Complexity-Performance Trade-off In Acoustic-to-Articulatory Inversion

Aravind Illa and Prasanta Kumar Ghosh

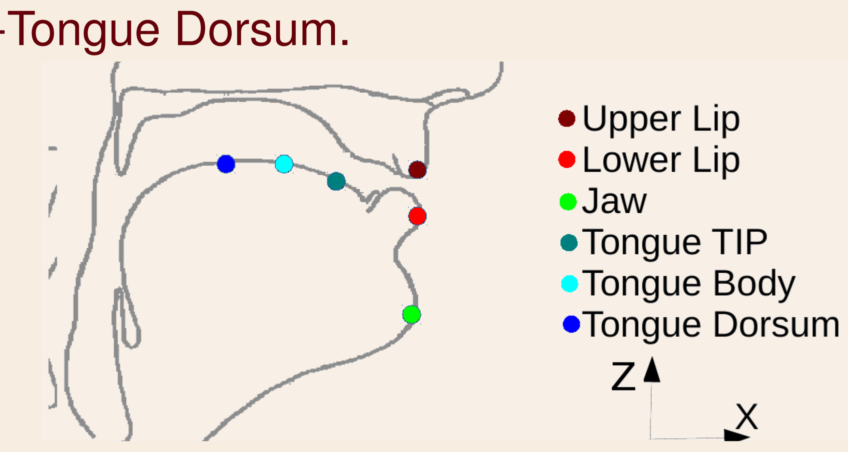
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India-560 012

Introduction

- Estimating articulatory motion from speech acoustics is known as acoustic-to-articulatory inversion (AAI) [1].
- Articulatory movements are known to vary slowly in nature, in order to preserve these characteristics in the predicted articulatory trajectories, these are further post-processed using different techniques like low-pass filtering (LPF), Kalman filtering and maximum likelihood parameter generation (MLPG).
- **Motivation**: To systematically compare AAI performance across different models, namely Gaussian mixture model (GMM), deep neural networks (DNN), convolution neural network (CNN) and bidirectional long-short term memory network (BLSTM), with respect to model complexities and post-processing techniques.
- **Key findings**:
  - Among the three post-processing methods, we observed that **MLPG** performs better than Kalman and Low-pass filtering.
  - Among all the AAI models, **BLSTM** yields the best complexity performance trade-off, which are followed by CNN, GMM and DNN with MLPG.

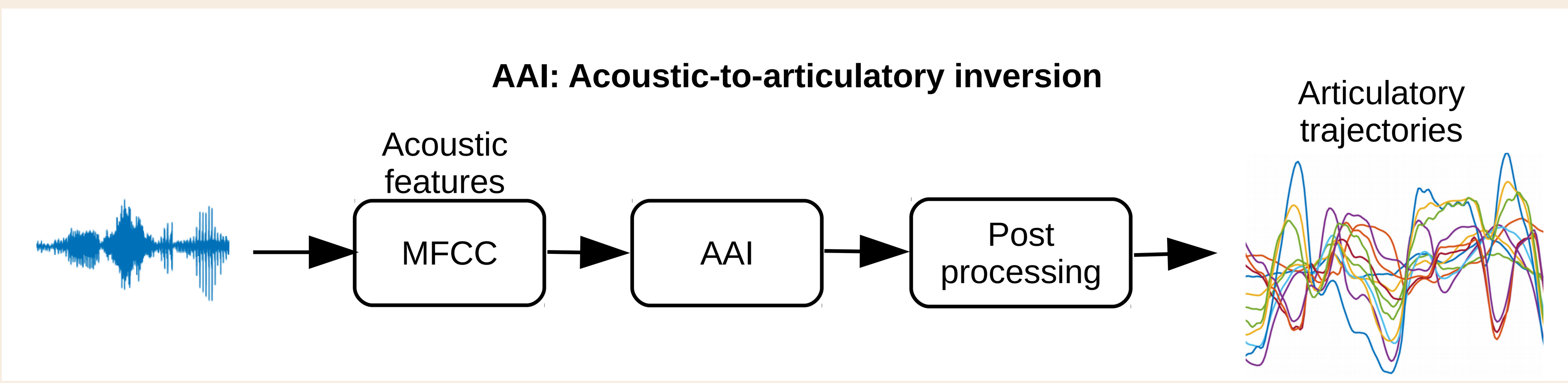
Data Collection

- Articulatory movement data recorder: → **EMA AG501**.
- Available sampling rate: 250 Hz and 1250 Hz.
- **Speech Stimuli**: 460 phonetically balanced English sentences from the MOCHA-TIMIT corpus [3] are chosen as the stimuli for data collection.
- **Six** sensors are connected: UL-Upper Lip, LL-Lower Lip, Jaw-Jaw, TT-Tongue Tip, TB-Tongue Body, TD-Tongue Dorsum.



- From the **six** sensors, we obtain **12**-dimensional articulatory features (AFs) namely, UL_x, UL_z, LL_x, LL_z, Jaw_x, Jaw_z, TT_x, TT_z, TB_x, TB_z, TD_x, TD_z.
- We collected data from 20 speakers comprising 10 males and 10 females in an age group of 20-28 years.

AAI & Experimental setup



- **Data processing and feature extraction**:
  - **Acoustics:**
    - 13-dim MFCC feature vector with frame length 20ms and shift being 10ms.
  - **EMA:**
    - Low-pass filtered with a cutoff at 25Hz
    - Down-sampled from 250Hz to 100Hz
    - Removed the mean sensor position for each articulatory feature in every sentence.
- 460 sentences from all the subjects are divided into 3 sets, 368 sentences for training data, 46 sentences each for test and validation data.
- **Objective measure**: Mean square error between the original and the predicted articulatory trajectories.

Results

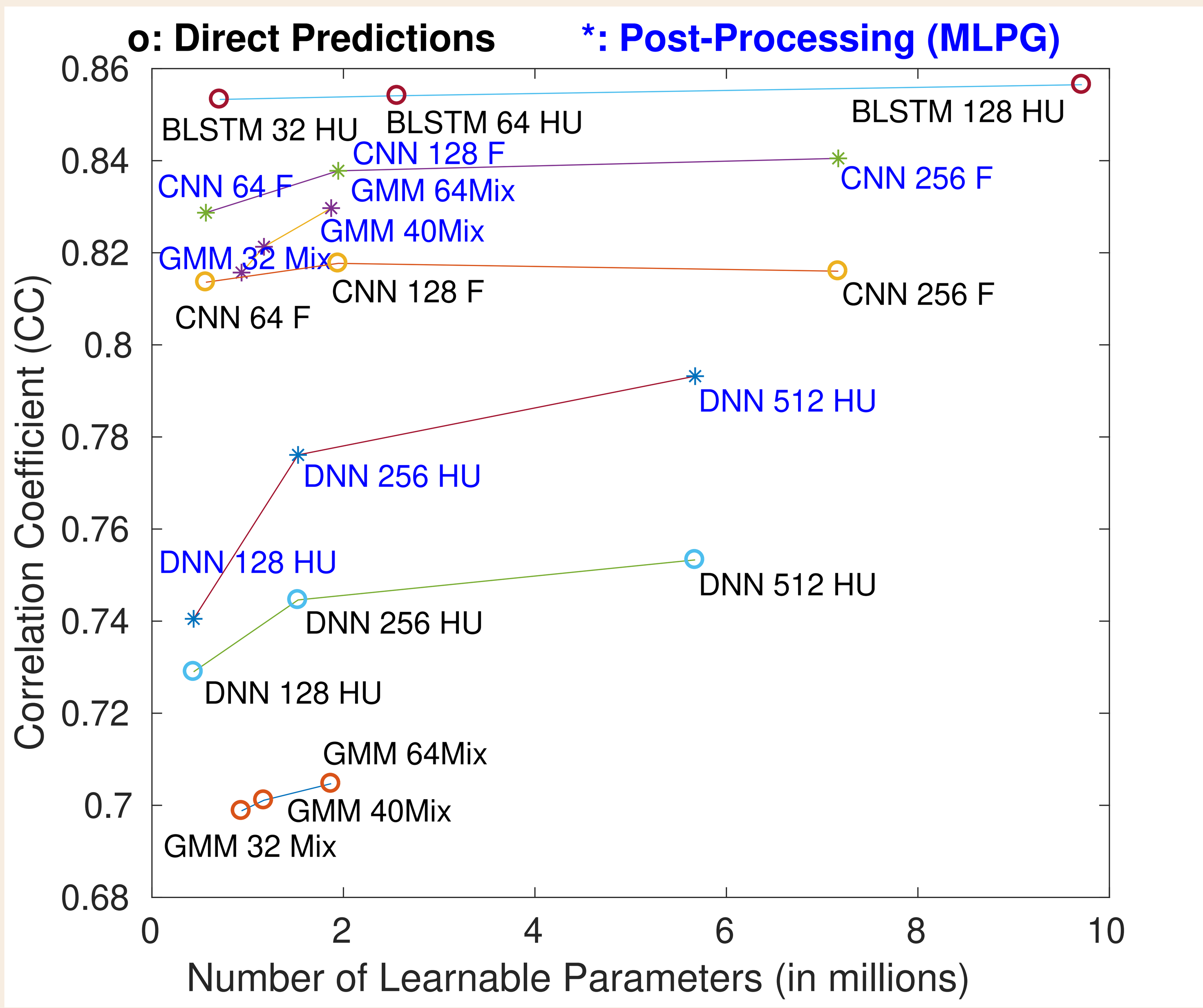
Choice of hyper-parameters:

- **GMM:** Full co-variance matrix with 32, 40 and 64 mixtures components
- **DNN:** 3-hidden layers with 126, 256 and 512 hidden units, last layer as linear regression layer.
- **CNN:** 3-hidden layers with 1-d convolutional filters of length 5, and number of filters in each layer was varied from 64, 128 and 256, last layer as linear regression layer.
- **BLSTM:** 3-hidden layers with 32, 64 and 128 hidden LSTM units, last layer as linear regression layer.
- **Evaluation metric**: Correlation coefficient (CC) [2]

Comparison of AAI post-processing techniques in terms of average CC:

	Direct	Kalman	LPF	MLPG
GMM	0.7047	0.7799	0.7862	**0.8297**
DNN	0.7533	0.7862	0.7897	**0.7932**
CNN	0.816	0.8268	0.8274	**0.8405**

AAI performance vs model complexity



Conclusion

- Among the three post-processing methods, we observed that MLPG performs better than Kalman and Low-pass filtering.
- Among all the AAI models, BLSTM yields the best complexity performance trade-off, which are followed by CNN, GMM and DNN with MLPG.
- Future work: Investigation on the demand of acoustic-articulatory data and complexity-performance trade-off in unseen case speaker evaluation across different models.

References

1. Korin Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping.," in Proceedings of the ICSLP, Pittsburgh, 2006, pp. 577-580.
2. Aravind Illa and Prasanta Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," Proc. Interspeech, pp. 3122–3126, 2018.
3. A. Wrench, MOCHA-TIMIT, speech database, Department of Speech and Language Sciences, Queen Margaret University College,Edinburgh, 1999

Acknowledgment: Authors thank **all the subjects** who participated for the study. Authors thank the **Pratiksha Trust**, and Department of Science and Technology (DST), Government of India, for their support.

aravindi@iisc.ac.in, prasantg@iisc.ac.in