Poster ID: 4934

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], and Prasanta Kumar Ghosh[1]

[1]SPIRE Lab, Department of Electrical Engineering, Indian Institute of Science (IISc), Bengaluru, India

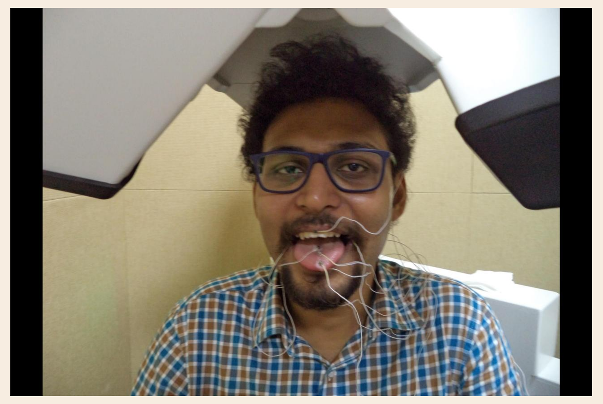
[2]Department of Speech Pathology and Audiology, NIMHANS, Bengaluru, India

Introduction

- **Dysarthria**: Speech disorder causing decline in speech clarity by affecting movements of articulators [1].
- **AAI**: Estimating articulatory movements from acoustic recordings [2].
- **Challenge**: Collecting acoustic-articulatory data, from patients with dysarthria, is tedious. BLSTM networks require a large amount of data to train for AAI [3].
- **Objective**: Perform AAI on dysarthric speech at low-resource conditions, using a rich cross-corpus.

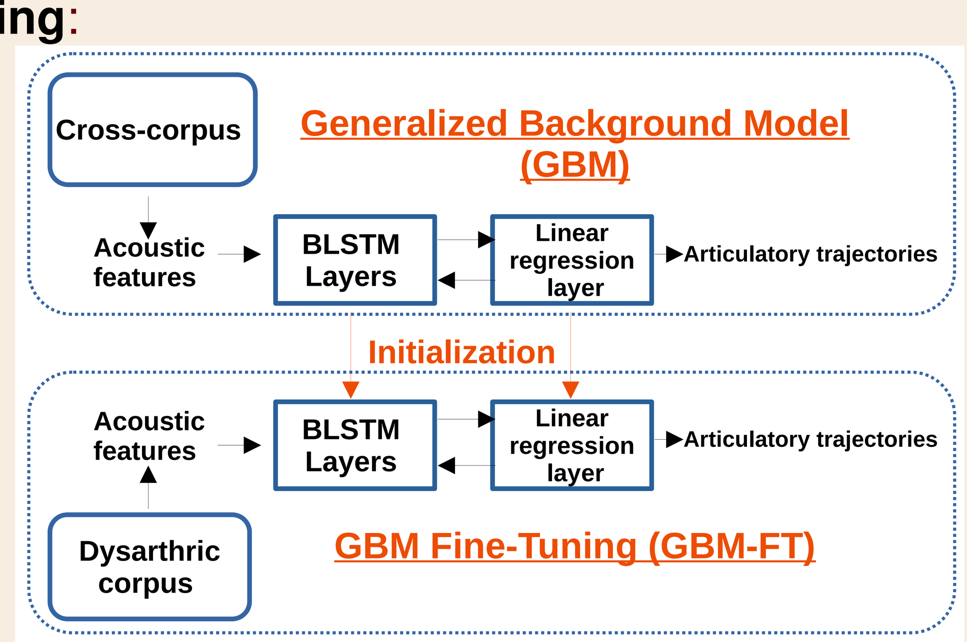
Data

- **Electromagnetic Articulograph (EMA)**: Articulatory movements of four articulators, using EMA AG501 at 100 Hz, are considered.
- **Cross-corpus**: Data from 38 healthy controls; speech stimuli: 460 sentences from the MOCHA-TIMIT; total data: ~11.4 hours.
- **Dysarthric corpus**: Data from 7 healthy controls(HC) and 13 patients(P); speech stimuli: reading a Kannada(Indian language) passage, rehearsed speech, and spontaneous speech; total data: ~1.16 hours.



Proposed Approach

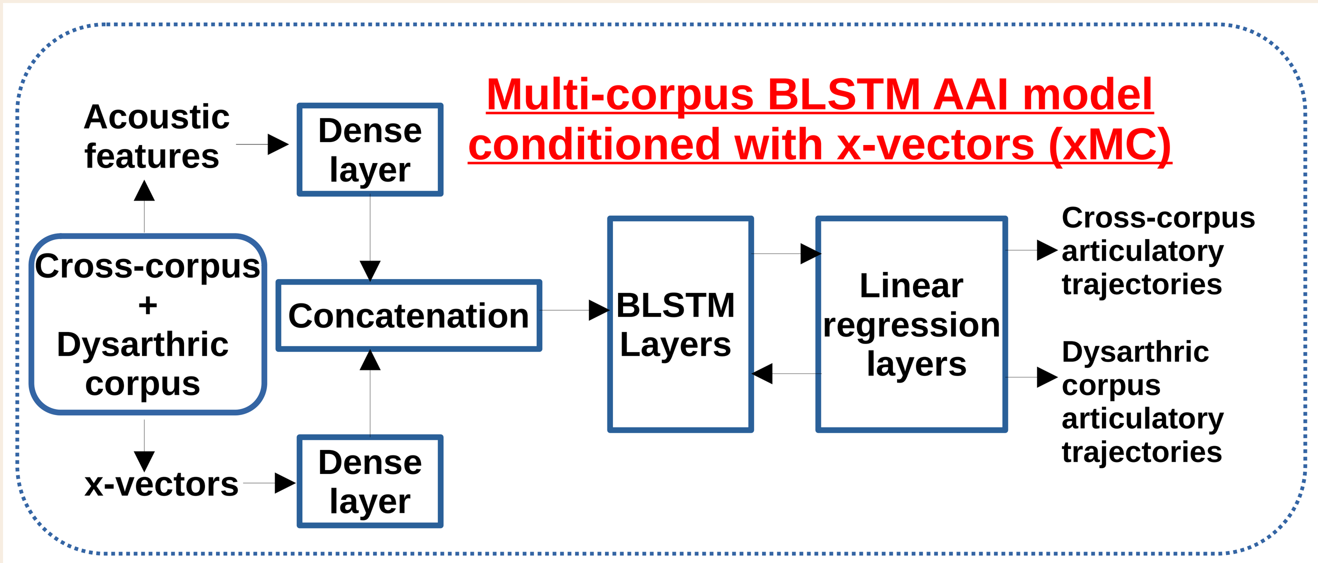
- **Transfer learning**:



- Following [1], we train a GBM which will serve as an initialization and fine-tune its weights(GBM-FT) on the dysarthric corpus to make it optimized for dysarthric speech.
- **Joint-training**: Experiments are done to account for multi-learning [5] and speaker conditioning [4], by pooling data from both the corpora.
- **Experimental Setup**:
  - 39-dims MFCCs(20ms window,10ms shift) as acoustic features.
  - All 38 subjects from the cross-corpus are used for experiments.
  - 5-fold cross validation setup in **seen** and **unseen** subject conditions.

Multi-corpus + Speaker Conditioned AAI (xMC)

- **Illustration of the multi-corpus AAI model conditioned with x-vectors**:



- Acoustic features and x-vectors [4] are fed into separate dense layers, and further sent to BLSTM layers after concatenation.
- The last layer of the BLSTM network is fed into two linear regression layers to obtain the first 8-dims of articulatory trajectories corresponding to the cross-corpus and the remaining 8-dims to that of the dysarthric corpus.
- **AAI models used in this work**:

AAI Model	Choice of hyperparameters
Randomly Initialised (RI) & Generalized Background Model (GBM)	3 BLSTMs (256 nodes), 1 linear regression layer.
Multi-corpus model (MC)	3 BLSTMs (256 nodes), 2 linear regression layers.
Speaker Conditioned (xSC)	3 BLSTMs (256 nodes), 1 linear regression layer.
Multi-corpus + Speaker Conditioned (xMC)	3 BLSTMs (256 nodes), 2 linear regression layers.

- **Baselines**: RI, GBM-FT, MC, and xSC AAI models.
- **Evaluation metric**: Pearson correlation coefficient between the ground-truth articulatory trajectories and their corresponding predicted articulatory trajectories.

Conclusions

- The rich cross-corpus database was beneficial to learn AAI for dysarthric speech, even though they were different in terms of speech stimuli, language, and age groups.
- The proposed multi-corpus AAI model conditioned with x-vectors(xMC) performed at par or better than the other baseline AAI models that used the cross-corpus.

References

- [1] Aravind Illa, et al., “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.
- [2] Korin Richmond, “Estimating articulatory parameters from the acoustic speech signal,” Ph.D. dissertation, University of Edinburgh, 2002.
- [3] Aravind Illa, et al., “Low resource acoustic-to-articulatory inversion using bi-directional long short term memory.,” in *Interspeech*, 2018, pp. 3122–3126.
- [4] Aravind Illa, et al., “Speaker conditioned acoustic-to-articulatory inversion using x-vectors,” in *Interspeech*, 2020, pp.1376–1380.
- [5] Nadee Seneviratne, et al., “Multi-corpus acoustic-to-articulatory speech inversion.,” in *Interspeech*, 2019, pp. 859–863.

Results & Discussions

- **Corpus dependent models**:

BLSTM nodes	RI				GBM	
	Seen		Unseen			
	HC	P	HC	P	HC	P
256	0.43	0.52	0.42	0.46	**0.5**	**0.5**

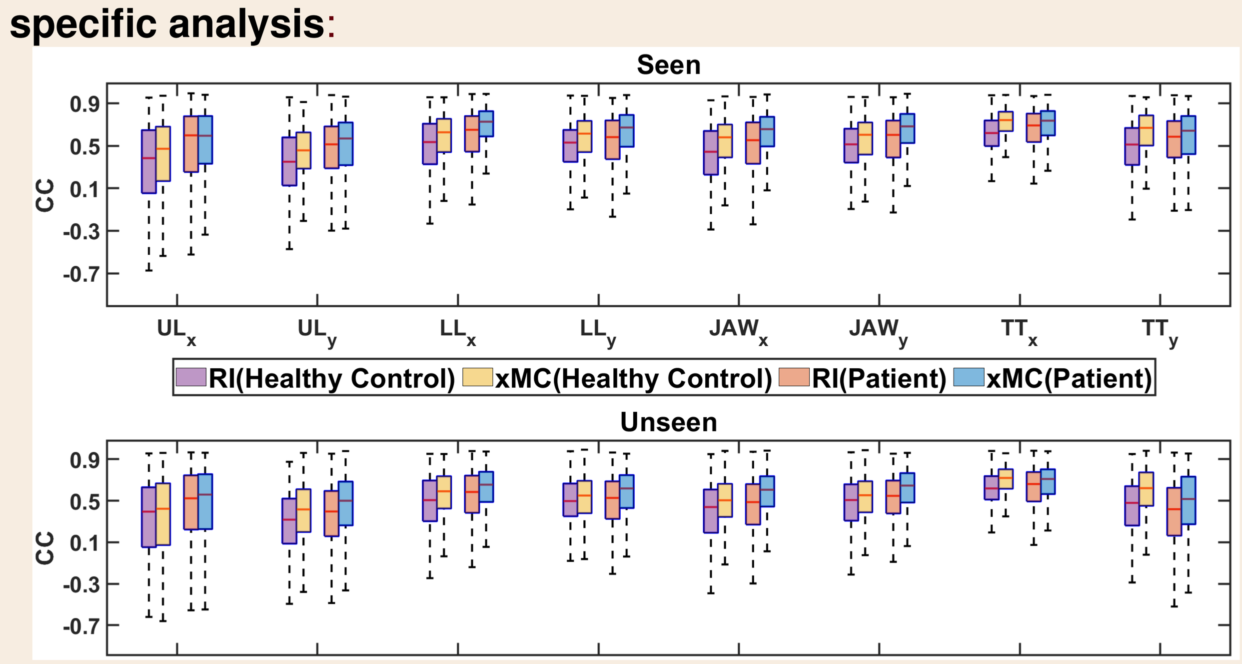
Making use of the cross-corpus was beneficial. Experiments were also done with different BLSTM nodes(32,64,128) to investigate if the RI model would overfit. It reached saturation at 256 BLSTM nodes.

- **Models using cross-corpus**:

	Seen									
	RI		GBM-FT		MC		xSC		xMC	
	HC	P	HC	P	HC	P	HC	P	HC	P
Avg	0.438	0.524	0.514	0.573	0.513	0.557	0.525	0.57	**0.538**	**0.593**
(Std dev)	(0.08)	(0.06)	(0.08)	(0.06)	(0.09)	(0.07)	(0.09)	(0.07)	**(0.08)**	**(0.07)**
	Unseen									
Avg	0.424	0.462	0.504	0.522	0.503	0.523	0.505	0.535	0.502	**0.538**
(Std dev)	(0.09)	(0.08)	(0.09)	(0.07)	(0.09)	(0.07)	(0.1)	(0.08)	(0.09)	**(0.07)**

Seen cases: xMC achieved improvements of ~13.16%(RI), ~3.49%(GBM-FT), ~6.46%(MC), and ~4.03%(xSC) for patients; Unseen cases: xMC>MC for patients, since conditioning with x-vectors leads to a better generalization to unseen speakers.

- **Articulatory specific analysis**:



(JAW_x and LL_x) and (TT_y and JAW_x) show maximum improvements for patients(seen, unseen subject conditions respectively).

- **Frequency characteristics**:

Articulatory Trajectories	Original		Seen				Unseen			
			xMC		RI		xMC		RI	
	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	7.38	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

The table reports cut-off frequencies(Hz) corresponding to 98% of the energy of original and predicted trajectories. Decline in speaking rate contributes to low values for patients.