Speech production information for enhanced speech recognition

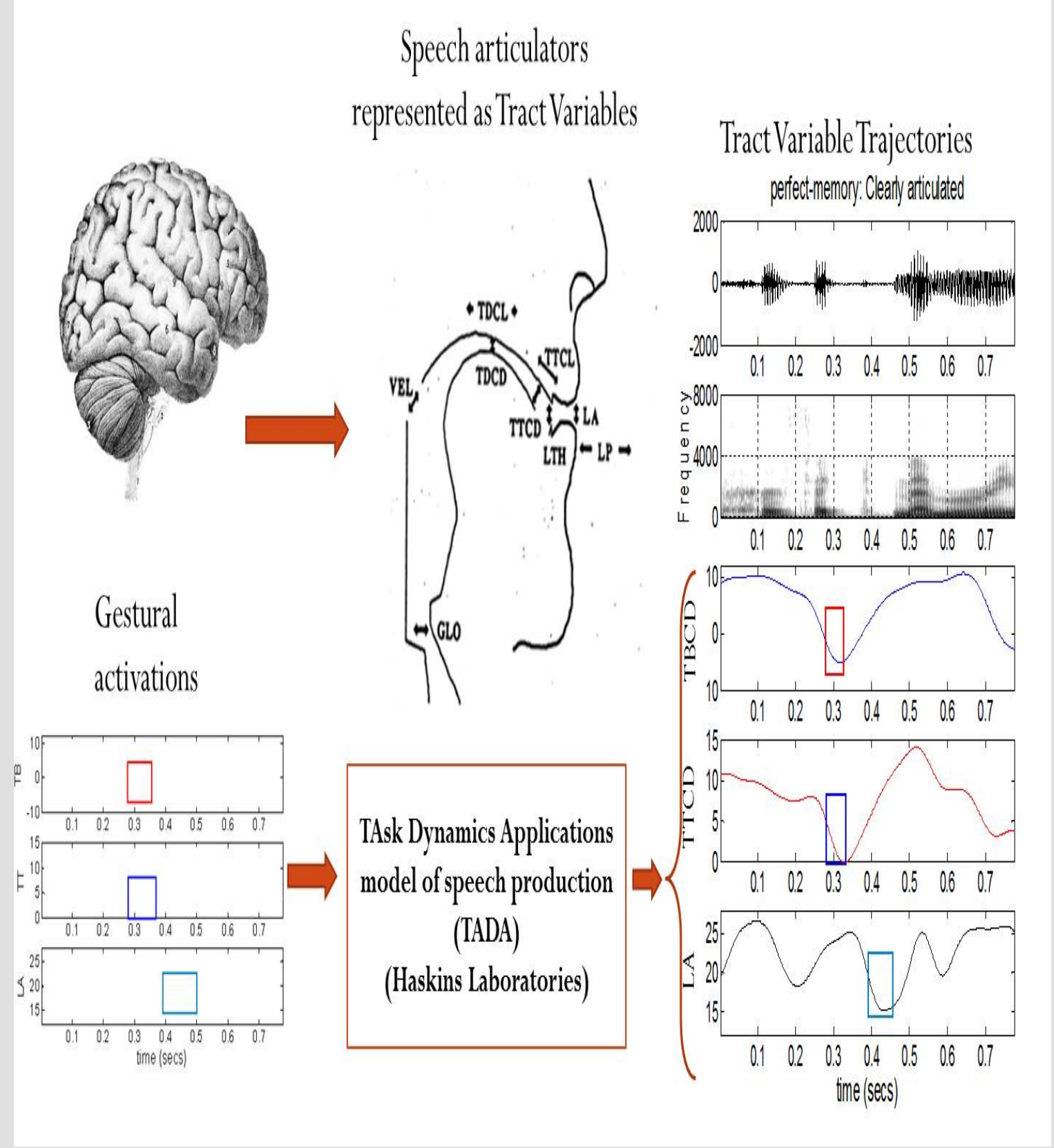
Ganesh Sivaraman, Carol Espy-Wilson

30 THIRTY YEARS OF SYSTEMS RESEARCH EXCELLENCE

INSTITUTE FOR SYSTEMS RESEARCH
A. JAMES CLARK SCHOOL OF ENGINEERING

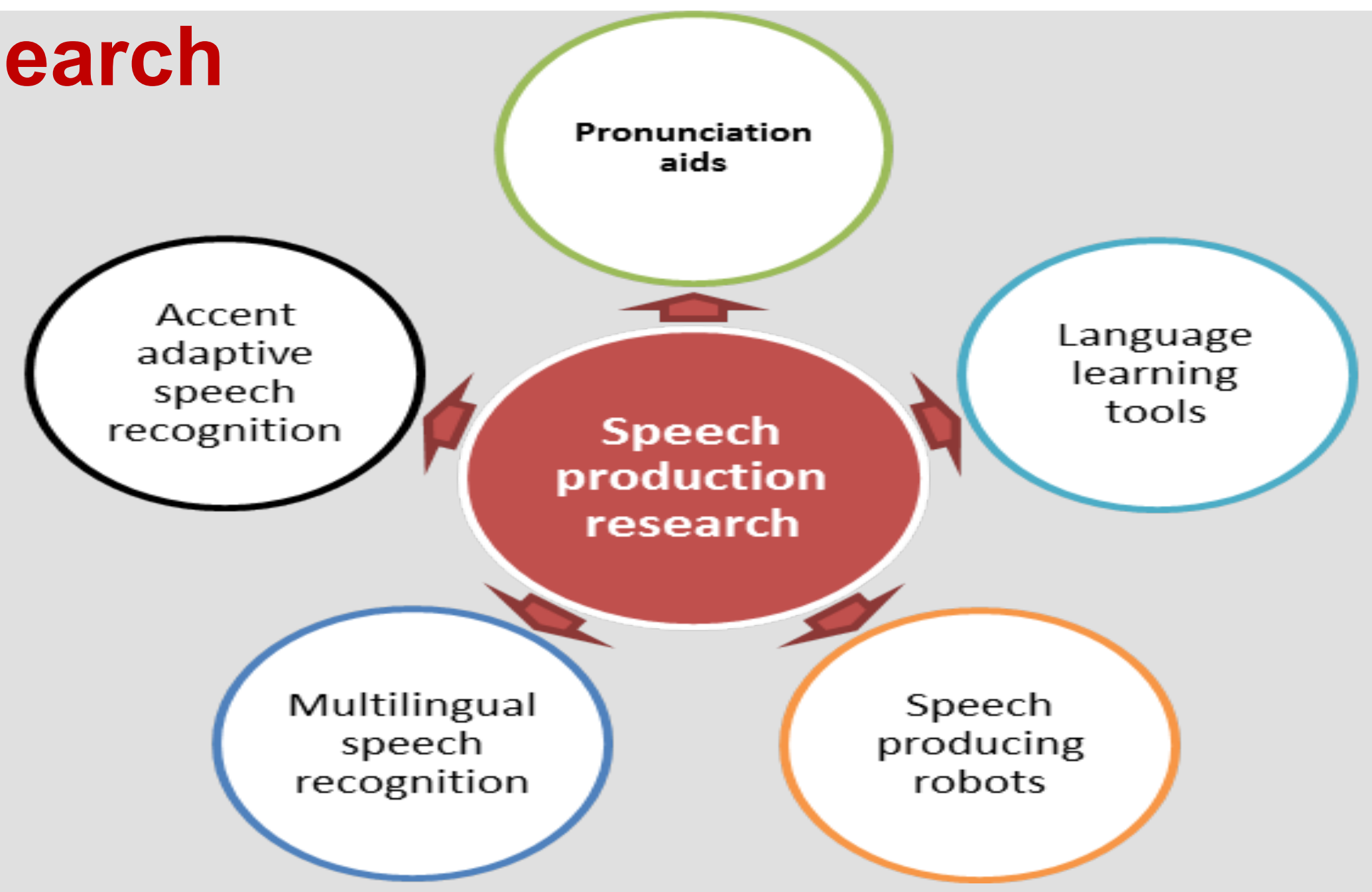
Motivation

- Traditional Automatic Speech Recognition (ASR) systems suffer from robustness due to acoustic variations. Phone based ASR systems do not adequately model (1) the temporal overlap of neighboring sounds due to co-articulation and (2) changes in acoustic properties due to reduction in articulatory precision (lenition). Di-phone and Tri-phone models which are attempts to model such variability require large training data and limit contextual influence to immediate neighbors.
- **Articulatory Phonology** proposes that human 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 as assimilations, insertions, deletions and substitutions.


Speech articulators represented as Tract Variables
Tract Variable Trajectories
perfect-memory: Clearly articulated
Gestural activations
TAsk Dynamics Applications model of speech production (TADA) (Haskins Laboratories)
TDCL
TDCD
TTCL
TTCD
VEL
LA
LTH
LP
GLO
TB
TT
LA
Frequency
TBCD
TTCD
time (secs)

Future of this research

- Apply the TV and gesture based features to **large vocabulary ASR** tasks. This work will involve annotating a Large vocabulary speech database with articulatory gestures.
- Explore novel ASR architectures incorporating gestures and Tract Variables
- Develop **multi-accent ASR** systems taking advantage of articulatory gestures.
- **Develop Multi-lingual ASR** systems using articulatory gestural models.
- Improve models of speech production through knowledge from articulatory recordings.
- Establish gesture based annotation of multiple languages.
- **Develop pronunciation correction** systems providing feedback to users in the form of articulatory corrections.
- Develop tools to enable **language learning** through articulatory speech aids.


Speech production research
Pronunciation aids
Language learning tools
Speech producing robots
Multilingual speech recognition
Accent adaptive speech recognition

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.

Parameterizing speech production

- The Task Dynamics & Applications model (TADA) represents speech production as a dynamical system of vocal tract constriction variables.
- Tract variables (TVs) are continuous time functions that specify the shape of the vocal tract in terms of constriction degree and location of the constrictors. Fig. 2 shows a block diagram of TADA [1].

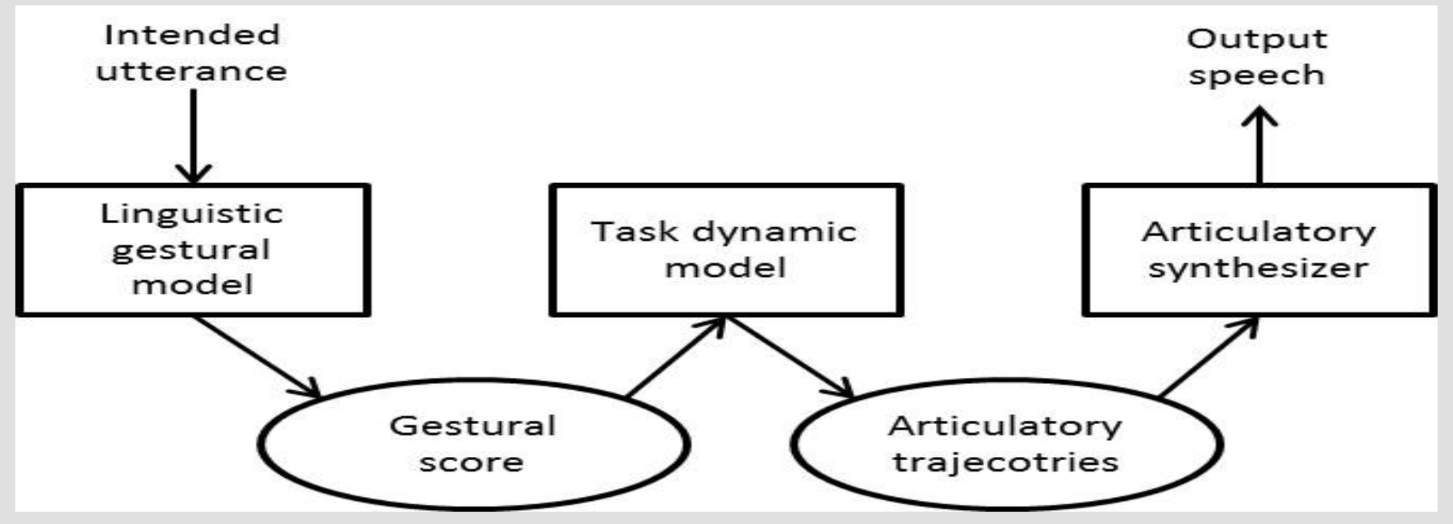
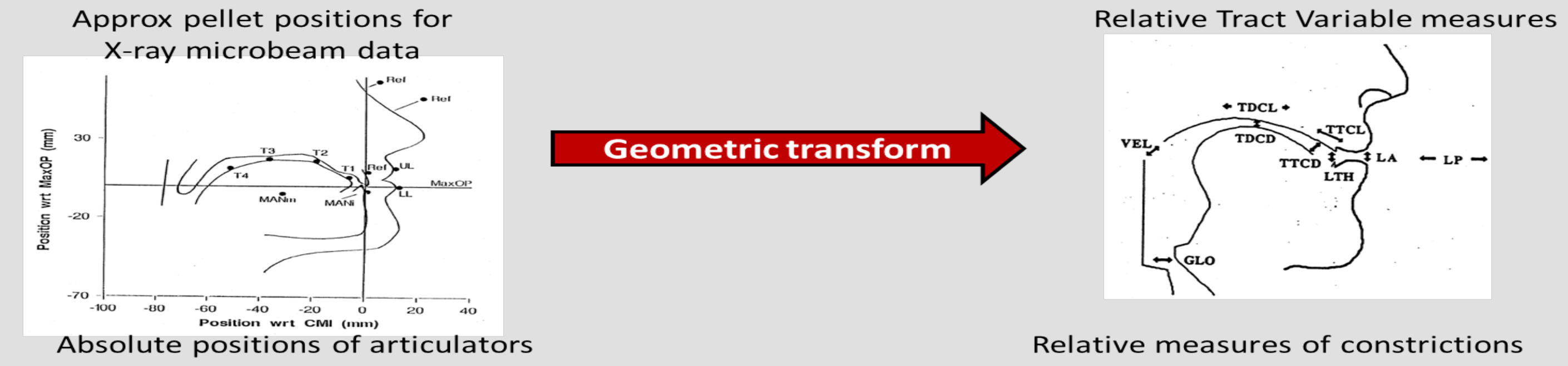


Fig 2

- Since there is no database that contains speech annotated withTVs, initial work focused on synthetic data generated by TADA [2].
- However, synthetic speech and the TADA model do not capture all of the variability observed in natural speech. Hence, our current focus is on natural articulatory data.
- Articulatory data from the X-ray microbeam database (XRMB) [3] consists of X-Y positions of pellets placed along the vocal tract which depend on the physiology of the speakers.
- We convert these absolute positions into TVs as outlined in Fig. 3.



Absolute positions of articulators dependent on speaker dimensions

Pellet positions

UL	Upper lip	UL_x, TL_y
LL	Lower lip	LL_x, UL_y
T1	Tongue tip	TT_x, TT_y
T3	Tongue body	TB_x, TB_y
T4	Tongue dorsum	TD_x, TB_y
MANi	Jaw	MANi_x, MANi_y

Geometric transform

Relative measures of constrictions independent of speaker dimensions

Tract Variables (TVs)

LA	Lip Aperture
LP	Lip protrusion
TBCL	Tongue body constriction location
TBCD	Tongue body constriction degree
TTCL	Tongue Tip constriction location
TTCD	Tongue Tip constriction degree

Fig 3

Speech inversion: From acoustics to articulations

- The amount of simultaneous acoustic and articulatory data is very limited and not easy to obtain. It is therefore essential to build models to estimate TVs from acoustics. These models are referred to as speech inversion systems.
- Multi layer feed forward neural networks were trained to estimate the TVs from contextualized Mel-frequency cepstral coefficients (MFCCs). We trained two such TV estimators using (1) synthetic XRMB data (generated from TADA) and (2) natural XRMB data.
- The correlation between estimated and groundtruth TVs for the natural and synthetic TV estimators are shown in Table 1.
- The comparatively poorer correlation results of the natural TV estimator is due to the challenging variability observed in natural speech.

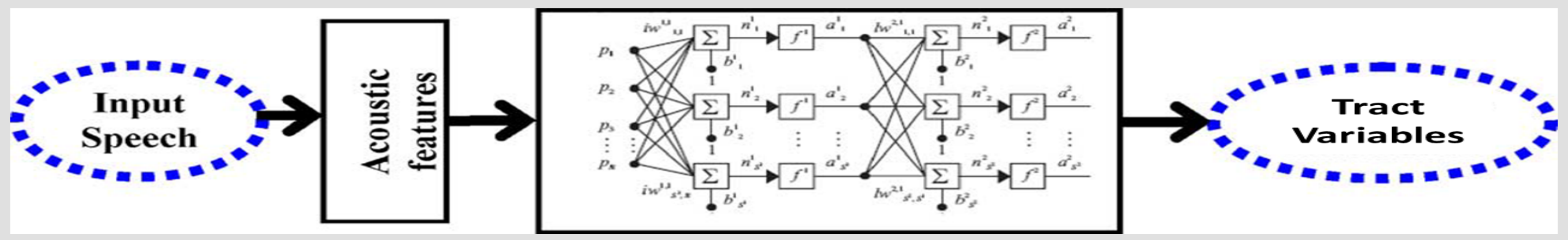


Fig 4

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

Tract Variables	LA	LP	TBCL	TBCD	TTCL	TTCD
Synthetic TV estimator	0.89	0.92	0.94	0.91	0.89	0.92
Natural TV estimator	0.66	0.56	0.78	0.59	0.65	0.76

- 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 examples of fast and normal rate speech analyzed by the natural TV estimator. Note that the TV estimator was not trained and tested on speech from the same talker. Also, note that the synthetic TV estimator did not perform nearly as well as the natural TV estimator.

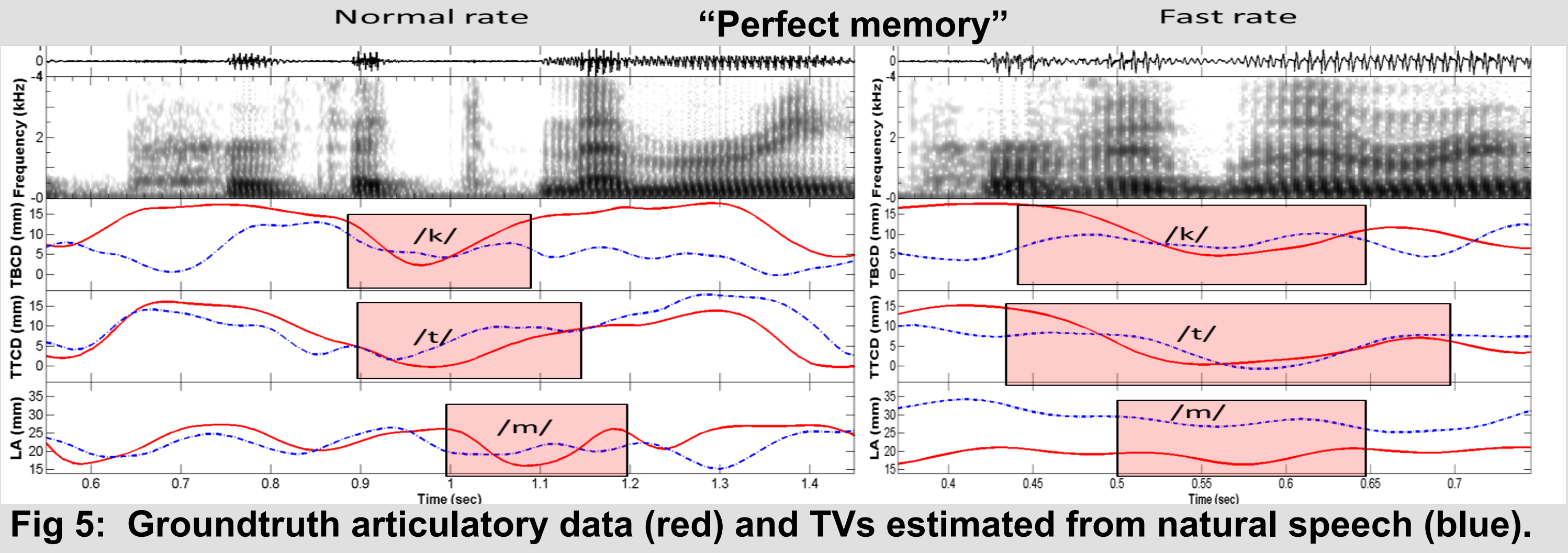


Fig 5: Groundtruth articulatory data (red) and TVs estimated from natural speech (blue).

Impact on speech recognition

- It is a challenging task to use acoustic features to recognize the place of articulation of stop consonants: labial (/p,b/), alveolar (/t,d/) or velar (/k,g/).
- For the task of stop consonant place of articulation classification, estimated TVs along with MFCCs provided accuracy close to the groundtruth TVs.
- We extracted statistical measures from the TVs and MFCCs and used them as features for the classification task. Table 2 shows the results of this experiment.

Table2: Average accuracy of place of articulation classification of stop consonants

	No Context	20ms context
Groundtruth_TV_statfeat	93.91%	93.84%
MFCC_statfeat	89.00%	91.04%
XRMB_TV_statfeat	83.40%	84.07%
{MFCC+XRMB_TV}_statfeat	**91.19%**	**92.54%**

- Table 3 shows the results of a small vocabulary keyword recognizer [4]
- 34 speakers, 6 word sequences. 6 different noise conditions

Table3: Keyword recognition accuracy on the CHiME 2 small vocabulary speech recognition task

Features	-6dB	-3dB	0dB	3dB	6dB	9dB	Avg
Baseline MFCC (39D) [HTK-MFCC]	49.33	58.67	67.50	75.08	78.83	82.92	68.72
MFCC+ModTV_pca (42D)	53.67	63.67	72.50	79.83	84.00	87.33	73.50

This work was supported by:
NSF Grant # IIS-1162525
NSF Grant # BCS-1436600

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