was collected at two places with different stimuli, languages, and age groups. The details are provided in the following subsections.

2.1. Cross-corpus

The cross-corpus acoustic-articulatory data were recorded with subjects speaking 460 phonetically balanced English sentences from the MOCHA-TIMIT [13]. The corpus includes 38 healthy control subjects, which consists of 21 males and 17 females with an age range between 20-28 years. The data collection was done at SPIRE Lab, Indian Institute of Science (IISc), Bengaluru, India. The subjects were students and research scholars at IISc with proficiency in English. None of the subjects were reported to have any speaking disorders. After removing silences, the total duration of acoustic-articulatory data was $\sim$11.4 hours, with an average duration of $\sim$18 minutes/subject.

2.2. Dysarthric corpus

The recording session for the collection of the acoustic-articulatory data comprising the dysarthric group was conducted at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India. The subjects were requested to sign a consent form before the recording sessions could start. The recording was approved by the ethics committee of NIMHANS. We also collected acoustic-articulatory data from a group of healthy control subjects, to compare with the dysarthric articulation. The healthy controls and patients were identical with respect to age. We assured that the speech stimuli were the same for the two groups. Data of 7 healthy controls (4 males and 3 females) and 13 ALS patients (7 males and 6 females) constitute this corpus, with Kannada, an Indian language, being their first native language. The healthy control subjects reported to have no speech disorders in the past.

To informally examine the variability and waning of articulation of ALS patients speaking Kannada, Babu et al. [14] suggested an articulation test. We accept this test as a base, and employ three standardized speech stimuli, which include reading a Kannada passage (T1), rehearsed speech (T2), and spontaneous speech (T3), with each stimulus being performed by the subject for 2-3 times. T1 dealt with subjects reading out a passage about history, with the assistance of a clinician. The main objective of T2 was to request the subjects to speak, "*Nanna hesaru XYZ. Nanu iga Bengaluru nalli iddene. (My name is XYZ. I am now in Bengaluru)*", with XYZ being the first name of the subject. In T3, subjects were given the liberty to speak whatever they wished to, with the majority of them talking about their home, work, and lifestyle. This task provides an insight into the natural way of speaking, and the articulation, by a subject. Silences from all collected speech recordings were manually removed.

Table 1 provides the speech stimulus-specific utterance duration for healthy controls and patients used in this work. Column "Duration Range" reports the range of utterance duration in seconds, stimulus wise. Column "Total Duration (Duration/Subject)" indicates the total duration of speech utterances alongside the duration per subject. Note that not all the subjects were able to perform all the tasks that have been mentioned above, due to discomfort caused by the sensors. In summary, the average total duration of data collected is 3.3 minutes/patient, and 3.74 minutes/healthy control.

Tasks	Healthy Controls		Patients	
	Duration Range	Total Duration (Duration/Subject)	Duration Range	Total Duration (Duration/Subject)
T1	0.524-39.69	824.23 (117.7)	0.427-34.3	790.92 (60.84)
T2	0.545-8.713	158.37 (22.62)	0.586-24.2	566.60 (43.58)
T3	0.538-62.15	588.44 (84.06)	0.468-38.46	1245.07 (95.77)

Table 1. Speech stimulus-specific duration (in seconds) for the 7 healthy controls and the 13 ALS patients constituting the dysarthric corpus.

3. PROPOSED APPROACH

Acoustic-to-articulatory inversion (AAI) is a regression problem. The relationship between the input speech acoustics and the output articulatory movements is known to be complex and non-linear [15, 16]. Neural networks have been shown to perform well in learning complex non-linear functions. In [8, 17], it has been shown that BLSTM networks provide the state-of-the-art performance for AAI. Hence, in this study, we deploy a BLSTM network with three hidden layers and time-distributed regression layers for AAI. A BLSTM network demands a large amount of acoustic-articulatory data due to its network complexity. Thus, the available acoustic-articulatory data from the dysarthric patients might not be sufficient. Hence, we utilize the cross-corpus data using transfer learning and joint-training techniques.

Several techniques have been proposed in the literature to utilize multi-corpus, multi-speaker, and low-resource acoustic-articulatory data. To overcome the limitation on the amount of acoustic-articulatory data from a target subject, a low-resource AAI model [8] was proposed using transfer learning, where a generic background AAI model (GBM AAI) was trained by pooling the data from all speakers, and then the GBM AAI was fine-tuned (GBM-FT) on the low-resource data from the same corpus. An alternative approach to GBM-FT was to provide speaker-specific information as auxiliary features along with the acoustic features for learning the rich acoustic-to-articulatory mappings of multiple speakers, known as speaker conditioned AAI [18, 19]. As auxiliary features, one-hot representation [18] of train subjects or x-vector embeddings [19] were utilized. On the other hand, across the corpora, the articulatory measurements differed to a large extent due to the differences in the placement of sensors and the measurement devices, which limit most of the previous works on AAI to a single corpus. In [20], a multi-task learning AAI was trained using three different articulatory corpora (three sets of target output, one for each corpus).

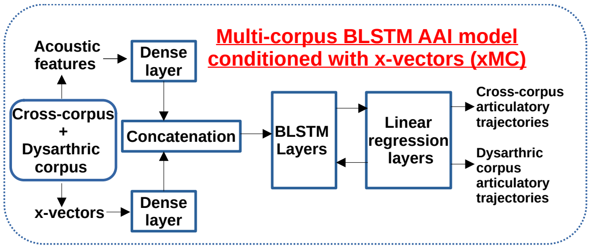
Multi-corpus BLSTM AAI model conditioned with x-vectors (xMC)
Acoustic features
Dense layer
Cross-corpus + Dysarthric corpus
Concatenation
BLSTM Layers
Linear regression layers
Cross-corpus articulatory trajectories
Dysarthric corpus articulatory trajectories
x-vectors
Dense layer

Fig. 1. Block diagram of the proposed multi-corpus BLSTM AAI model conditioned with x-vectors (xMC).

In this work, we propose an AAI model by combining speaker conditioning [19] and multi-task learning [20] approaches to utilize multi-speaker and multi-corpus data. The proposed speaker conditioned multi-corpus AAI model is shown in Fig. 1. The multi-corpus AAI model is conditioned with x-vectors [21] along with the acoustic features, which is denoted as xMC. The extracted x-vector is replicated for every frame of an utterance, which equals the total number of acoustic feature vectors. The x-vectors and the acoustic features are fed as inputs to separate dense layers and then concatenated. The concatenated vector is fed to the BLSTM layers with the output of the last layer being fed to two time-distributed linear regression layers, where, the first eight dimensions correspond to the articulatory trajectories of the cross-corpus, and the remaining eight correspond to that of the dysarthric corpus.

4. EXPERIMENTAL SETUP

Pre-processing and extraction of features: The recorded acoustic-articulatory data is pre-processed for both the corpora considered in this work. The speech waveforms are down-sampled from 44.1 kHz to 16 kHz. We use 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) as acoustic features, which have been shown to be optimal for AAI [22, 23]. MFCCs are computed with a window length of 20 ms and a shift of 10 ms using the Kaldi toolkit [24]. On the other hand, the articulatory trajectories are low-pass filtered with a 25 Hz cutoff frequency to avoid high-frequency noise incurred due to measurement error of EMA. The articulatory data is down-sampled from 250 Hz to 100 Hz to obtain a one-to-one correspondence with MFCC vectors. We further perform utterance level mean and variance normalization across each dimension for both acoustic and articulatory features. The Kaldi toolkit [24] is used to compute the x-vector using a pre-trained model trained on the VoxColeb database [25]. For the experiments, we consider a 5-fold cross-validation setup in both seen and unseen subject conditions. The healthy controls and patients from the dysarthric corpus are equally distributed across the 5 folds. In the unseen condition, 16 subjects from four groups are considered for training, and the remaining 4 subjects are considered for testing in a round-robin fashion. In the seen condition, the dysarthric corpus is split into 10 groups covering each stimulus uniformly. Each fold consists of 8 groups for training and the remaining 2 groups for testing in a round-robin fashion. From the cross-corpus, we consider all the 38 subjects for training.

Baseline AAI models: As a baseline, we train an AAI model using only the dysarthric corpus with randomly initialized weights, indicated as RI AAI. To analyze the benefits of transfer learning, we consider the GBM-FT [8], multi-corpus (MC) [20], and xSC [19] as baseline AAI models to compare the performance of the proposed xMC AAI model.

Training and network parameters: For GBM AAI training, the cross-corpus data, with 38 subjects, is divided into the train (34 subjects) and validation sets (4 subjects). We perform training of the xSC network, as mentioned in [19], by pooling data from both cross-corpus and dysarthric corpus. Experiments with the multi-corpus AAI model (MC) are conducted with 3 BLSTM layers (256 BLSTM units), and two output layers (separately for cross and dysarthric corpus), which are time-distributed linear regression layers (8-dimensions). For the evaluation results of MC and xMC, the articulatory trajectories obtained corresponding to the dysarthric corpus, are considered. For training the RI, GBM-FT, and xSC models, we use the mean-squared error (MSE) as the loss function. For the MC and xMC models, we use a custom loss function for training, where the MSE loss computed at the output of the corresponding corpus (to which the utterance belongs) backpropagates the error to update the weights while masking the other corpus' output estimates. Network parameters are optimized using Adam optimizer for all the experiments. We perform early stopping based on the validation loss. Experiments are done with Keras [26], with Tensorflow [27] as the backend.

Number of BLSTM nodes	RI AAI				GBM AAI	
	Seen		Unseen		Healthy Controls	Patients
	Healthy Controls	Patients	Healthy Controls	Patients		
32	0.42 (0.07)	0.48 (0.06)	0.40 (0.07)	0.42 (0.07)	0.49 (0.09)	0.50 (0.09)
64	0.42 (0.08)	0.49 (0.06)	0.40 (0.08)	0.43 (0.07)	0.50 (0.09)	0.51 (0.08)
128	0.45 (0.07)	0.52 (0.06)	0.42 (0.08)	0.46 (0.07)	0.50 (0.08)	0.51 (0.08)
256	0.43 (0.08)	0.52 (0.06)	0.42 (0.08)	0.46 (0.08)	0.50 (0.1)	0.50 (0.09)

Table 2. Average (standard deviation) of the CC values for the GBM AAI and RI AAI models on the dysarthric corpus in seen and unseen conditions.

Performance metric: To evaluate the performances of different AAI models, we use the Pearson Correlation Coefficient (CC) [8, 22] between the ground-truth and its corresponding predicted articulatory trajectories.

5. RESULTS AND DISCUSSION

5.1. Evaluation of corpus dependent AAI models:

Table 2 reports the average (standard deviation) CC values on the dysarthric data (test set) evaluated on the corpus dependent RI and GBM AAI models. Unlike the RI AAI model, the GBM AAI model does not have seen and unseen subject conditions, as the dysarthric data is completely unseen. Thus, for a fair comparison, we compare the GBM and RI AAI models' performance in unseen subject conditions. It is observed that the GBM AAI model (with 256 BLSTM nodes) performs better than the RI AAI model (with 256 BLSTM nodes) with a relative improvement of $\sim$19.05% and $\sim$8.7% for healthy controls and patients, respectively. As the training data for RI ($\sim$1.16 hrs) is less than that for GBM ($\sim$11.4 hrs), there is a chance of over-fitting of the RI AAI model as both model networks have equal complexity. To investigate this, we experiment with different numbers of BLSTM nodes (hidden units) for the RI and GBM AAI models. The performance of the RI AAI model increases as the number of BLSTM nodes increases from 32 to 128 and saturates at 256, which indicates that the RI AAI model does not overfit to the training data. It is observed that the GBM AAI model performs better than the RI AAI model consistently for all choices of the number of BLSTM nodes in unseen subject conditions. Thus, the GBM AAI model which utilizes the cross-corpus data performs better than the RI AAI model irrespective of the mismatch between the train and test data related to language, speech stimuli, and age. Hence, using the rich acoustic-articulatory data from a cross-corpus for training helps in accurate articulatory movements prediction compared to using limited training data in RI AAI model. In the next subsection, we present the results by utilizing acoustic-articulatory data from both the corpora.

5.2. Evaluation of AAI models utilizing cross-corpus:

Table 3 reports the average CC (standard deviation) obtained on the dysarthric data (test set) using the different AAI models experimented in this work. Compared to the RI model, all the AAI models utilizing cross-corpus show a significant improvement in performance in both seen and unseen conditions. Thus, experimental results reveal that utilizing rich acoustic-articulatory data, from the cross-corpus, to learn an AAI model is beneficial for low-resource dysarthric corpus. In seen conditions, for patients, the proposed xMC AAI model outperforms the RI, GBM-FT, MC, and xSC AAI models with relative improvements of $\sim 13.16\%$, $\sim 3.49\%$, $\sim 6.46\%$, and $\sim 4.03\%$, respectively. For healthy controls,

Speech Stimuli	Seen										Unseen									
	RI AAI		GBM-FT AAI		MC AAI		xSC AAI		xMC AAI		RI AAI		GBM-FT AAI		MC AAI		xSC AAI		xMC AAI	
	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients
T1	0.472 (0.09)	0.506 (0.09)	0.551 (0.09)	0.572 (0.08)	0.534 (0.11)	0.548 (0.1)	0.541 (0.11)	0.574 (0.08)	0.569 (0.1)	0.588 (0.1)	0.460 (0.11)	0.448 (0.09)	0.535 (0.1)	0.537 (0.1)	0.529 (0.1)	0.512 (0.08)	0.527 (0.12)	0.531 (0.09)	0.53 (0.01)	0.527 (0.09)
T2	0.480 (0.09)	0.534 (0.06)	0.557 (0.09)	0.591 (0.06)	0.556 (0.09)	0.557 (0.07)	0.575 (0.09)	0.583 (0.08)	0.588 (0.09)	0.602 (0.06)	0.444 (0.09)	0.452 (0.09)	0.534 (0.12)	0.523 (0.07)	0.553 (0.09)	0.525 (0.07)	0.543 (0.11)	0.545 (0.11)	0.537 (0.12)	0.543 (0.08)
T3	0.389 (0.72)	0.528 (0.05)	0.462 (0.08)	0.564 (0.06)	0.476 (0.08)	0.56 (0.07)	0.488 (0.09)	0.563 (0.07)	0.488 (0.08)	0.59 (0.06)	0.386 (0.09)	0.473 (0.08)	0.464 (0.08)	0.517 (0.07)	0.457 (0.09)	0.529 (0.07)	0.469 (0.09)	0.533 (0.08)	0.461 (0.09)	0.542 (0.07)
Avg (Std Dev)	0.438 (0.08)	0.524 (0.06)	0.514 (0.08)	0.573 (0.06)	0.513 (0.09)	0.557 (0.07)	0.525 (0.09)	0.57 (0.07)	**0.538** **(0.08)**	**0.593** **(0.07)**	0.424 (0.09)	0.462 (0.08)	0.504 (0.09)	0.522 (0.07)	0.503 (0.09)	0.523 (0.07)	0.505 (0.1)	0.535 (0.08)	0.502 (0.09)	**0.538** **(0.07)**

Table 3. Average (standard deviation) of the CC values for the AAI models using the cross-corpus, on the dysarthric data in seen and unseen subject conditions.

in the seen conditions, the xMC AAI model performs the best with a maximum relative improvement of $\sim$22.83% over the RI AAI model. In unseen conditions, for healthy controls, GBM-FT, MC, xSC, and xMC AAI models have similar performances, while for the patients, xSC and xMC show relative improvements of $\sim$2.49% and $\sim$2.86% when compared with GBM-FT and MC, respectively. This could be due to conditioning with x-vectors at the input, leading to a better generalization to unseen speakers [19]. Comparing the performances of the AAI models for different speech stimuli (T1, T2, and T3), it is observed that the CC value for task T3 is lower than that of T1 and T2 tasks for healthy control subjects. Unlike T1 and T2, in the case of T3, a subject can speak at his/her own pace and the speech content can vary from subject to subject. This brings higher variability within T3, and also contrasts with T1, T2, and cross-corpus which are recorded while the subjects read sentences/passage. These factors could lead to a performance drop in T3 compared to T1 and T2.

5.3. Articulatory specific analysis:

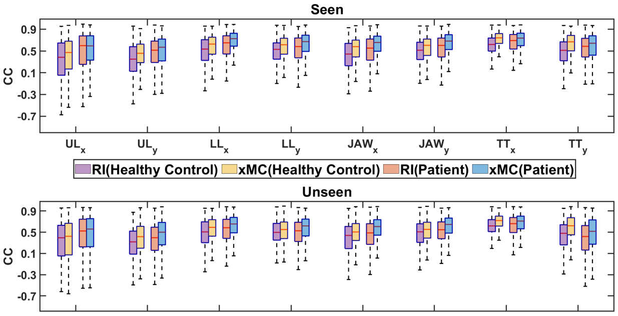

Fig. 2. CC values separately for each articulator predicted using RI and xMC AAI models in seen and unseen conditions for healthy controls and patients.

We compare the articulatory specific performances of the xMC AAI model and RI AAI model. Fig. 2 illustrates the CC values for the articulatory trajectories predicted from the RI and xMC AAI models in seen and unseen conditions for healthy controls and patients. It is observed that the xMC AAI model performs better than the RI AAI model for all the articulators in both seen and unseen conditions. For patients, in the seen conditions, we observe a maximum relative improvement for JAW_x (21.18%) and LL_x (17.69%) articulators, whereas in the unseen conditions, TT_y (32.66%) and JAW_x (26.25%) show maximum relative improvement.

Table 4 reports the cut-off frequencies (f_c) corresponding to 98% of the energy of the original and predicted articulatory trajectories (Arti Traj) from RI and xMC in seen and unseen conditions for healthy controls and patients. From the f_c values of original trajectories, we observe a drop in f_c values in the case of patients compared to the healthy controls for all the articulators. A similar trend is observed for the predicted trajectories from the xMC AAI model in both seen and unseen conditions. However, in the case of the RI AAI model, the f_c values are found to be less compared to the original articulatory trajectories for both patients and healthy controls which could be due to the less variability in the dynamics of articulatory trajectories predicted from RI which leads to low-frequency characteristics.

Arti Traj	Original		Seen				Unseen			
			xMC AAI		RI AAI		xMC AAI		RI AAI	
	HC	P	HC	P	HC	P	HC	P	HC	P
UL_x	11.51	9.24	11.66	10.68	7.56	6.41	11.93	10.45	6.41	5.68
UL_y	9.76	8.88	13.59	12.36	8.61	7.87	13.46	11.83	7.72	7.29
LL_x	8.64	7.83	9.51	8.00	7.94	6.43	9.32	7.72	6.72	5.80
LL_y	9.42	8.61	10.38	8.65	8.50	7.03	10.12	8.02	7.42	6.37
JAW_x	8.86	8.80	9.90	8.38	8.85	7.08	9.84	7.86	7.40	6.19
JAW_y	8.87	8.47	10.07	8.29	8.79	7.01	9.72	7.83	7.35	6.21
TT_x	9.11	8.17	9.85	8.86	8.08	6.77	9.72	8.38	6.63	6.28
TT_y	9.30	8.50	9.86	9.71	7.69	7.00	9.73	9.24	7.11	6.42

Table 4. Cut-off frequencies (f_c) in Hz corresponding to 98% of the energy of the original and predicted articulatory trajectories from xMC and RI, in seen and unseen conditions for healthy controls (HC) and patients (P).

The low f_c values for the original articulatory trajectories of the dysarthric speech could be due to a decline in the speaking rate for dysarthric patients [28]. Also, due to speech impairment, there could be unclear pronunciation and lack of variability in the speech which leads to a reduction in acoustic and articulatory space. However, to understand how these factors could influence the performance of AAI needs further investigation.

6. CONCLUSIONS

In this work, we performed experiments with different AAI models on dysarthric speech, by utilizing the cross-corpus data. Experimental results revealed that the cross-corpus acoustic-articulatory data was beneficial to learn AAI for dysarthric patients even though both corpora were different in terms of age group, language, and speech stimuli among subjects. We further studied the benefit of conditioning the multi-corpus model with x-vectors. Experimental results revealed that the xMC AAI model performed better than the baseline RI for dysarthric patients and healthy controls, in seen and unseen conditions. Further, for dysarthric patients, we showed that the xMC AAI model performed better than or on par with baseline AAI models which utilize cross-corpus data. Our future work includes investigating the role of mismatch in language and speaking style (reading vs spontaneous) of cross-corpus on the performance of different AAI models.

7. REFERENCES

[1] Matthew C Kiernan, Steve Vucic, Benjamin C Cheah, Martin R Turner, Andrew Eisen, Orla Hardiman, James R Burrell, and Margaret C Zoing, "Amyotrophic lateral sclerosis," *The lancet*, vol. 377, no. 9769, pp. 942–955, 2011.

[2] Susan E Langmore and Mark E Lehman, "Physiologic deficits in the orofacial system underlying dysarthria in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 1, pp. 28–37, 1994.

[3] Aravind Illa, Deep Patel, BK Yamini, SS Meera, Shivashankar N, Preethish-Kumar Veeramani, Seena Vengalil, Kiran Polavarapu, Saraswati Nashi, Atchayaram Nalini, and Prasanta Kumar Ghosh, "Comparison of speech tasks for automatic classification of patients with amyotrophic lateral sclerosis and healthy subjects," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6014–6018.

[4] Raymond D Kent, Robert L Sufit, John C Rosenbek, Jane F Kent, Gary Weismer, Ruth E Martin, and Benjamin R Brooks, "Speech deterioration in amyotrophic lateral sclerosis: A case study," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 6, pp. 1269–1275, 1991.

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

[6] Cyndi Stein-Rubin and Renee Fabus, *A guide to clinical assessment and professional report writing in speech-language pathology*, Nelson Education, 2011.

[7] Korin Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, University of Edinburgh, 2002.

[8] Aravind Illa and Prasanta Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory.," in *Interspeech*, 2018, pp. 3122–3126.

[9] "3d electromagnetic articulograph," available online: http://www.articulograph.de/, last accessed: 07/04/2020.

[10] "EM 9600 shotgun microphone," available online: http://www.articulograph.de/, last accessed: 07/04/2020.

[11] Christian Kroos, "Using sensor orientation information for computational head stabilisation in 3D electromagnetic articulography (EMA)," in *Interspeech*, 2009, pp. 776–779.

[12] Ashok Kumar Pattem, Aravind Illa, Amber Afshan, and Prasanta Kumar Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *Computer speech & language*, vol. 47, pp. 157–174, 2018.

[13] A. Wrench, "MOCHA-TIMIT," Speech database, Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, 1999 .[Online].Available: http://sls.qmuc.ac.uk.

[14] RM Babu, N Ratna, and R Bettagiri, "Test of articulation in Kannada," *The JAIISH*, vol. 3, pp. 7–19, 1972.

[15] Katrin Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[16] Zhiyong Wu, Kai Zhao, Xixin Wu, Xinyu Lan, and Helen Meng, "Acoustic to articulatory mapping with deep neural network," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9889–9907, 2015.

[17] Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, Helen Meng, and Lianhong Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4450–4454.

[18] Aravind Illa and Prasanta Kumar Ghosh, "Closed-set speaker conditioned acoustic-to-articulatory inversion using bi-directional long short term memory network," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL171–EL176, 2020.

[19] Aravind Illa and Prasanta Kumar Ghosh, "Speaker conditioned acoustic-to-articulatory inversion using x-vectors," in *Interspeech*, 2020, pp. 1376–1380.

[20] Nadee Seneviratne, Ganesh Sivaraman, and Carol Y Espy-Wilson, "Multi-corpus acoustic-to-articulatory speech inversion.," in *Interspeech*, 2019, pp. 859–863.

[21] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[22] Prasanta Kumar Ghosh and Shrikanth Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[23] Aravind Illa and Prasanta Kumar Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5931–5935.

[24] Daniel Povey et al., "The Kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, 2011.

[25] "Kaldi VoxCeleb pretrained models, available online: https://kaldi-asr.org/models/m7, last accessed:13/06/2020," .

[26] François Chollet, "Keras," 2015, available online: https://github.com/fchollet/keras.

[27] Martín Abadi et al., "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283.

[28] Antje S Mefferd, Gary L Pattee, and Jordan R Green, "Speaking rate effects on articulatory pattern consistency in talkers with mild ALS," *Clinical linguistics & phonetics*, vol. 28, no. 11, pp. 799–811, 2014.