Articulatory representations to address acoustic variability in speech

Ganesh Sivaraman, Prof. Carol Espy-Wilson

UNIVERSITY OF MARYLAND 1856
INSTITUTE FOR SYSTEMS RESEARCH
A. JAMES CLARK SCHOOL OF ENGINEERING

Motivation

- **The Motor theory of speech perception** argues that when perceiving speech, listeners access their own knowledge of how phonemes are articulated
- Most speech recognition systems incorporate acoustic features inspired by the human auditory system; however, there is little representation of the motor pathway.
- **Articulatory Phonology** proposes that speech can be decomposed into a constellation of articulatory gestures and it provides a unified framework for understanding how spatiotemporal changes in the pattern of underlying speech gestures lead to acoustic consequences that are typically reported deletion, insertions and substitutions.
- The objective of this research is to propose robust articulatory representations of speech and develop an acoustic-articulatory joint model for improved speech recognition

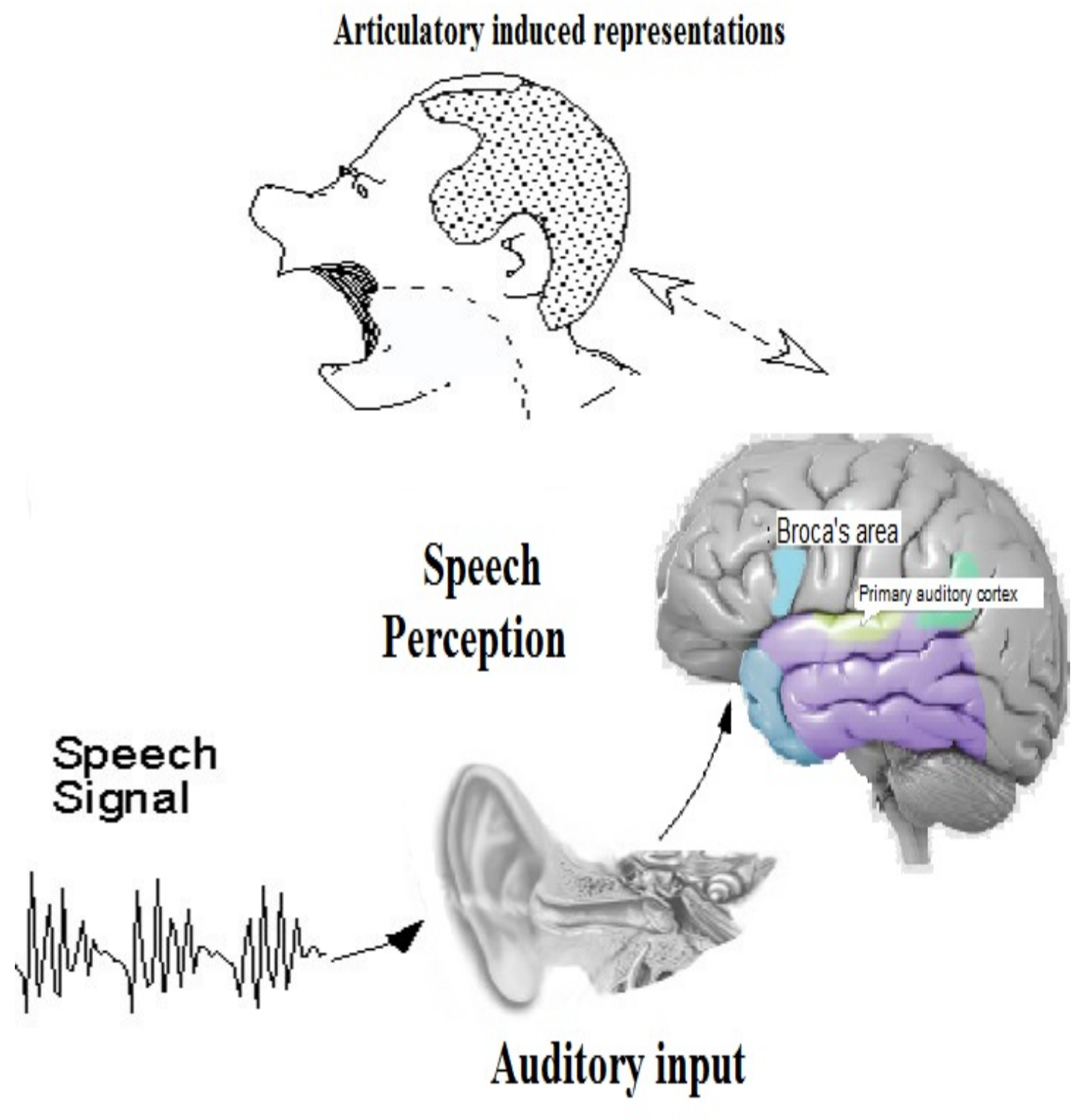

Fig 1

Speech inversion: From acoustics to articulations

- The amount of simultaneous acoustic and articulatory data is very limited and not easy to obtain. Hence it is essential to build models to estimate TVs from acoustics. These models are referred to as **speech inversion systems.**
- Feed forward deep neural networks were trained to estimate the TVs from contextualized Mel-frequency cepstral coefficients (MFCCs) (160 ms of speech to capture the full gesture). Figure 2 shows the architecture of the speech inversion system.
- The Wisconsin X-ray Microbeam (XRMB) dataset was used to train the inversion system.
- The training set contained 36 different speakers. The model was tested on 5 speakers from the XRMB dataset not contained in the training set.
- This is the first such speaker independent speech inversion system trained on so many speakers!
- The correlation between estimated and groundtruth TVs are shown in Table 1.
- Fig. 3 compares the estimated and groundtruth TVs which show a close match during critical gestures.

Fig 2

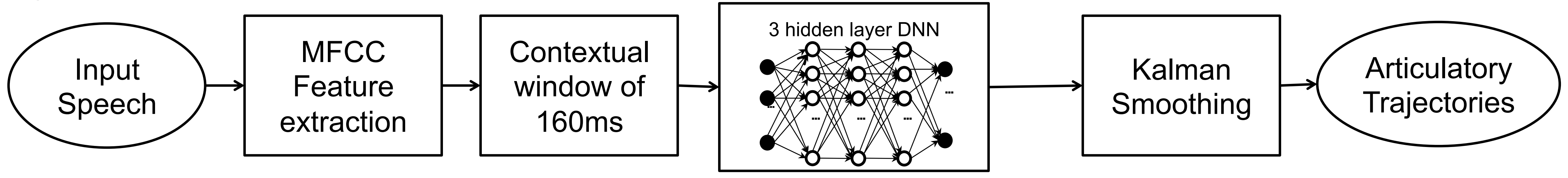

Table 1: ***Test set correlation values for TV estimators trained on natural and synthetic XRMB database***

Tract Variables	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
Correlation between actual and estimated TVs	0.79	0.67	0.88	0.75	0.76	0.86	**0.78**

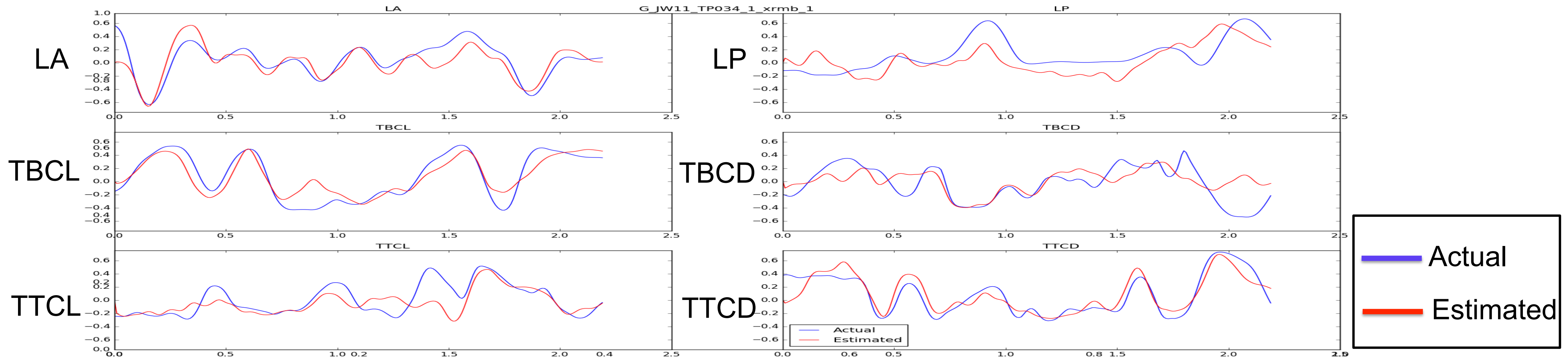

Fig 3: Groundtruth TVs (blue) and estimated TVs (red) for "Combine all of the ingredients in a large bowl" (unseen by the training data)

Analyzing speech variability with estimated articulatory trajectories

- Acoustic variability due to coarticulation and lenition are most observed in fast spoken speech. We recorded simultaneous acoustic and articulatory data from two subjects speaking at normal and fast rates.
- The speech inversion systems were successful in estimating the TVs in such challenging scenario. Figure 5 shows the TVs for an utterance of "perfect memory" at fast and normal rate estimated by the TV estimator. Note that the TV estimator was not trained and tested on speech from the same talker.

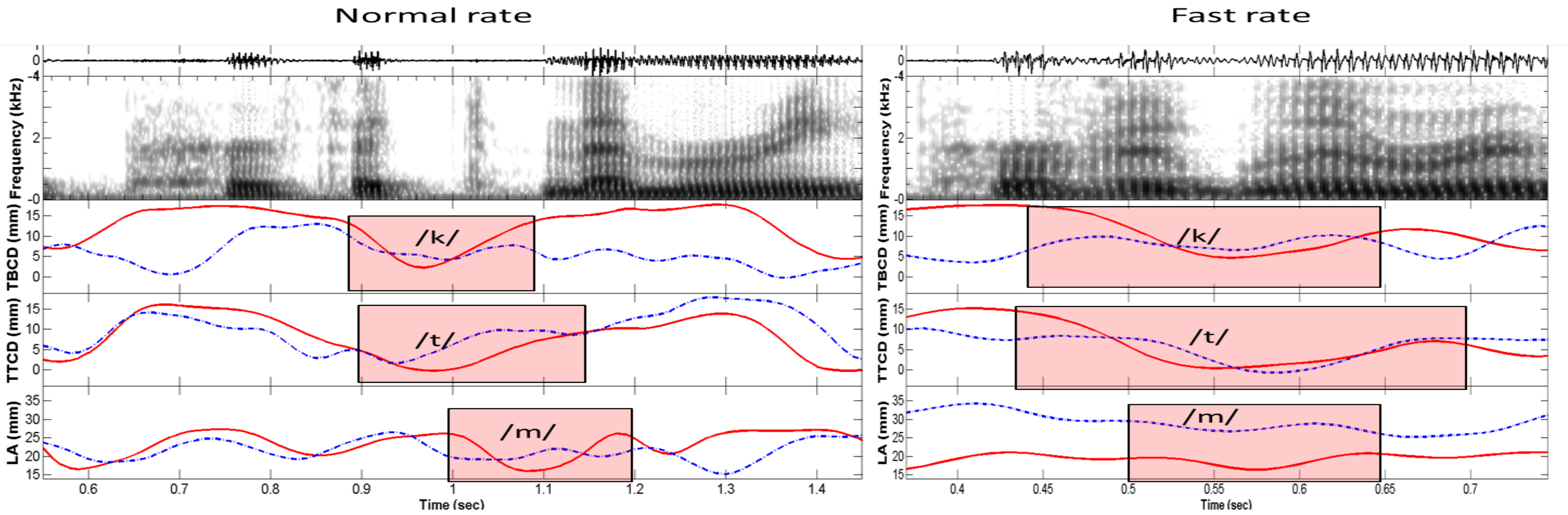

Fig 5: Groundtruth TVs (red) and TVs estimated from TV estimator (blue) for "perfect memory" utterance at fast and normal speaking rates. The boxes show the /k/, /t/, and /m/ gestures.

Application to Speech recognition

- Automatic Speech recognition (ASR) experiments were performed on the Wall Street Journal (WSJ1) corpus.
- The Baseline system took Gammatone Filterbank (GFB) features as input and used of a DNN acoustic model
- WSJ1 Training set: 78 hours of news broadcast speech from 284 speakers
  WSJ1 Test set: 0.8 hours of speech from
  20 speakers not in the training set
- TVs were estimated for the training and test set using the XRMB TV estimator.
- Note that the XRMB TV estimator is a completely different speech corpus from subjects not seen in the WSJ1 dataset.
- Figure 6 shows the ASR architecture used to combine acoustic and articulatory features.
- Table 2 shows the Word Error Rate (WER) and Phone Error Rate (PER) obtained in the WSJ1 test set.

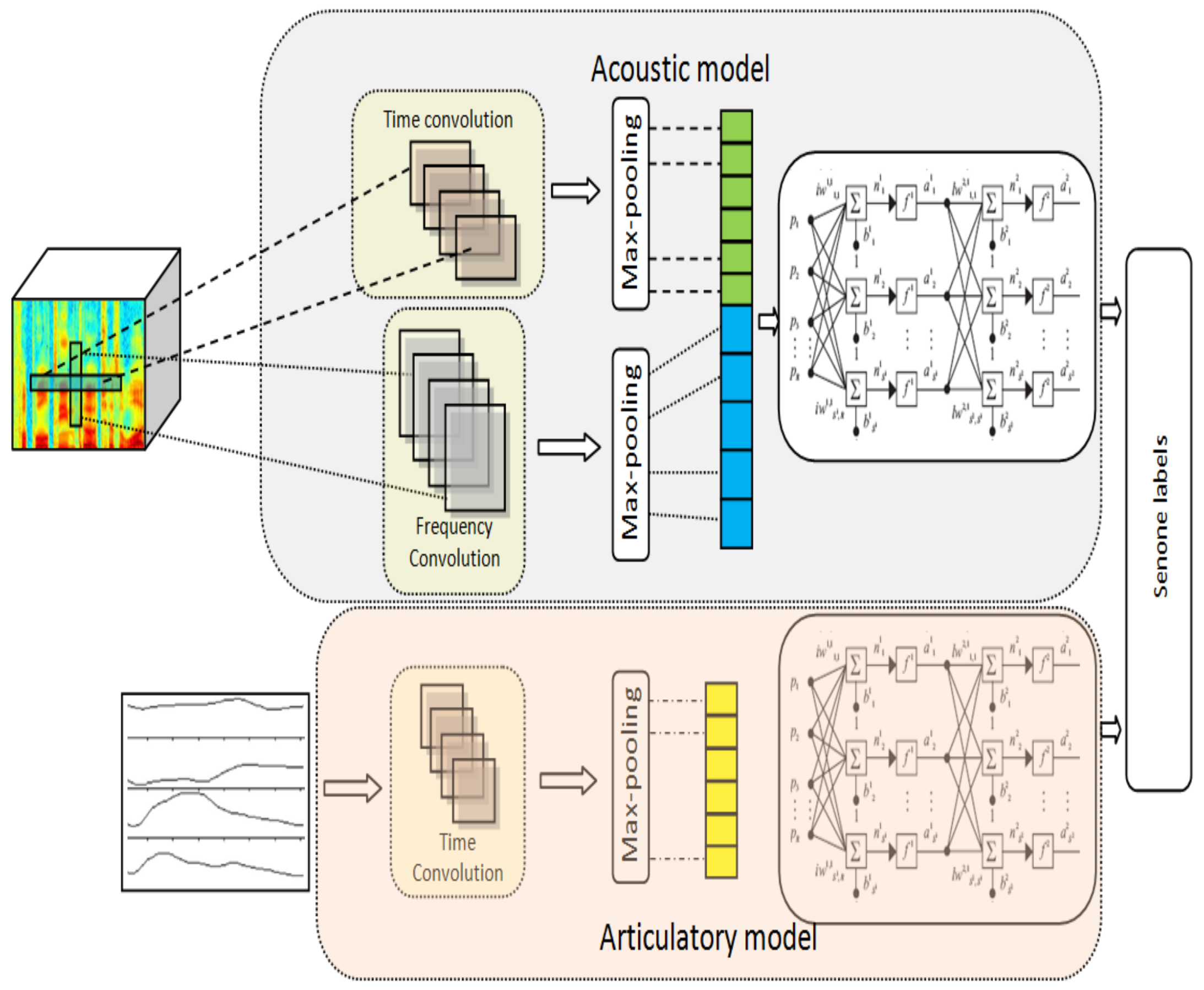

Fig 6: Hybrid Convolutional Neural Network (HCNN) architecture. Time and frequency convolutions on acoustic features and time convolutions on TVs

Table 2: Results from the WSJ1 ASR experiments

Features	Acoustic model	WER	PER
GFB	DNN_5x1024	6%	16.40%
GFB+xrmbTV	HCNN	5.70%	15.70%

References

1. Browman, C. and Goldstein, L. (1990) "Gestural specification using dynamically-defined articulatory structures", J. of Phonetics, Vol. 18, pp. 299-320.
2. V. Mitra, "Articulatory information for robust speech recognition," Ph.D. dissertation, University of Maryland, College Park, Maryland, 2010.
3. J. R. Westbury, "Microbeam Speech Production Database User''s Handbook," IEEE Pers. Commun. - IEEE Pers. Commun., 1994.
4. G. Sivaraman, V. Mitra, and C. Y. Espy-Wilson, "Fusion of acoustic, perceptual and production features for robust speech recognition in highly non-stationary noise," in Proc. CHiME-2013, Vancouver, Canada, June 2013, pp. 65–70.
5. Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman, Mark Tiede, Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition, Speech Communication, Volume 89, May 2017, Pages 103-112

This work was supported by:
NSF Grant # IIS-1162525
NSF Grant # BCS-1436600
The International Graduate Research Fellowship 2016
The Future Faculty Fellowship

I would like to thank Dr. Vikramjit Mitra for his help in conducting most of the speech recognition experiments performed in this research.