ANALYSIS OF ACOUSTIC TO ARTICULATORY SPEECH INVERSION FOR NATURAL SPEECH

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

[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

Abstract

- The objective of this research is to train efficient acoustic to articulatory speech inversion systems on natural speech that provide reliable articulatory features for unseen test speakers.
- We constructed two such systems using feedforward neural networks. One was trained using natural speech data from the XRMB database and the second using synthetic data generated by the Haskins Laboratories TADA model that approximated the XRMB data.
- XRMB pellet trajectories were first converted into vocal tract constriction variables (TVs), providing a relative measure of constriction kinematics (location and degree).
- TV-estimators were tested using previously collected acoustic data on the utterance "perfect memory" spoken at slow, normal, and fast rates.
- The TV estimator trained on XRMB data (but not on TADA data) was able to recover the tongue tip gesture for /t/ in the fast utterance despite the gesture occurring partly during the acoustic silence of the closure.
- The XRMB system (but not the TADA system) could **distinguish between bunched and retroflexed /r/.**
- Speaker dependent TV estimators were trained an tested on matched and mismatched speaker conditions.

The X-ray Microbeam database

- The X-ray microbeam (XRMB) database [1] consists of continuous speech along with time aligned coordinates of pellets placed at various points along the vocal tract as sown in Fig, 1.

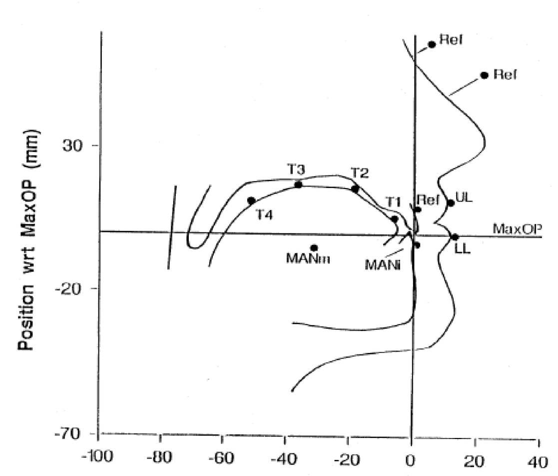


Fig 1

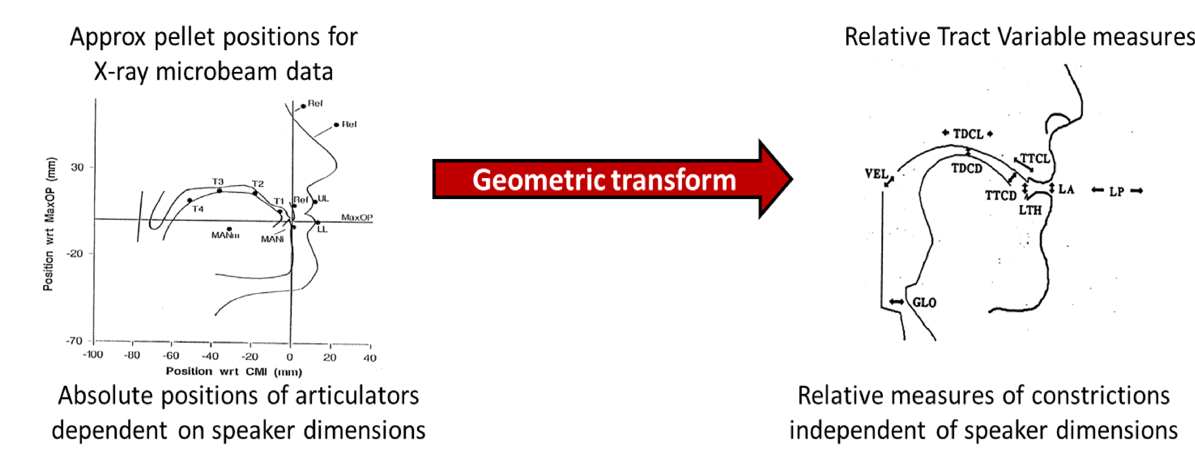
- The database consists of 2500 speech utterances with pellet trajectories across 57 different speakers.
- The database provides X and Y coordinates of the pellets as the subject produces the desired utterances.
- However, these pellet positions are closely connected to speaker anatomy and head positions leading to cross speaker variability.
- A total of 1720 utterances across 46 speakers were successfully transformed to TVs. This formed the natural XRMB database

XRMB pellet trajectories to Tract 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.
- The pellet trajectories from XRMB data are geometrically transformed into relative TV measures.
- The XRMB data after transformation is represented in terms of six TVs as shown in Fig. 2.

Synthetic XRMB database with synthetic TVs from TADA

- The Task Dynamics and Applications (TADA) system [2] from Haskins laboratories along with HLSyn [3] was used to generate synthetic speech along with time aligned TVs for the XRMB sentences.
- The synthetic speech was then warped to align with the natural XRMB utterances. The same warping function was used to warp the synthetic TVs.
- The details of this synthetic XRMB data generation are given in [4].



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, TD_y
MANi	Jaw	MANi_x, MANi_y

Geometric transform

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 2

Speech Inversion systems

- Multi layer feed forward neural networks were trained to estimated the TVs from contextualized MFCCs.
- We trained two such TV estimators – (1) on synthetic XRMB data, (2) on natural XRMB data.
- The correlation between estimated and groundtruth TVs for the natural and synthetic TV estimators are shown in Table 1.

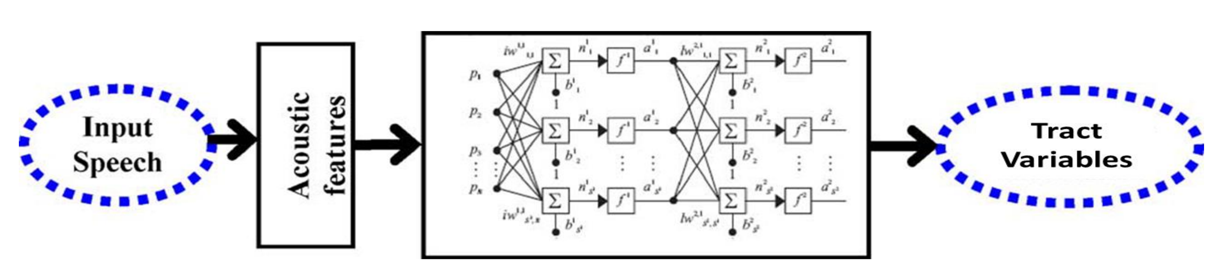


Fig 3

- The correlation between estimated and groundtruth TVs for the synthetic TV estimator are much better than that of the natural TV estimator because the **synthetic speech is single speaker with very less variability in production.**

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

Tract Variables	Synthetic TV estimator	Natural speech TV estimator
LA	0.8881	0.7101
LP	0.9150	0.5721
TBCL	0.9419	0.6999
TBCD	0.9086	0.5630
TTCL	0.8876	0.5944
TTCD	0.9223	0.7303

Uncovering coarticulation using TV estimators

- Speaker dependent TV estimators were trained using one of the speakers' data from the XRMB database.
- The speaker dependent TV estimator was used to obtain the TVs for the utterance "perfect memory" for the two cases – (1) clearly spoken (2) fast spoken
- Figure 5 shows the estimated TVs (LA, TBCD, TTCD) for clearly spoken utterance and Figure 6 shows the same TVs for fast spoken utterance.

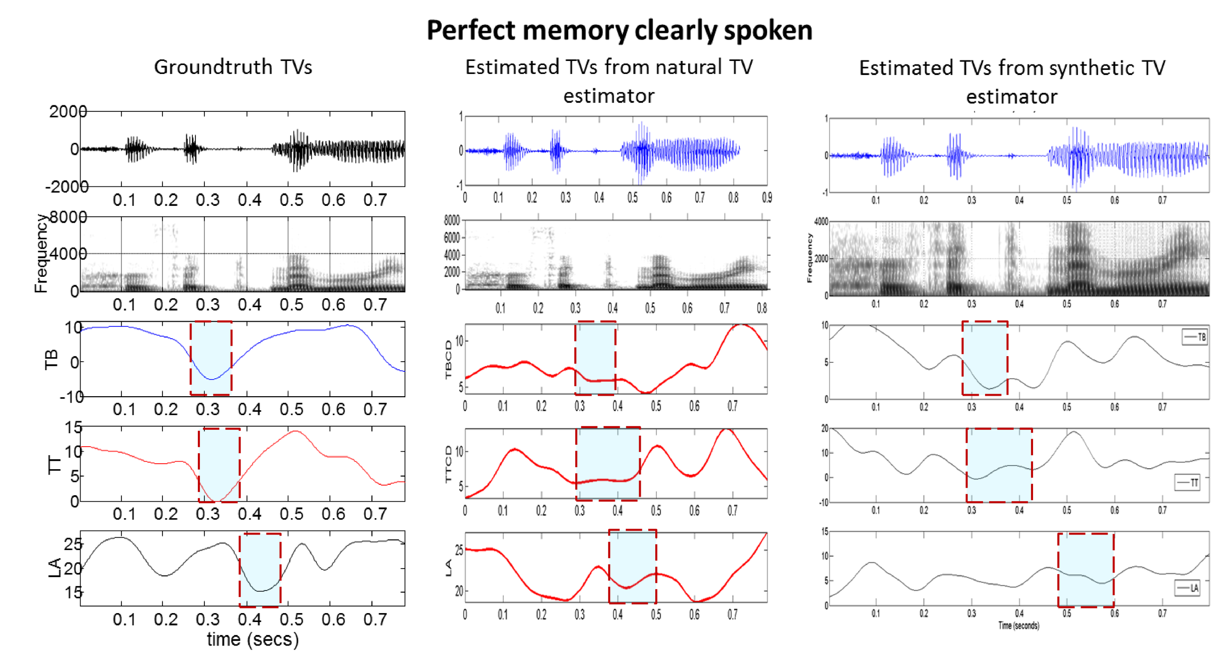


Fig 4

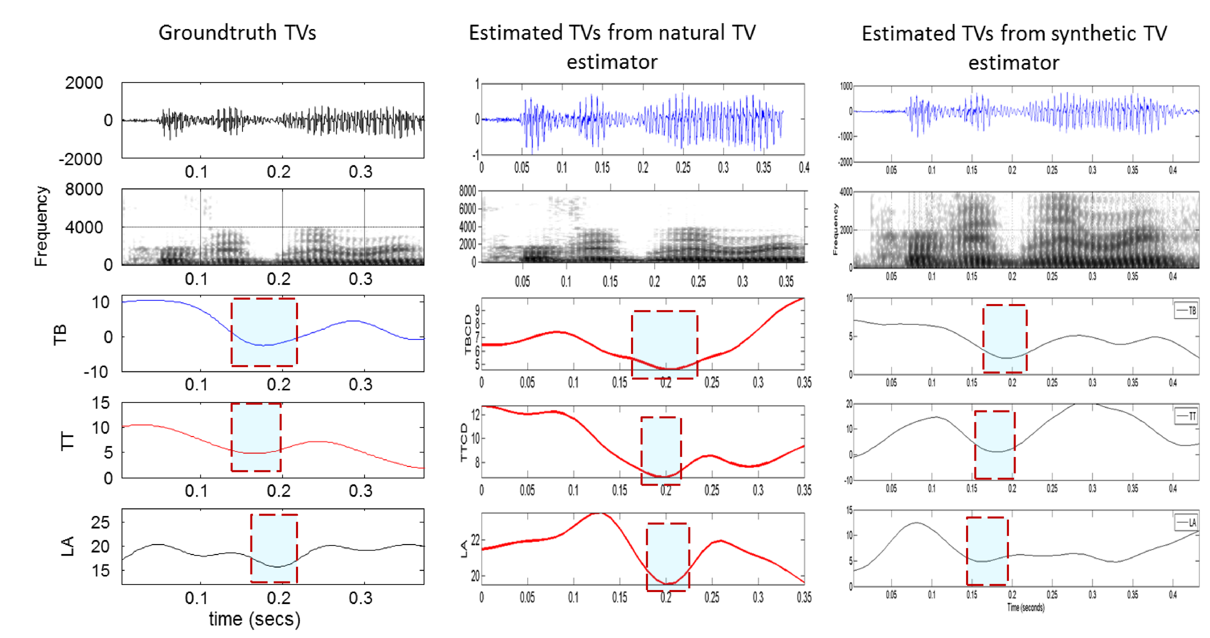


Fig 5

Uncovering bunched and retroflexed /r/ using the TV estimator

- We used the TV estimator to analyze utterances containing bunched and retroflexed production of the /r/ sound as determined from MRI [5].
- Figure 6 shows the tongue positions for the bunched and the retroflexed productions of /r/
- As seen in Figure 7, the natural TV estimator correctly uncovers the tongue tip and tongue body constrictions for the two types of /r/ productions.

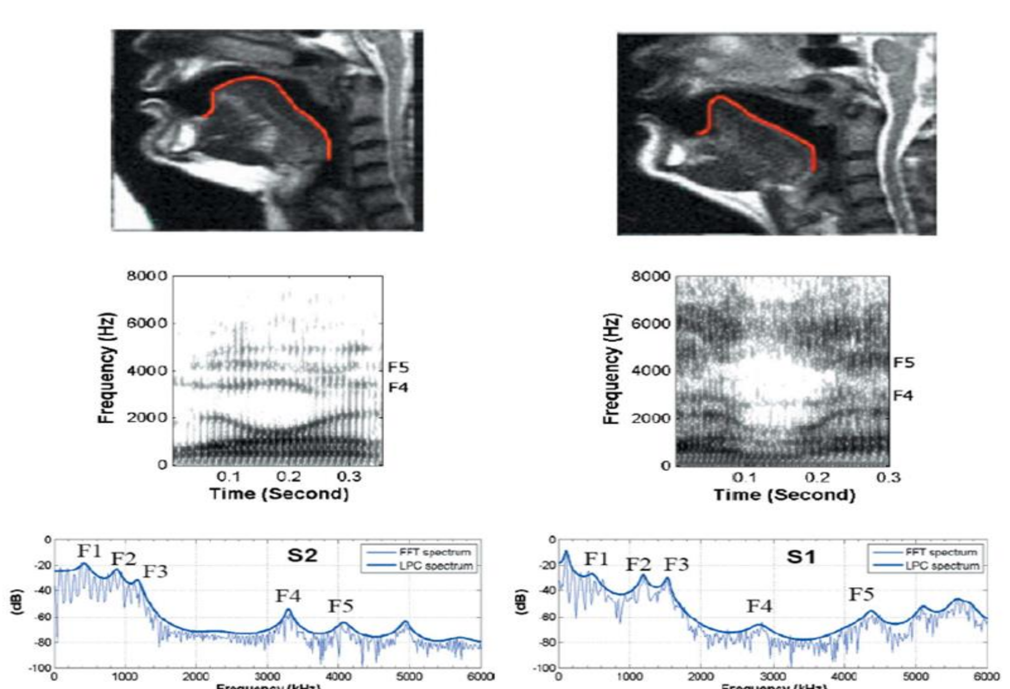


Fig 6

- In [5], the authors show that the spacing between F5 and F4 is an acoustic signature for tongue shape. F5-F4 for retroflex /r/s is around 1340Hz And for bunched /r/s is around 740Hz
- We low pass filtered the speech to remove the F4 and F5 frequencies from the spectrum. We then passed the filtered speech through the TV estimator.
- We can see from figure 8 that the TV estimator does not show the correct tongue constrictions for the bunched and retroflexed /r/, providing further evidence that F5 and F4 information provides the distinction between bunched and retroflexed /r/.

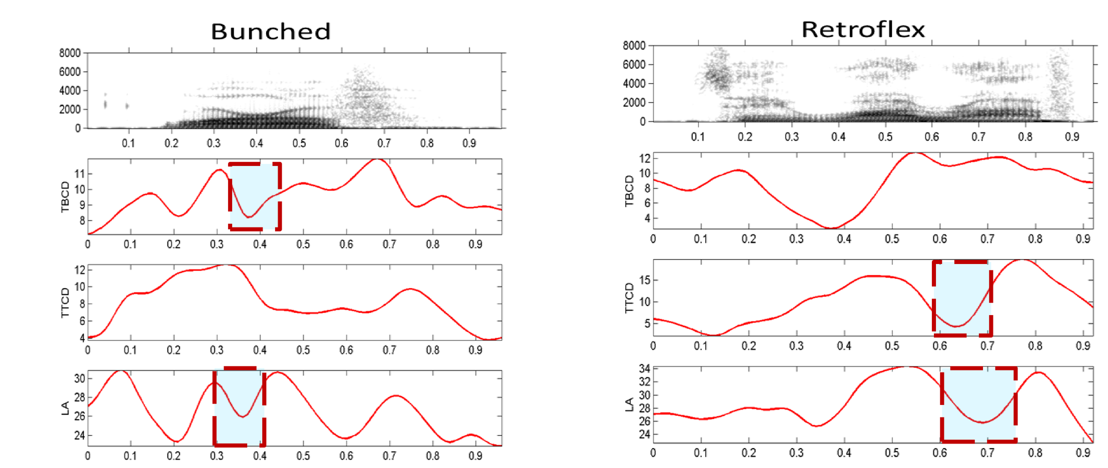


Fig 7

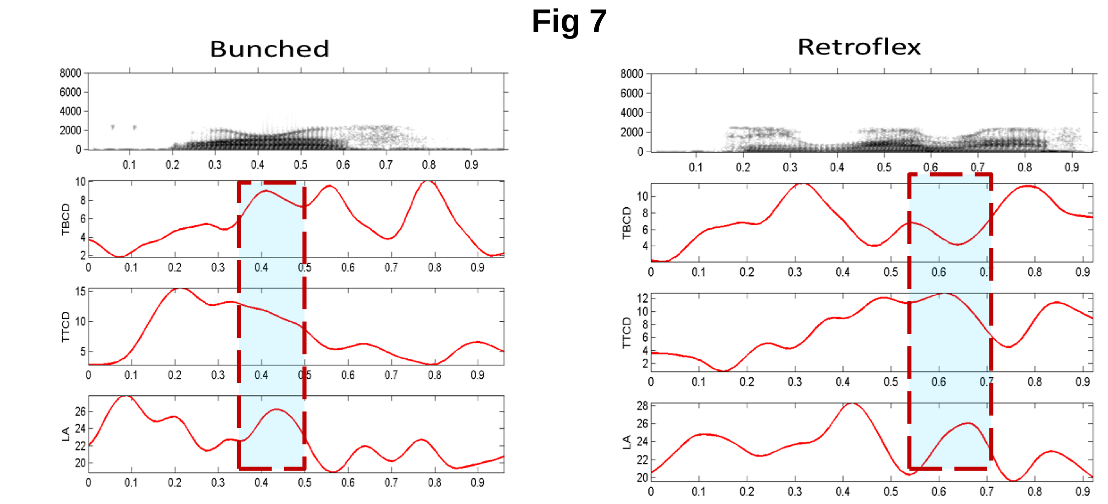


Fig 8

Analysis of Speaker dependent TV estimators

- 10 speakers containing at least 40 utterances were selected from the XRMB database (5 males & 5 females) to perform this experiment.
- Speaker dependent TV estimators were trained on each of the 10 speakers.
- Each TV estimator was tested in the matched speaker and mismatched speaker conditions.
- As a crude approximation to speaker normalization, each speaker's MFCCs and TVs were globally normalized using the mean and variance across all speakers.
- The results of these experiments are shown in Table 2.

Table 2: ***Results of matched and mismatched tests on speaker dependent TV estimators***

Normalization scheme	Test condition	LA	LP	TBCL	TBCD	TTCL	TTCD	Mean correlation
Independently normalized MFCC and TVs	Mismatched	0.2164	0.3565	0.5741	0.2555	0.2591	0.5194	0.3635
	Matched	0.6193	0.6728	0.8372	0.6880	0.7547	0.7935	0.7276
Independently normalized MFCC and globally normalized TVs	**Mismatched**	**0.2330**	**0.3703**	**0.5779**	**0.2702**	**0.2547**	**0.5140**	**0.3700**
	Matched	**0.6208**	**0.6861**	**0.8377**	**0.6977**	**0.7523**	**0.7946**	**0.7315**
Globally normalized MFCC and globally normalized TVs	Mismatched	0.2214	0.3751	0.5878	0.2458	0.2611	0.5320	0.3705
	Matched	0.5701	0.6694	0.8254	0.6852	0.7417	0.7709	0.7104
Speaker Independent estimator		0.8222	0.5748	0.7301	0.5564	0.5999	0.7323	0.6693

Conclusions

- The TV estimators trained on natural speech successfully uncover the hidden gestures in coarticulation for fast spoken speech.
- The difference between bunched and retroflexed /r/ are clearly shown by the TV estimator in spite of the fact that the utterances came from speakers unseen by the TV estimator.
- There is still a considerable amount of variability across speakers as indicated by the poor mismatched test correlations of the TV estimators.

Future directions

- Explore phone class specific TV estimators for natural speech
- Explore acoustic and articulatory variability at the phone level across different phone contexts.
- Explore speaker normalization schemes in the acoustic and articulatory domain.

References

[1] J. R. Westbury, "Microbeam Speech Production Database User''s Handbook," IEEE Pers. Commun. - IEEE Pers. Commun., 1994.

[2] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," Ecol. Psychol., vol. 1, no. 4, pp. 333–382, Dec. 1989.

[3] H. M. Hanson and K. N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn.," J. Acoust. Soc. Am., vol. 112, no. 3 Pt 1, pp. 1158–82, Sep. 2002.

[4] 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.

[5] Zhou, X. H., Espy-Wilson, C. Y., Boyce, S., Tiede, M., Holland, C., and Choe, A. (2008). "A magnetic resonance imaging-based articulatory and acoustic study of 'retroflex' and 'bunched' American English /r/," J. Acoust. Soc. Am. 123, 4466–4481.

Acknowledgement: This research was supported by NSF Grant #IIS-1162046