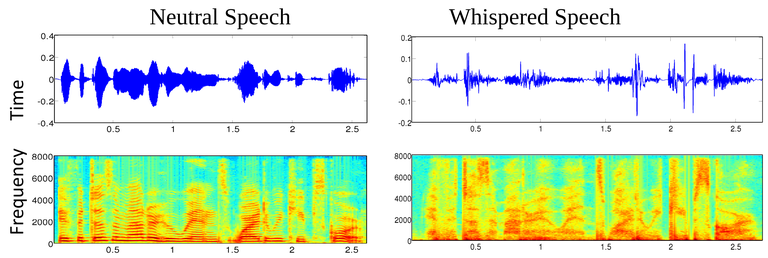
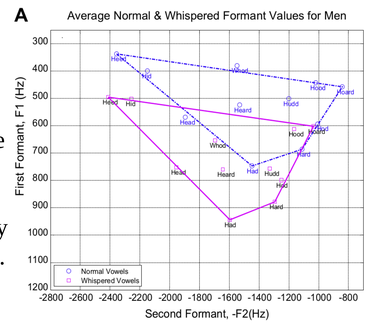
A COMPARATIVE STUDY OF ACOUSTIC-TO-ARTICULATORY INVERSION FOR NEUTRAL AND WHISPERED SPEECH

Aravind Illa, Nisha Meenakshi, Prasanta Kumar Ghosh
Indian Institute of Science (IISc), Bangalore-560012, India.
aravindsp@ee.iisc.ernet.in, gnisha@ee.iisc.ernet.in, prasantg@ee.iisc.ernet.in

Introduction



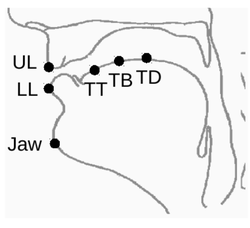
- Whispered speech is known to have different characteristics in acoustics and articulation compared to neutral speech.
- How well does whispered speech encode the articulatory information?
- We have chosen Acoustic-to-Articulatory Inversion (AAI) to address this problem.

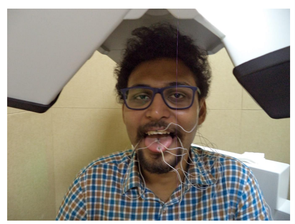


Major Observation: Whispered speech encodes most of the articulatory information!

Data Collection

Upper Lip: UL
Lower Lip: LL
Jaw: Jaw
Tongue TIP:TT
Tongue Body: TB
Tongue Dorsum: TD





- Articulatory movement data recorder: EMA AG501
- Six sensors are connected to the subject to obtain twelve articulatory trajectories.
- Stimuli: 460 phonetically balanced English sentences from the MOCHA-TIMIT corpus.
- Duration of the collected data, after removing silences before and after the sentences, are reported in the table below:

	Neutral (Min)	Whisper (Min)
M1	18.90	21.87
M2	24.38	25.85
F1	21.38	24.52
F2	20.23	21.57

Objective and Methodology

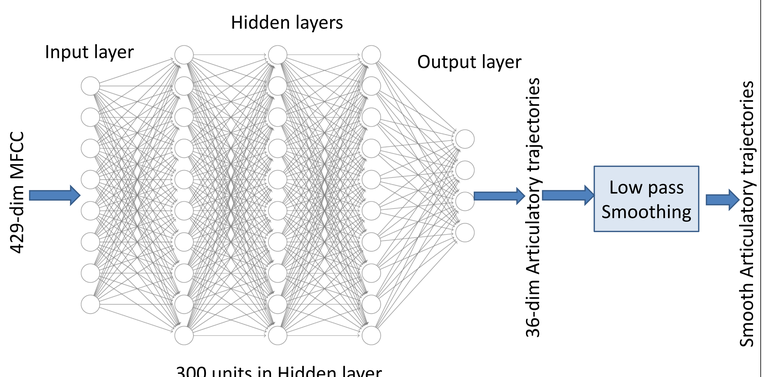
- To experimentally compare the accuracy with which the articulation can be recovered from the acoustics of both neutral and whispered speech.
- AAI is a regression problem, where the relationship between acoustics to articulators is known to be non-linear. So a DNN is trained to learn the non-linear relationship [1].
- Matched train-test evaluation: To know how much articulatory information is encoded in whisper speech acoustics in comparison to neutral speech by evaluating the performance of AAI on a test set using matched models.
- Mismatched train-test evaluation: To understand the differences in the acoustic-to-articulatory map for neutral and whispered speech. Cross model performance evaluation metric:

$$\text{Percentage drop in correlation } (PDCC) = \frac{\rho_o^i - \rho_c^i}{\rho_o^i} \times 100$$

where, ρ_o^i and ρ_c^i is be the correlation coefficients for the i-th articulator in the matched (o) and mis-matched (c) conditions respectively.

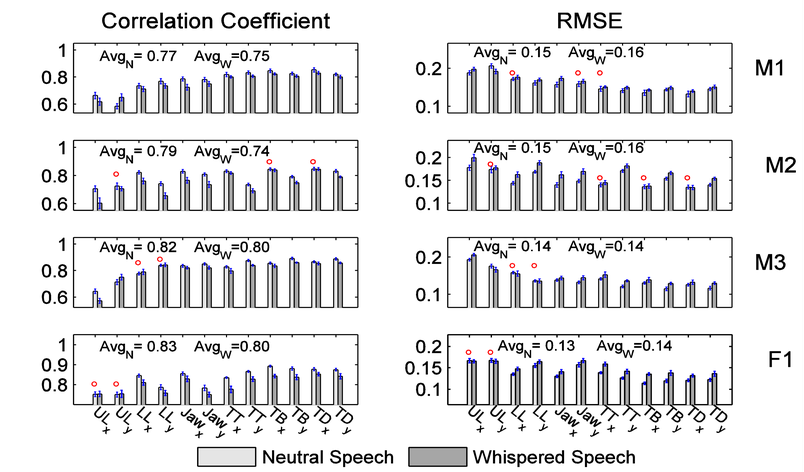
Experimental set-up

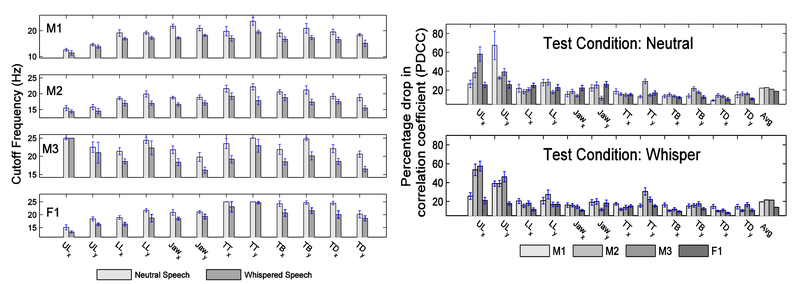
- Subject-wise AAI in a 10-fold cross-validation setup.
- The recorded speech is down sampled to 16kHz and 39-dim MFCC is computed for a window size of 20ms and a frame shift of 10ms followed by cepstral mean subtraction and variance normalization. To incorporate the contextual information, five frames before and after every frame are concatenated resulting in a 429-dimensional input feature vector.
- The 12-dimensional articulatory data is low-pass filtered with a cutoff frequency of 25Hz for all articulators, further the data is down-sampled to 100 Hz. Finally, we subtract the mean and divided by the standard deviation(SD) within every utterance for each dimension of the articulatory feature vector.
- The predicted articulatory features from DNN are jagged in nature, Since articulatory trajectories are smooth in nature, we low-pass filter each articulatory trajectory predicted by the DNN [2].



10-fold cross validation setup (8-train, 1-test, 1-validation sets)

Results and Conclusions





- We observe that the **articulator movement is smoother for whispered speech** compared to that of the neutral speech.
- Drop in AAI performance is observed in the mismatched train-test evaluation. This suggest that **acoustic to articulatory mapping** of whispered speech **is different** from that of neutral speech.
- Experiments also reveal that although the **information of the articulatory movements** is retained in whispered speech, it is **encoded differently**, compared to that in neutral speech.
- *Further investigation:* Required to examine the manner in which articulation during whisper speech could be different from that for neutral speech and **develop an adaptation technique**.

References

1. Zhiyong Wu, Kai Zhao, Xixin Wu, Xinyu Lan, and Helen Meng, "Acoustic to articulatory mapping with deep neural network," Multi-media Tools and Applications, vol. 74, no. 22, pp. 9889–9907, 2015.
2. 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.

Acknowledgement: We thank all subjects who participated in EMA data collection.