Analysis of coarticulated speech using estimated articulatory trajectories

Ganesh Sivaraman[1], Vikramjit Mitra[2], Mark Tiede[3], Elliot Saltzman[4], Louis Goldstein[5], Carol Espy-Wilson[1]

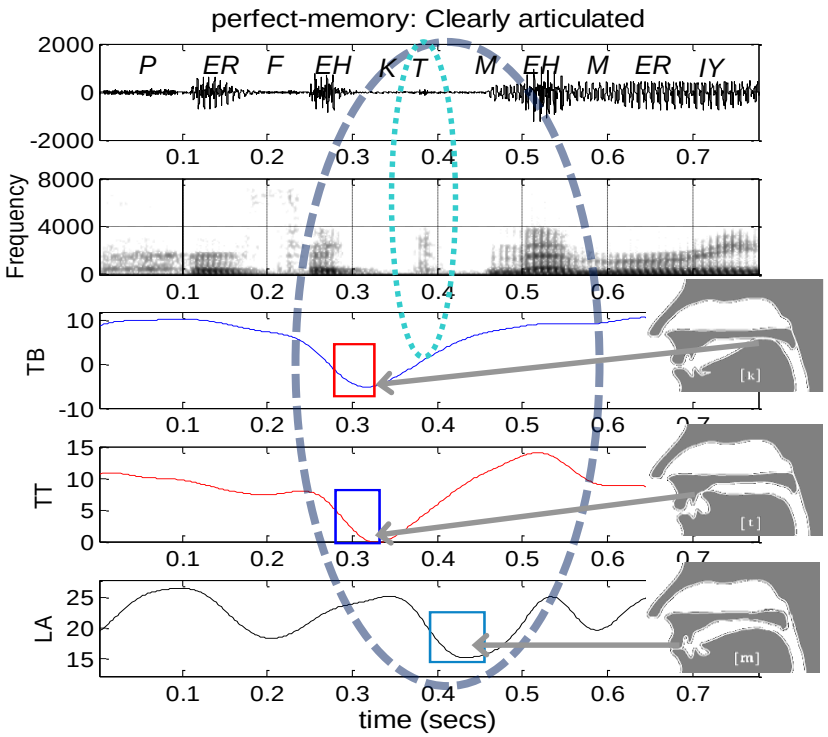
[1]University of Maryland College Park, MD,
[2]SRI International, Menlo Park, CA,
[3]Haskins Laboratories, New Haven, CT,
[4]Boston University, Boston, MA,
[5]University of Southern California, Los Angeles, CA

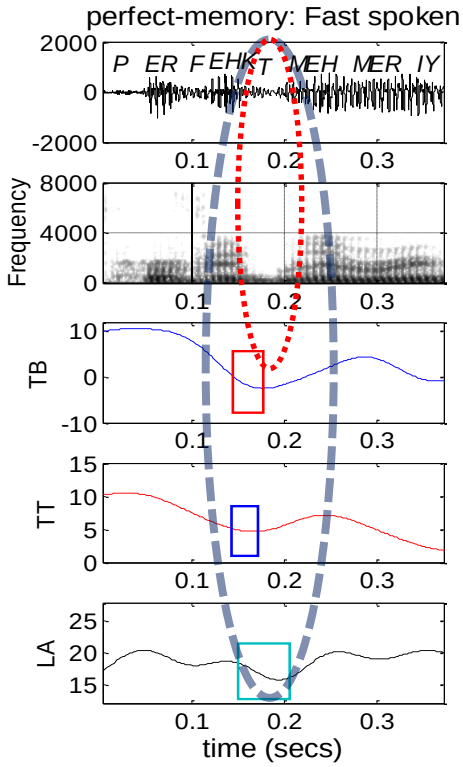
Overview

- Coarticulation: A significant source of speech variability
- Articulatory data collection
- Conversion of articulatory data to Tract Variables (TVs)
- Speech inversion
- Examples of estimated TVs for coarticulated utterances
- Discussions and Future directions

Coarticulation: A significant source of speech variability

- Coarticulation is the overlap of vocal tract gestures of one sound with that of another leading to variability in acoustics





State-of-the-art ASR systems on fast speech

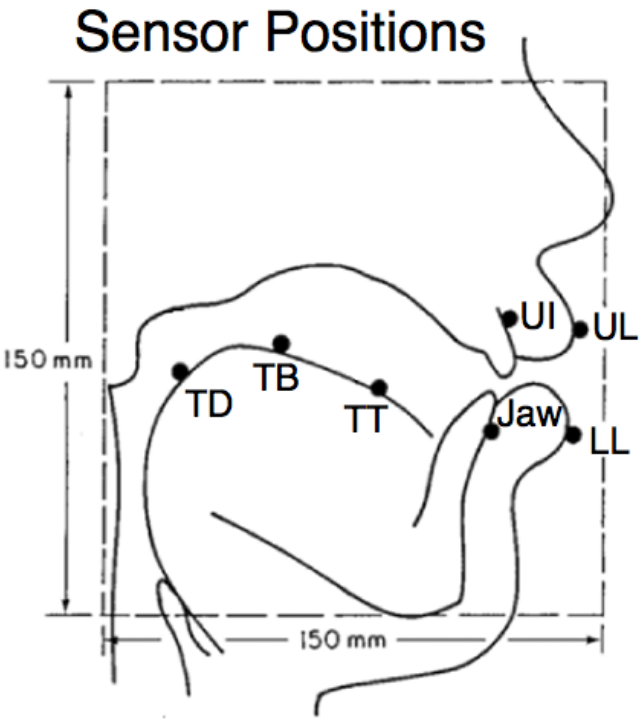
SRI International's DECIPHER® Speech Recognition system
GMM-HMM model with 4-gram phone based language model

SRI's conversational speech recognition
Trained on conversational speech with the SRI Language Model

Ground Truth	SRI phonetic recognizer	SRI word recognition
flask stood (normal production)	f l ae s k dh ey d	the empty **flasks hidden** under tinge tray
flask stood (fast production)	f l ae s t uh d	the empty **flustered** and that the tin tray
workman's (normal production)	w er m eh n d	the beam jot down on the **work manned** had
workman's (fast production)	v er b ih n t	they've been cut down on the **work been** type
perfect memory (normal production)	p er f ih k ah m er iy	she had a **perfect memory** for details
perfect memory (fast production)	v er g eh r iy	share a **part of the river** details

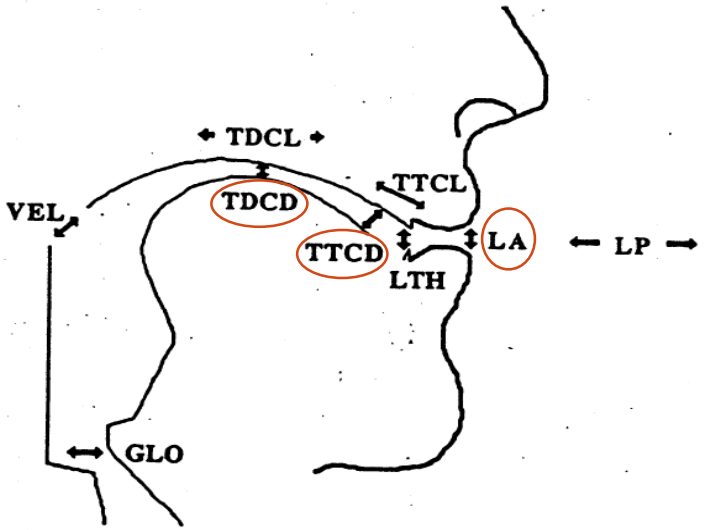
Articulatory data collection: The EMA-IEEE dataset

- **Corpus:** 720 IEEE Sentences (Harvard sentences)
- **Subject:** Female native speaker of American English in her mid-twenties.
- **Speaking rate:** Normal and Fast rates
  Normal rate of approximately 2.9 syllables/sec.
  Fast rate approximately 20% faster.
- **Instrument:** WAVE EMA system (Northern Digital)
- **Sensor placements:** As show in the figure



Tract Variables: What are they?

- Tract Variables (TVs) [Browman & Goldstein,1992] are measures of constriction degree and location executed by articulators in the vocal tract.
- Idea drawn from the Task Dynamics and Applications model (TADA) [Nam et al. (2004)] of speech production

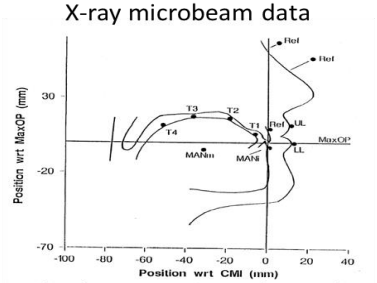


- Browman, C. and Goldstein, L. (1992) "Articulatory Phonology: An Overview", Phonetica, 49: 155-180.
- Nam, H., Goldstein, L., Saltzman, E. and Byrd, D. (2004) "Tada: An enhanced, portable task dynamics model in matlab", J. Acoust. Soc. of Am., **115**(**5**), 2, pp. 2430.

Why convert to TVs?

- TVs are a more speaker independent representation than pellet positions.
- Use the TADA theoretical framework to analyze the phonological phenomena.

Pellet positions for
X-ray microbeam data



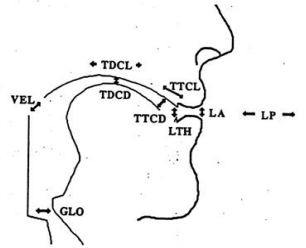
Absolute positions of articulators
dependent on speaker dimensions

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 Tract Variable measures

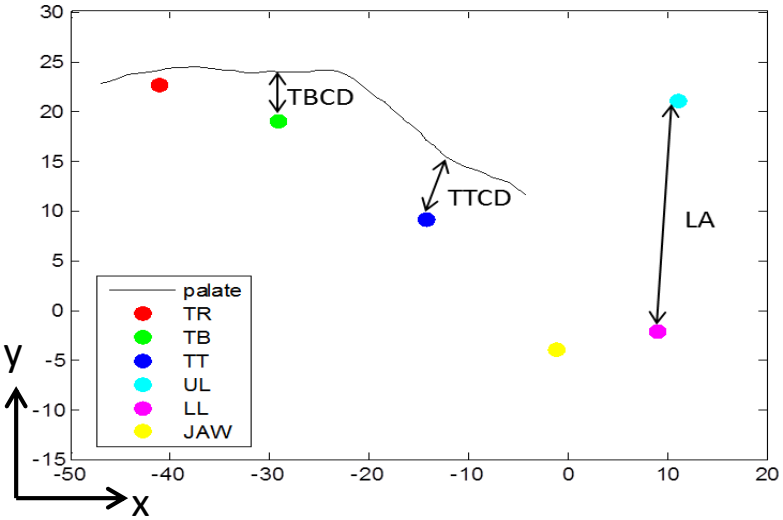


Relative measures of constrictions
independent of speaker dimensions

Tract Variables (TVs)

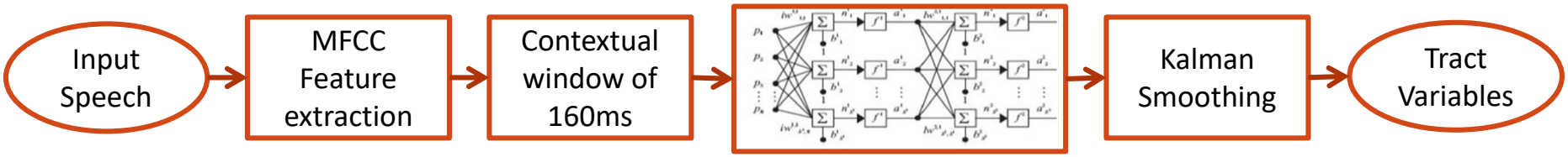
LA	Lip Aperture
LP	Lip protrusion
TBCL	Tongue body constriction location
TBCD	Tongue body constriction degree
TTCL	Tongue Tip constriction location
TTCD	Tongue Tip constriction degree

Conversion of articulatory data to Tract Variables (TVs)



- Geometric transformations
- $LA=(ULx-LLx)^2+(ULy-LLy)^2+(ULz-LLz)^2$
- $TBCD=Min\{Distance(TB,palate)\}$
- $TTCD= Min\{Distance(TT,palate)\}$

Speech inversion



- Function mapping approach to speech inversion
- Artificial neural networks (ANN) suitable for the highly non-linear and non-unique mapping from acoustics to TVs [V.Mitra et al. 2010]
- Input features: Contextualized MFCCs (13 coeffs x 17 frames)
- Outputs: 6 TVs (LA, LP, TBCL, TBCD, TTCL, TTCD)
  - 3 TVs for EMA-IEEE dataset.
- Single Hidden layer networks
- 100 – 500 nodes in hidden layer
- Scaled conjugate gradient algorithm for training.

- V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving Tract Variables From Acoustics: A Comparison of Different Machine Learning Strategies.," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 6, pp. 1027–1045, Sep. 2010.

Various speech inversion systems

- The XRMB data *[J. R. Westbury 1994]* and the EMA-IEEE data were used to create 4 different speech inversion systems
- Speaker dependent systems were also trained for each speaker in the XRMD dataset

TV estimator name	Training dataset	No. of speakers	Hours of training data
X_NORM	XRMB utterances converted to TVs using an algorithm outlined in [H. Nam, V. Mitra et.al. 2012]	46	4 hours
XF_NORM	Female speakers' utterances from XRMB database converted to TVs	25	2.42 hours
XM_NORM	Male speakers' utterances from XRMB database converted to TVs	21	1.55 hours
E_IEEE	Single female speaker EMA data converted to TVs	1	1.03 hours

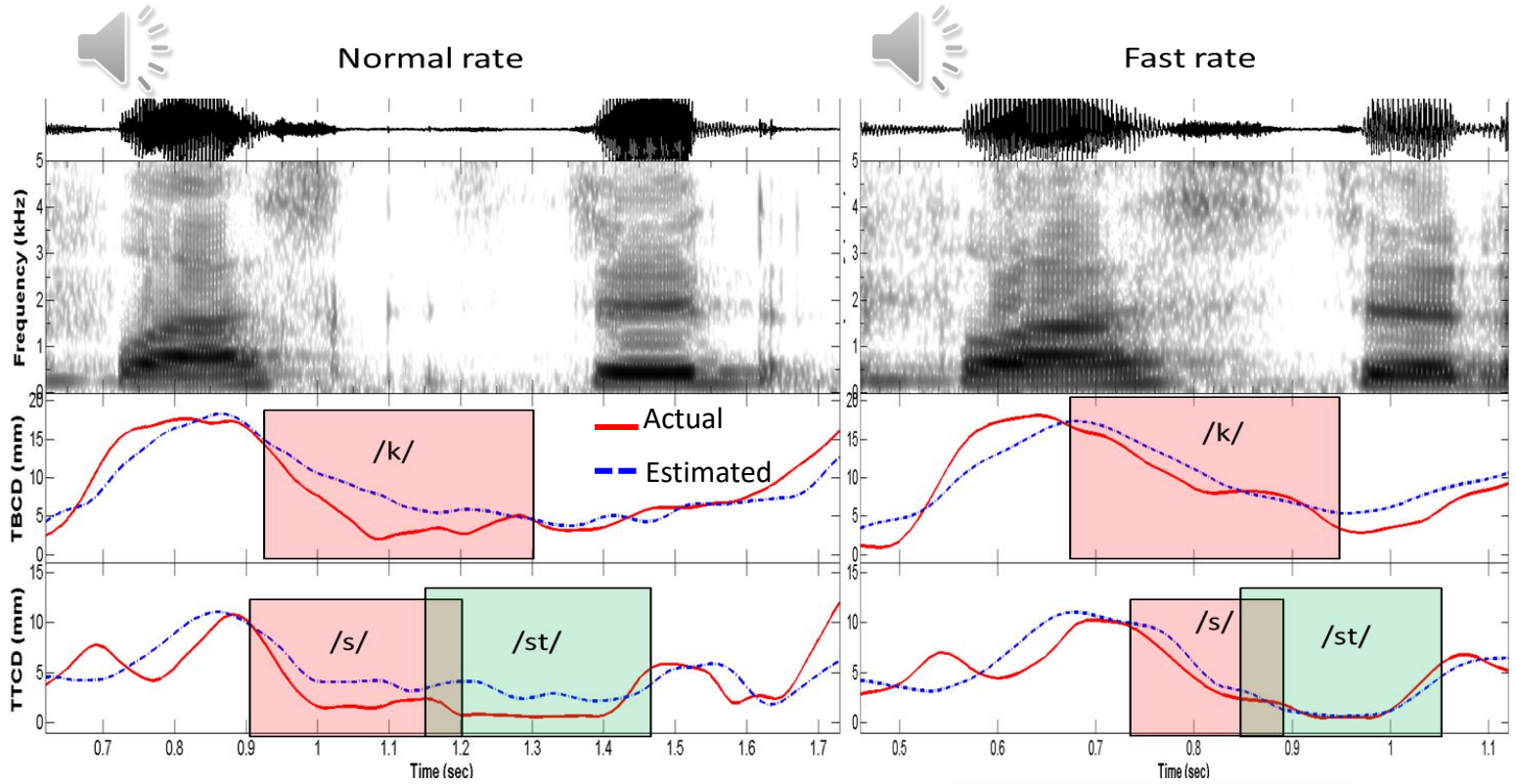
- The Pearson Product Moment Correlation (PPMC) between actual and estimated TVs for the test set was used to evaluate the trained systems.
- PPMC varies from -1 to 1. A value of 1 signifies perfect correlation

TV estimator name	LA	TBCD	TTCD	LP	TBCL	TTCL
X_NORM	0.66	0.59	0.76	0.56	0.78	0.65
XF_NORM	0.72	0.66	0.79	0.62	0.82	0.66
XM_NORM	0.68	0.64	0.78	0.57	0.83	0.72
E_IEEE	0.64	0.80	0.72	NA	NA	NA

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

Example 1

- **“flask stood” -** The empty **flask stood** on the tin tray

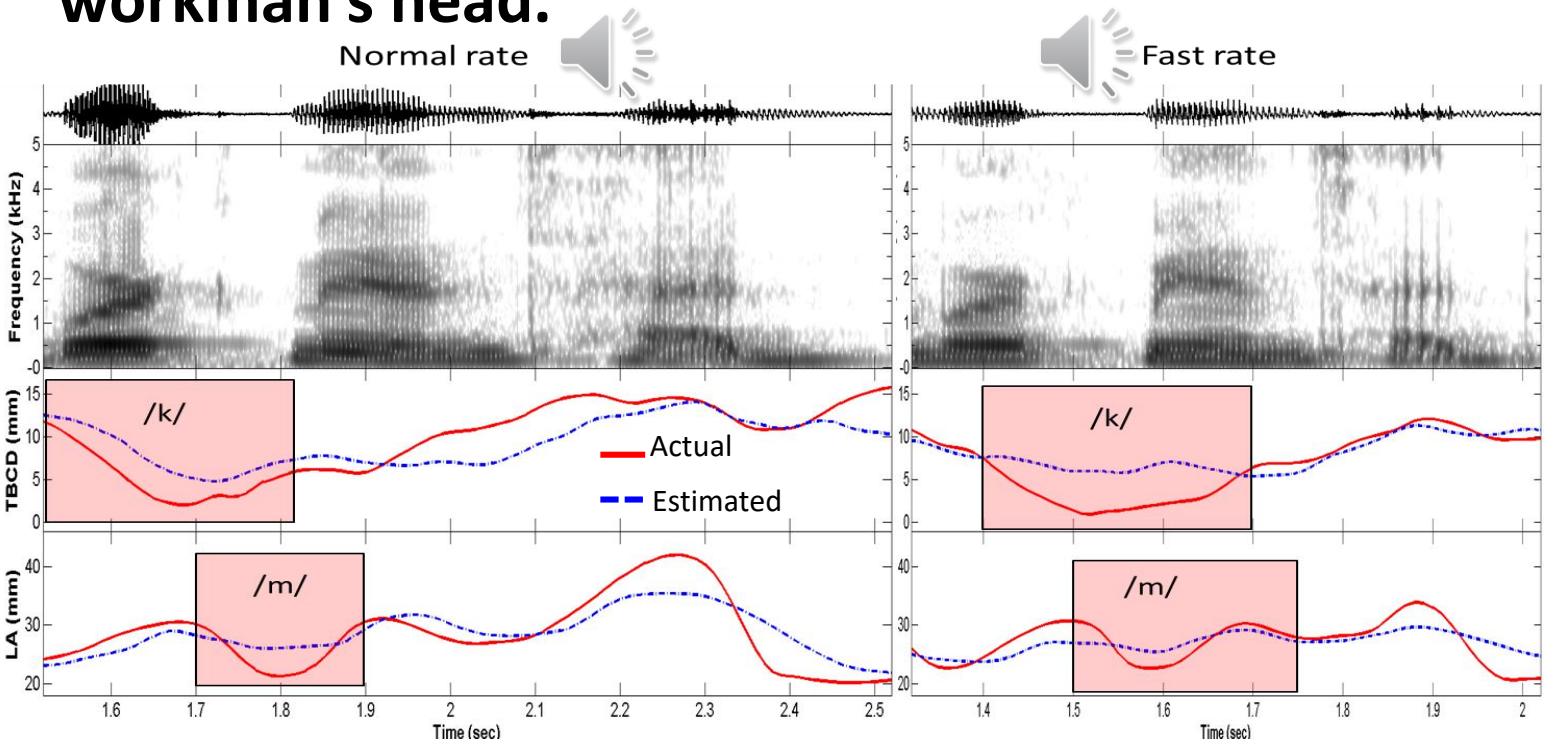


TV estimator name	flask stood	
	normal	fast
X_NORM	0.56	0.59
XF_NORM	0.56	0.59
XM_NORM	0.56	0.59
E_IEEE	**0.86**	**0.82**

	flask stood	
E_IEEE	fast	normal
LA	0.85	0.77
TBCD	0.87	0.86
TTCD	0.85	0.83
Average	0.86	0.82

Example 2

- "**workman's head**" - The beam dropped down on the **workman's head.**

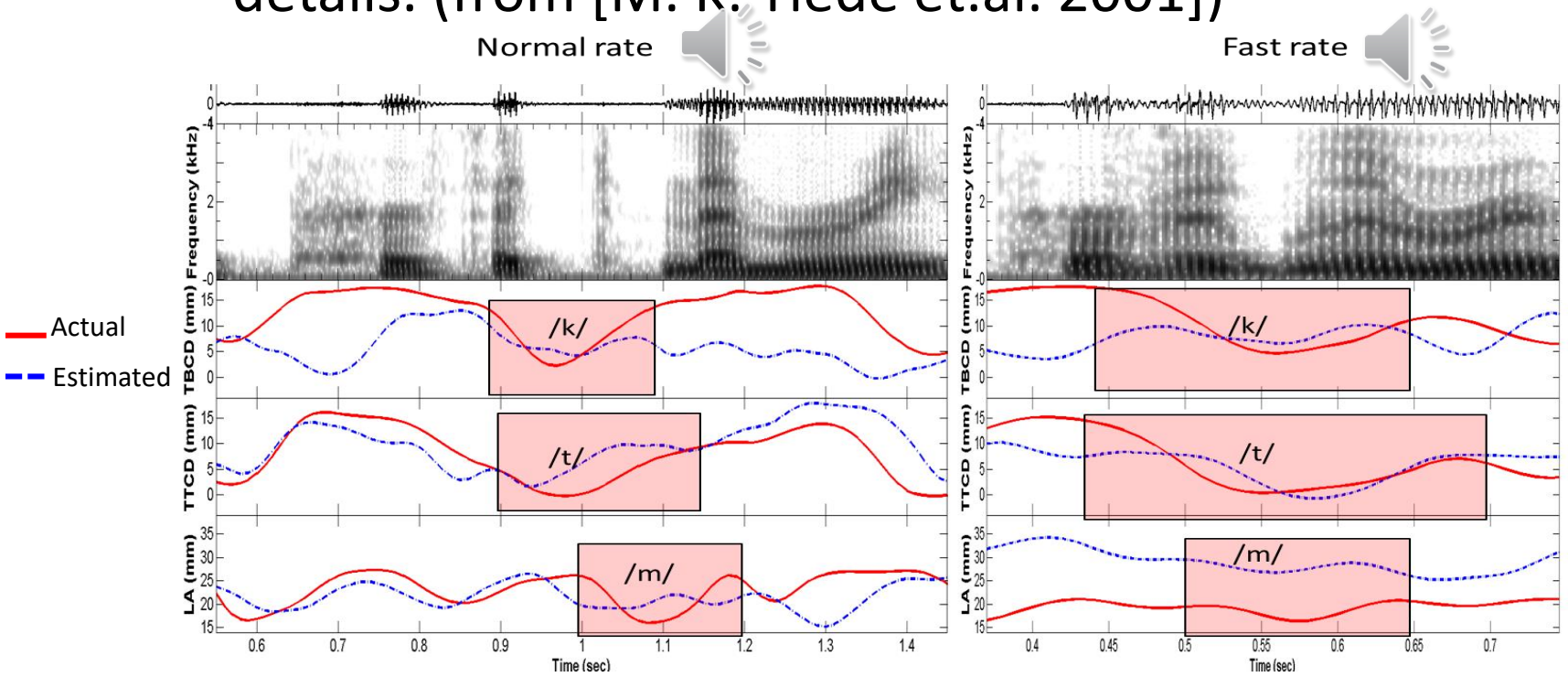


TV estimator name	Workman's head	
	normal	fast
X_NORM	0.61	0.75
XF_NORM	0.55	0.72
XM_NORM	0.59	0.63
E_IEEE	**0.75**	**0.79**

	Workman's head	
E_IEEE	normal	fast
LA	0.68	0.77
TBCD	0.88	0.81
TTCD	0.68	0.78
Average	0.75	0.79

Example 3

- **“perfect memory”** - She had a **perfect memory** for details. (from [M. K. Tiede et.al. 2001])



TV estimator name	Perfect memory	
	normal	fast
X_NORM	0.40	0.51
XF_NORM	0.28	0.55
XM_NORM	0.44	0.58
E_IEEE	0.18	0.44

	JW29	JW28
E_IEEE	normal	fast
LA	0.60	0.57
TBCD	0.52	0.32
TTCD	0.57	0.57
Average	0.56	0.48

Conclusion, Discussions and Future directions

- Speech inversion systems, can reliably uncover hidden or coarticulated gestures.
- Speaker dependent TV estimators are more accurate than speaker independent estimators. A **speaker normalization scheme** needs to be implemented in order to efficiently use articulatory data from different sources.
- A well defined scheme is needed to convert pellet data to TVs
- Data from more speakers needs to be collected for a thorough analysis of coarticulation in fast spoken speech.

Questions? Comments?