Augmenting acoustic phonetics with articulatory features for phone recognition.

Ganesh Sivaraman[1], Vikramjit Mitra[2], Hosung Nam[3], Elliott Saltzman[4], Carol Espy-Wilson[1]

[1] Institute for Systems Research, University of Maryland, College Park USA, [2]Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA, [3]Korea University, Seoul, South Korea, Physical Therapy & Athletic Training, Boston University, Boston, Massachusetts, USA

Introduction

- ❑ In articulatory phonetics, a phoneme's identity is specified by its articulator-free (manner) and articulator-bound (place) features.
- ❑ Previous studies have shown that acoustic-phonetic features (APs) can be used to segment speech into broad classes determined by the manner of articulation of speech sounds.
- ❑ This effort is to extend previous efforts [3] to develop a landmark system by adding in components to recognize place of articulation.
- ❑ The objective of this research is to test the performance of estimated articulatory trajectories for place of articulation classification.
- ❑ In the first stage, the speech signal is segmented into broad classes using ideal phonetic transcriptions into 5 broad classes (Vowels – V, Fricatives – Fr, Sonorant Consonants – SC, Stops – ST and Silence – SIL).
- ❑ A single feature vector composed of Mel Frequency Cepstral Coefficients (MFCCs) and estimated articulatory trajectories (estTV) were extracted from the broad class segments. Fixed length feature vectors were obtained from variable length segments using a statistical parameterization of the MFCCs and estTVs.
- ❑ The combination of MFCCs with estTVs provided an average of 2% relative improvement in recognition of the place features compared to MFCCs alone.

System description

- ❑ Figure 1 shows the diagram of the system implemented for this work.
- ❑ Table 1 summarizes the place of articulation classes under each broad class.

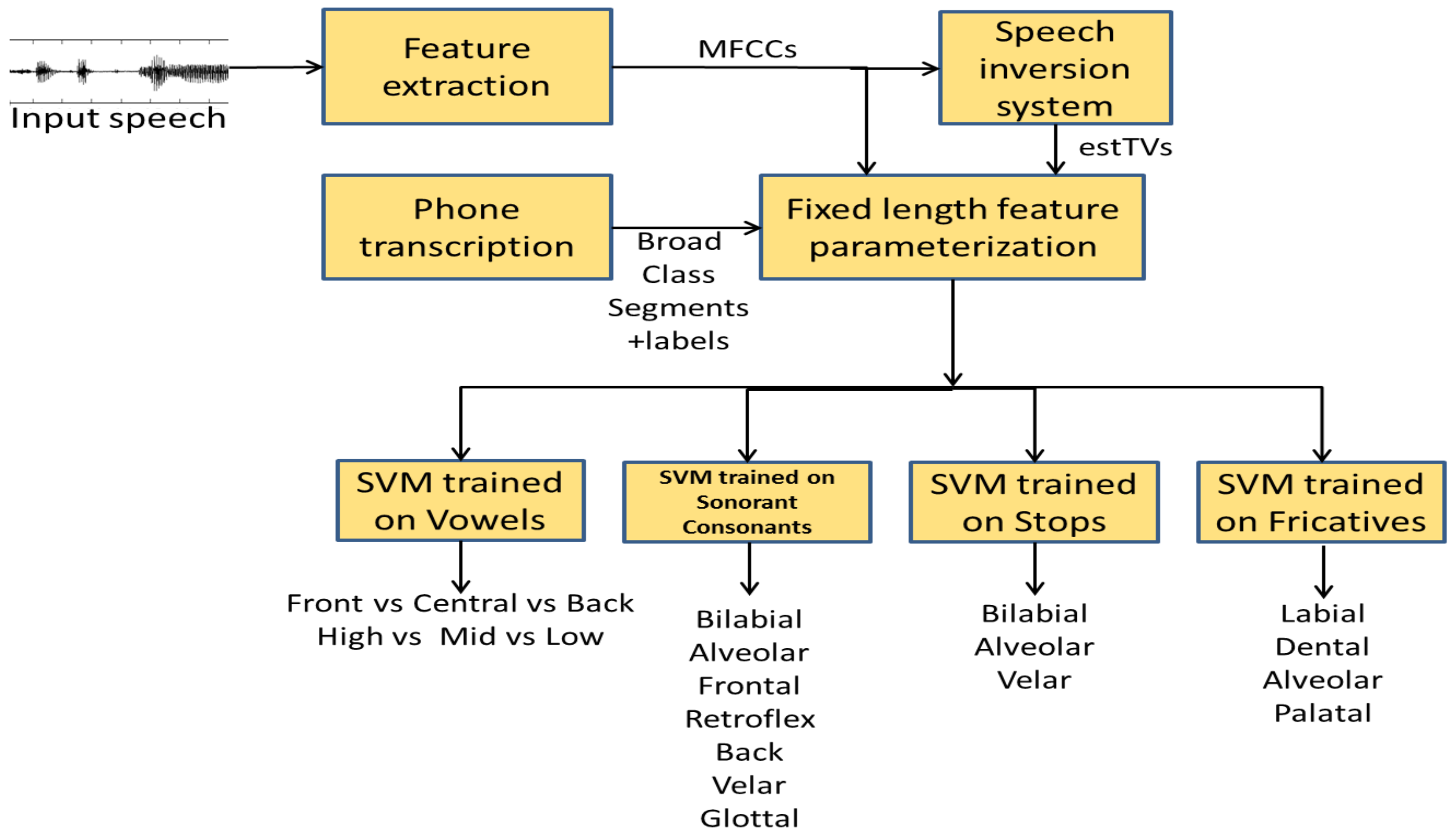

Figure 1

Table 1: Place of articulations for various broad classes

Broad class	Places of articulation
Vowels (V)	Back vs Central vs Front High vs Mid vs Low
Stops (ST)	Bilabial vs Alveolar vs Velar
Fricatives (Fr)	Labial vs Dental vs Alveolar vs Palatal
Sonorant Consonants (SC)	Bilabial vs Alveolar vs Frontal vs Retroflex vs Back vs Velar vs Glottal

Estimating Articulatory features

- ❑ The X-ray Microbeam (XRMB) data was used to train the acoustic to articulatory speech inversion systems.
- ❑ Pellet trajectories were converted to Tract Variables (TVs) using the geometric transformations described in [1]
- ❑ Multi-layer feed forward neural networks were trained to estimate the TVs from contextualized MFCCs with a context window of 160ms.
- ❑ 1720 utterances (35 females, 25 males) from the XRMB database were used to train the speech inversion system.
- ❑ The correlation between estimated and groundtruth TVs are shown in Table 2.

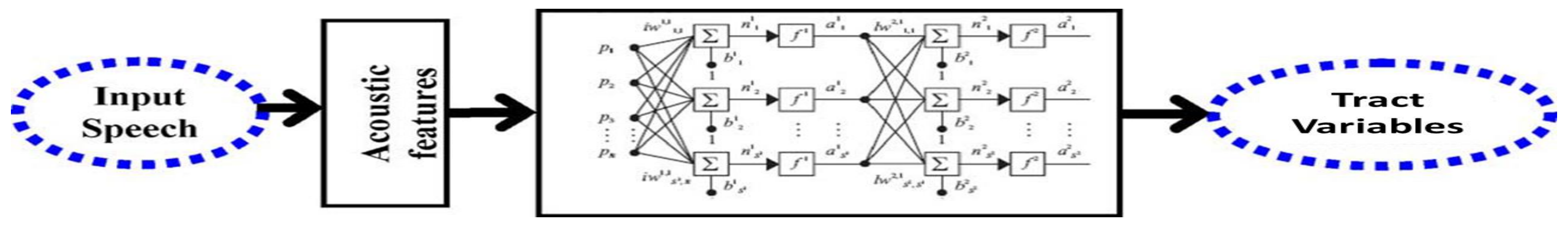

Figure 2

Table 2: Correlations of estimated TVs with groundtruth TVs

Tract Variables	LA	LP	TBCL	TBCD	TTCL	TTCD
Correlation	0.66	0.56	0.78	0.59	0.65	0.76

- ❑ Figure 3 shows the plot of estimated TVs for an utterance "Gwen planted green beans" from the TIMIT database.
- ❑ Note that the TV estimator was trained on XRMB database which is completely different from the TIMIT dataset chosen for this work.

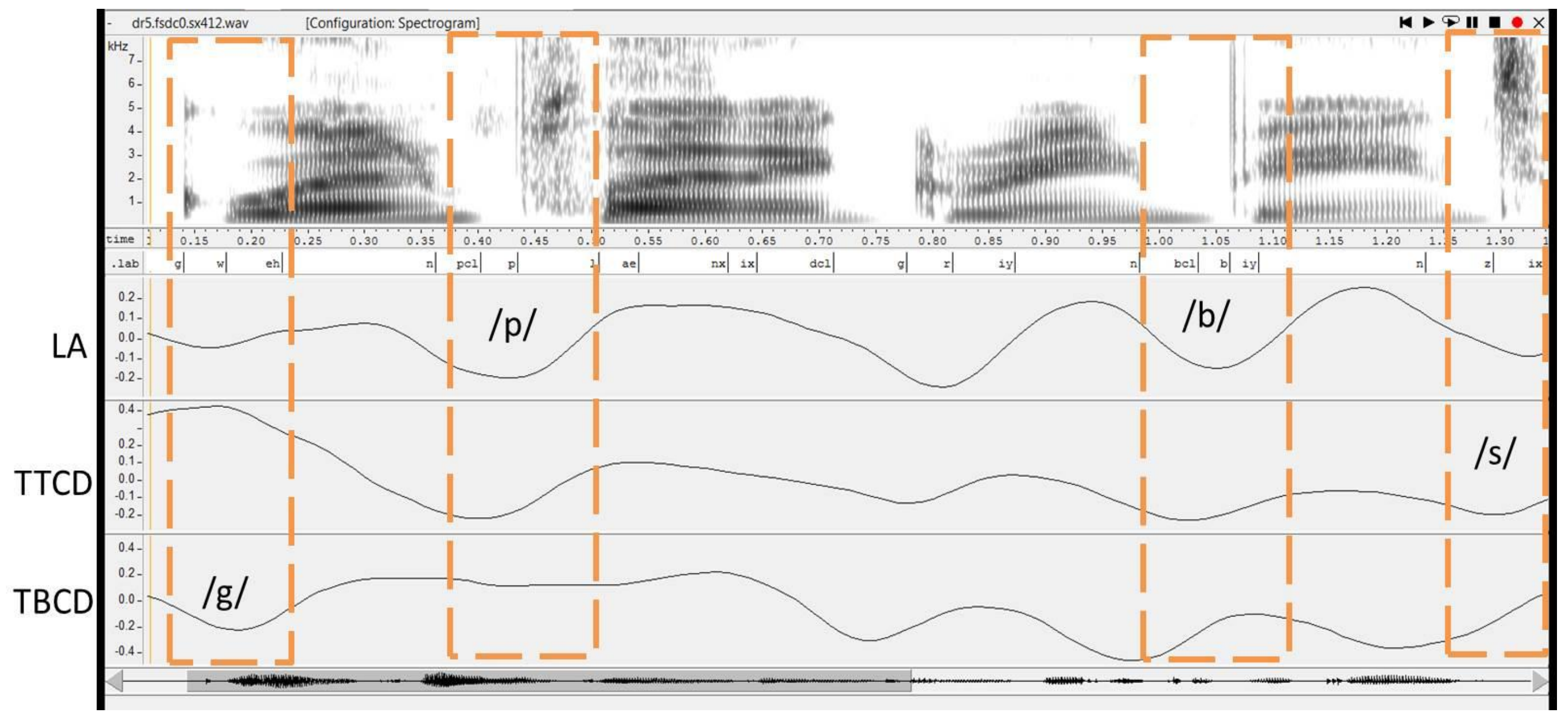

Figure 3

Fixed length feature parameterization

- ❑ The broad class segments of the TIMIT utterances obtained from the transcription were of different lengths.
- ❑ Classification of place using SVMs required a fixed length feature vector that accurately summarized the MFCCs and estTV features in the broad class segment.
- ❑ Table 3 shows the features obtained from each component of MFCCs and estTVs for each broad class segment.

Table 3: Fixed length features

	Description	Formula
1	Min	Min{Feat$_i$(t)} across time t
2	Max	Max{Feat$_i$(t)} across t
3	Mean	Mean{Feat$_i$(t)}
4	Max slope	Max($dFeat_i/dt$)
5	Min slope	Min($dFeat_i/dt$)
6	Min absolute slope	Min($\lvert dFeat_i/dt \rvert$)

Results

- ❑ Support Vector Machines (SVM) with Radial basis function kernels were trained to perform the sub-classification of broad class segments.
- ❑ A separate SVM was trained for each broad class.
- ❑ The bar charts in Figure 4 shows the classification accuracies for each broad class.

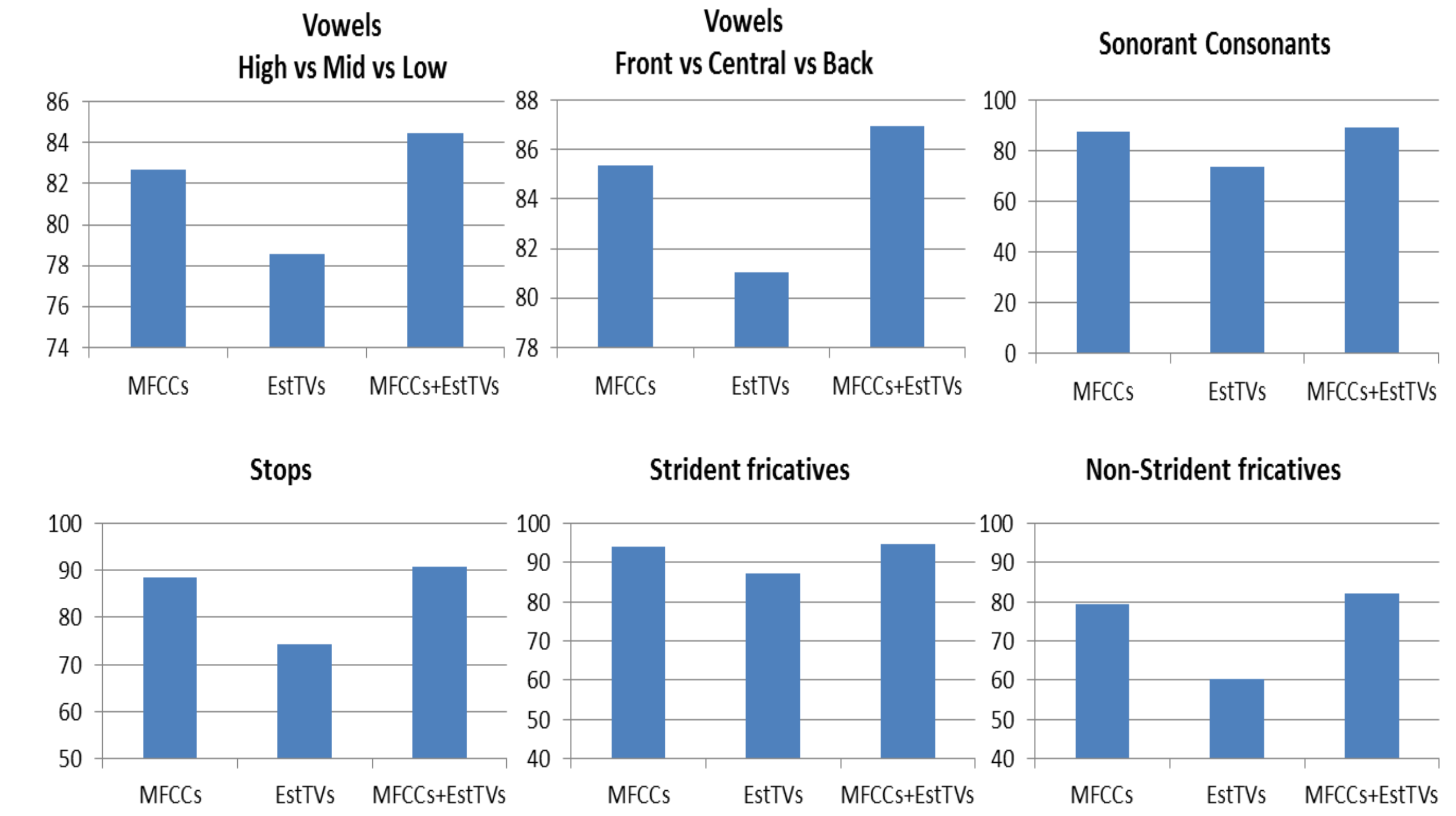

Figure 4

Conclusions

- ❑ Augmenting acoustic features with estimated articulatory features provides an average of 2% relative improvement in accuracy for Vowels, Stops and Sonorant consonants.
- ❑ For fricatives, the improvement is marginal. Adding contextual information can help in improving the classification accuracy.
- ❑ Articulatory features alone do not provide superior performance compared to acoustic features.
- ❑ Overall, articulatory features combined with acoustic features are robust for classifying the place of articulation of phonemes.

Future directions

- ❑ Use an automatic landmark detection system [2] to segment utterances into broad classes.
- ❑ Instead of using a speech inversion system trained on complete utterances, we plan to train specific speech inversion systems for each broad class. Features from such tuned speech inversion systems might be more accurate than a generic system. In fact, we expect the results may increase by 5%.
- ❑ Combining broad class probability estimates with place of articulation estimates to decode phone sequences.
- ❑ Train an articulatory gesture recognition system and combine gestural estimates with TVs and MFCCs to perform a phone recognition task.

References

1. H. Nam, V. Mitra, M. Tiede, M. Hasegawa-Johnson, C. Espy-Wilson, E. Saltzman, and L. Goldstein, "A procedure for estimating gestural scores from speech acoustics.," J. Acoust. Soc. Am., vol. 132, no. 6, pp. 3980–9, Dec. 2012.
2. V. Mitra, "Articulatory information for robust speech recognition.", PhD Thesis, Dept. of Electrical and Computer Engineering, University of Maryland, Colege Park, USA 2010.
3. Juneja, A. and Espy-Wilson, C. (2008) "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition", J. Acoust. Soc. of Am., 123(2), pp. 1154-1168.

The authors would like to thank Dr. Amit Juneja for his help in training the landmark detection system
Acknowledgement: This research was supported by NSF Grant #IIS-1162046