PROSODIC VARIATION WITHIN SPEECH PLANNING AND EXECUTION – INSIGHTS FROM REAL-TIME MRI

Vikram Ramanarayanan*, Dani Byrd^, Louis Goldstein^ and Shrikanth S. Narayanan*^

*Speech Analysis and Interpretation Laboratory, Ming Hsieh Department of Electrical Engineering

^Department of Linguistics, University of Southern California, Los Angeles, CA

`<vramanar,dbyrd,louisgol>@usc.edu,shri@sipi.usc.edu`

[Supported by NIH]

MOTIVATIONS

Some interesting planning- and execution-related research questions:

I. How does the cognitive load on the speech planner vary as speaking style becomes more informal?

II. How does the articulatory execution of this plan vary, and what are its acoustic consequences?

III. How can this understanding be applied to speech technology domains?

FORMULATION

Try to look at the problem from point of view of prosodic planning

I. SHAPING:

Quantify (constriction-forming events) for different speaking styles?

II. VARIABILITY:

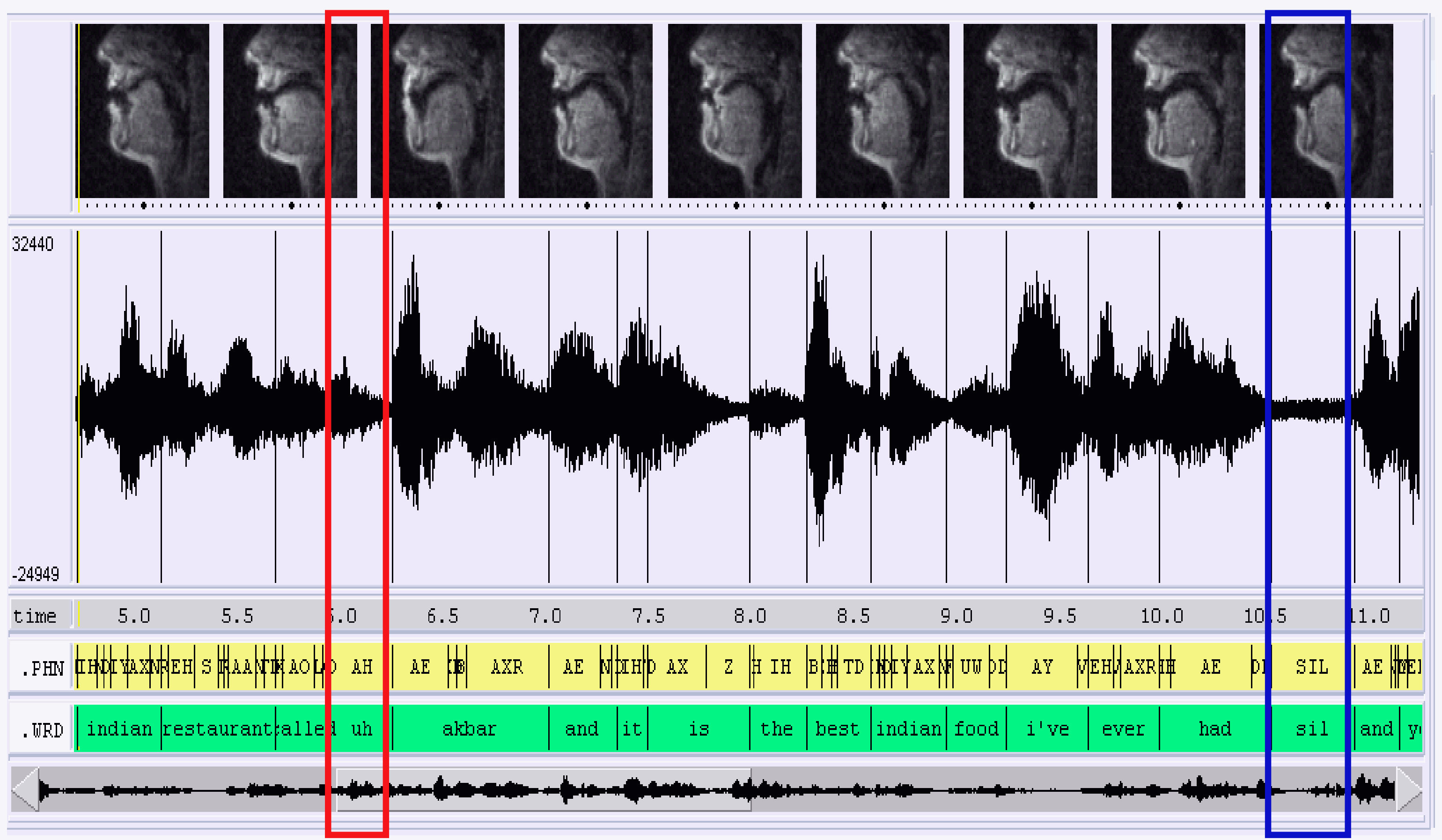
Reduction patterns in VCV sequences in read & spontaneous speech?

III. KINEMATICS:

Differences in timing and speed of critical articulators?

THEORETICAL FOUNDATIONS

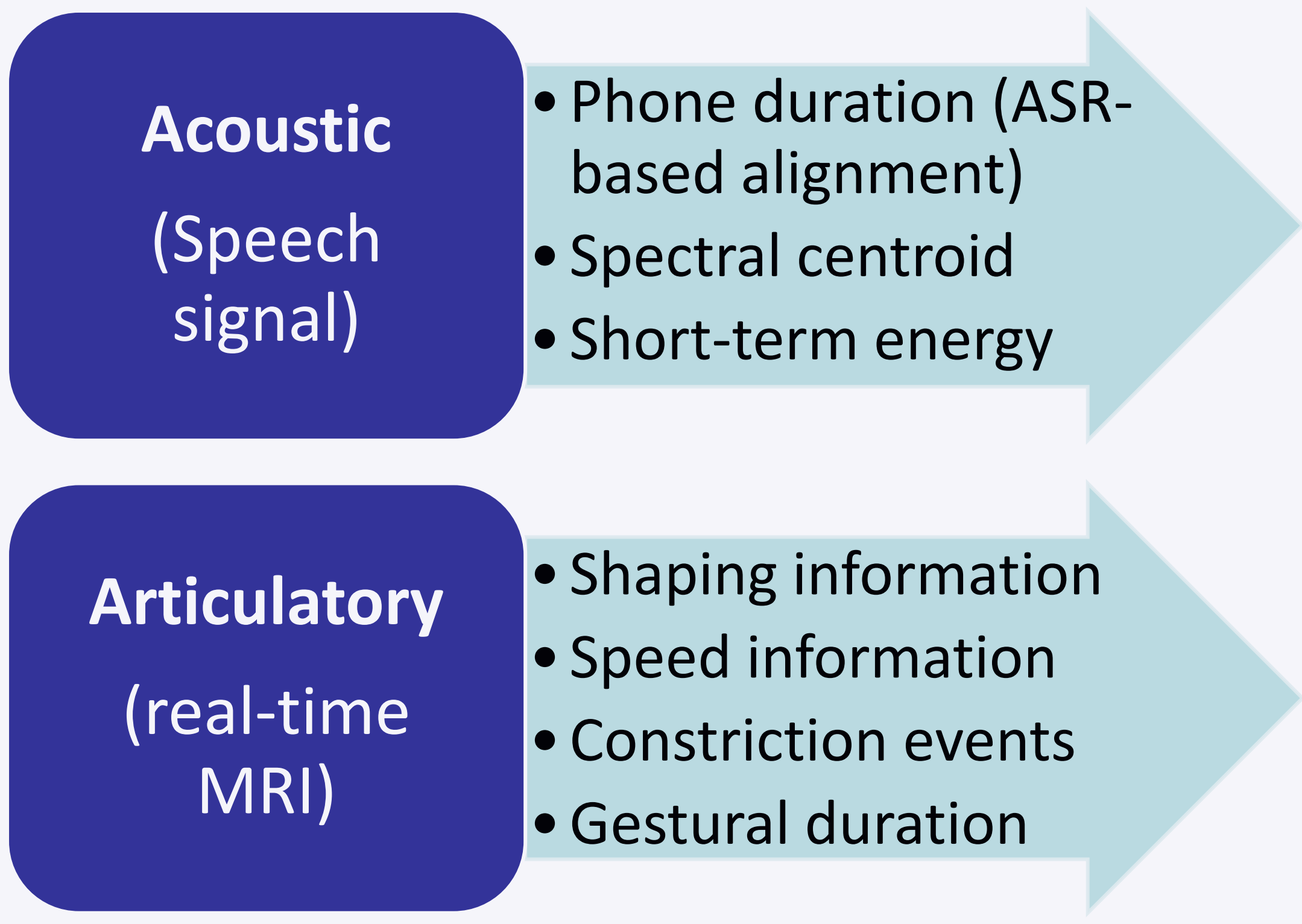
- **Articulatory Phonology (Browman & Goldstein, 1992)**
  - Speech act decomposable into atomic units of vocal tract action – "Gestures"
  - E.g.: a set of 'articulators' in the vocal tract (see the Task Dynamics model, Saltzman and Munhall, 1989)
- **Prosodic (π) Gestures (Byrd & Saltzman, 2003)**
  - phrase junctures – phonologically planned intervals of controlled local slowing of speech around a phrase edge
- **Planned pausing – slowing down of speech 'clock'**
- **Unplanned pauses interfere with articulators reaching their targets**



First pass – SONIC ASR - problems in spontaneous speech-MRI scan noise

Manual second pass to verify segmentation accuracy

MEASURES EXTRACTED



Acoustic (Speech signal)

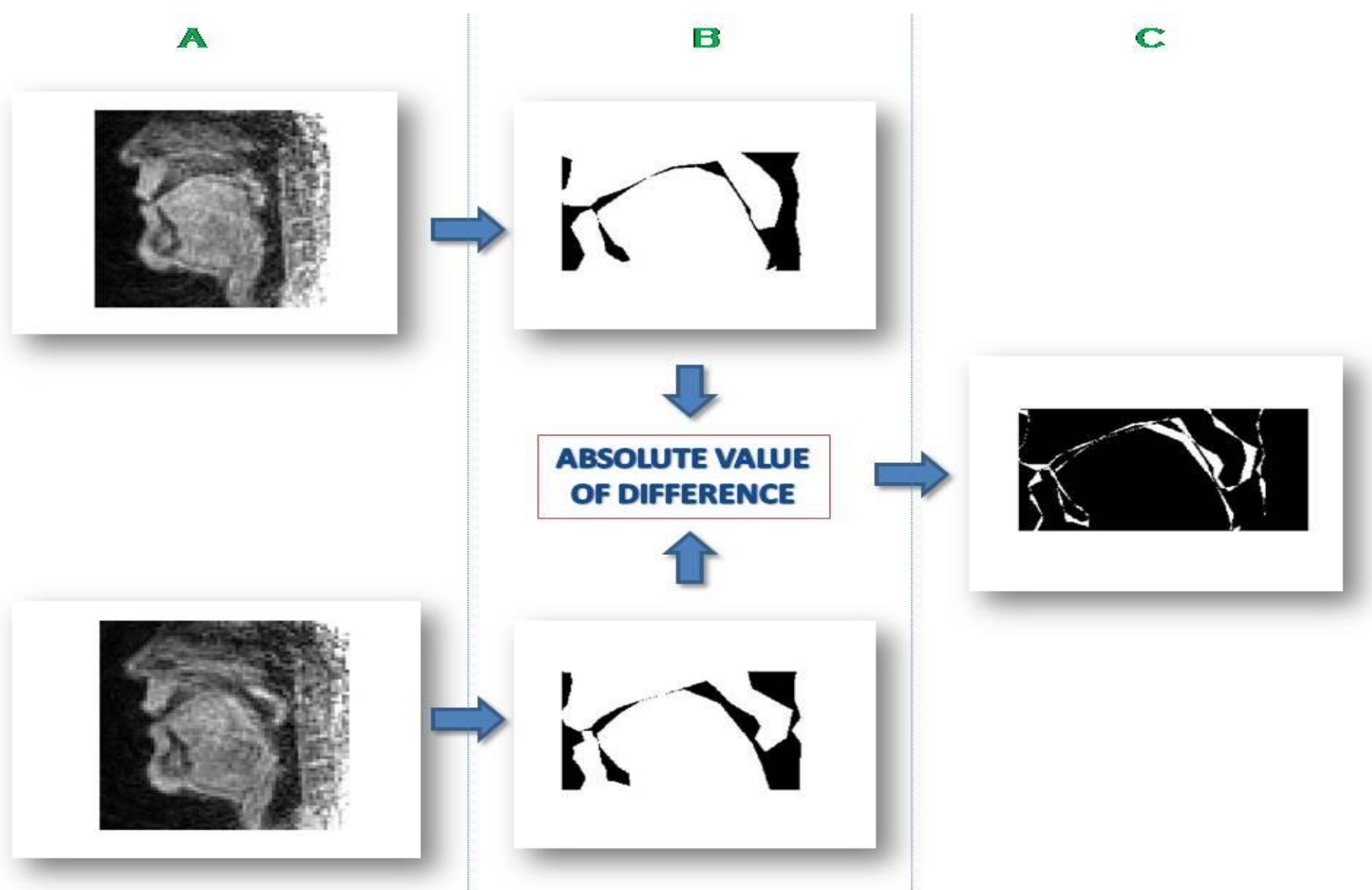
- Phone duration (ASR-based alignment)
- Spectral centroid
- Short-term energy

Articulatory (real-time MRI)

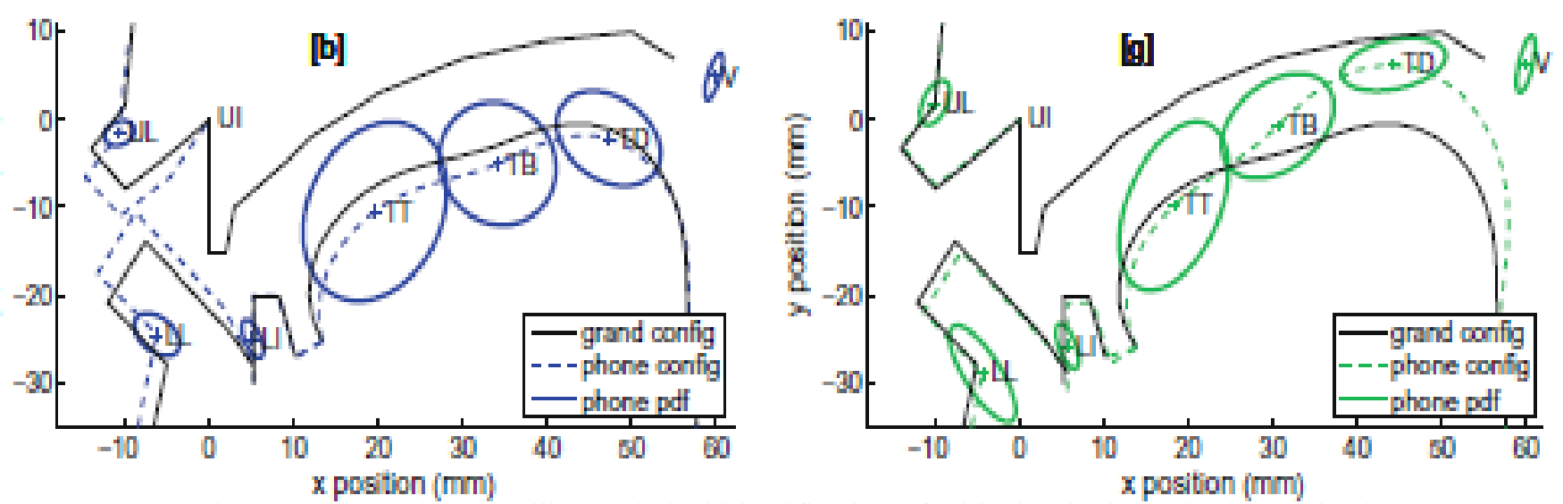
- Shaping information
- Speed information
- Constriction events
- Gestural duration

Articulatory information provides vital complementary information to acoustics!

SPEED MEASURE



SHAPING AND CONSTRICTION MEASURES

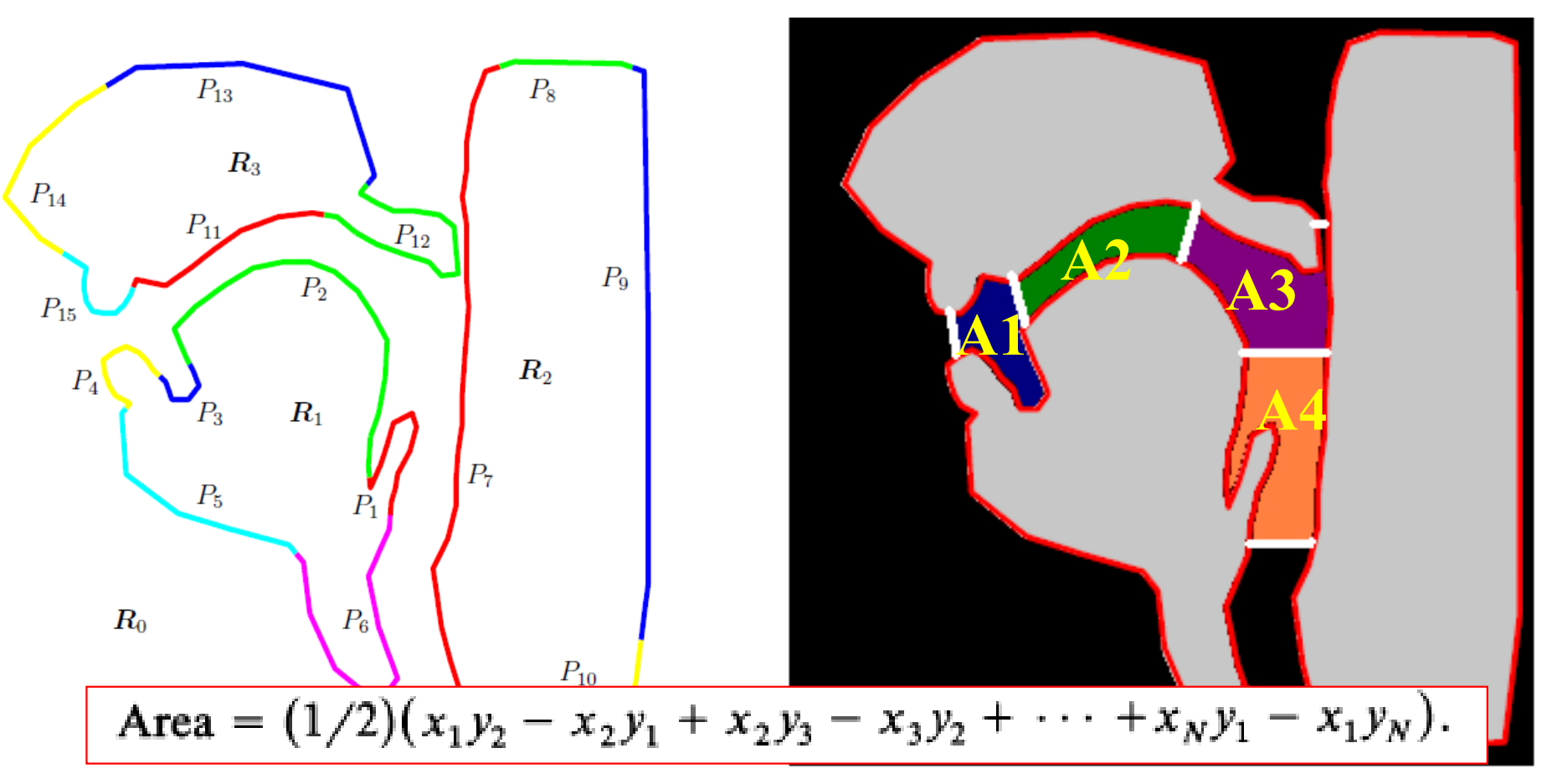


P.J.B. Jackson and V.D. Singampalli, "Statistical identification of critical articulators in the production of speech", *Speech Comm.*, 51(8): 695-710, August 2009.

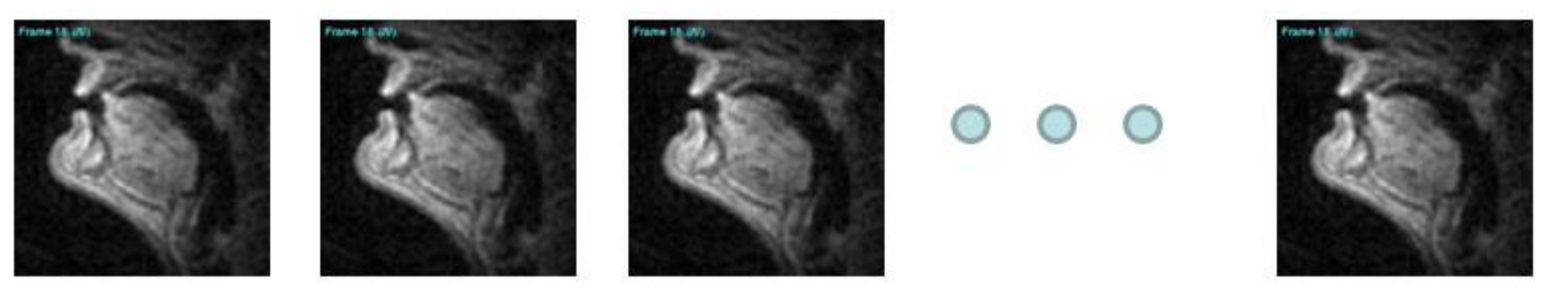
Critical articulator behavior – constrained

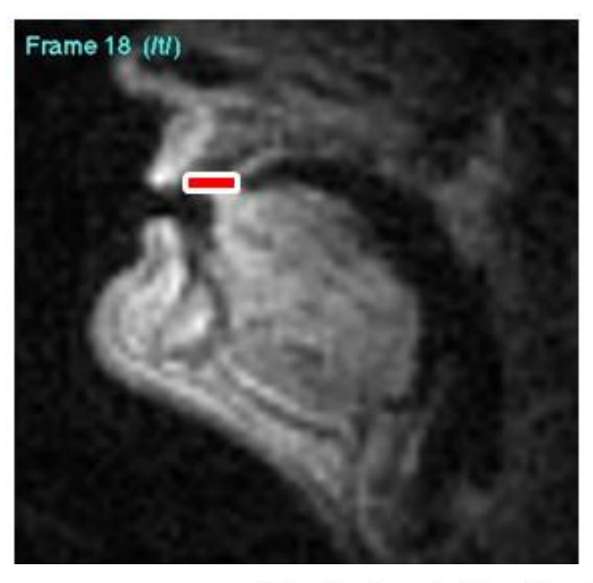
Dependent and *redundant* articulators – NOT constrained!

<u>Idea</u>: In addition to tract variables, incorporate information about vocal tract *areas* !



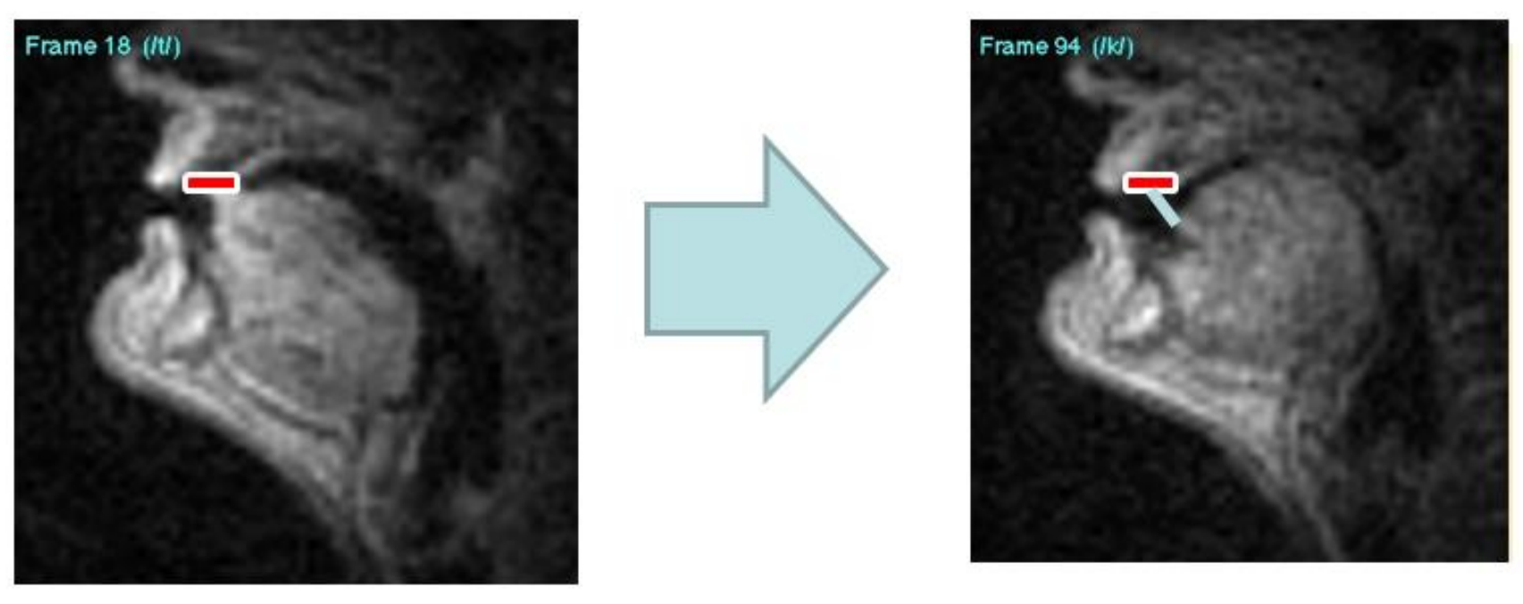
$$\text{Area} = (1/2)(x_1y_2 - x_2y_1 + x_2y_3 - x_3y_2 + \cdots + x_Ny_1 - x_1y_N).$$





STEP 1: Find palate-contact coordinate values for all sounds in the utterance that involve the Tongue Tip as a CRITICAL articulator (e.g., /t/, /d/) and use these points to form a "point-cloud distribution"

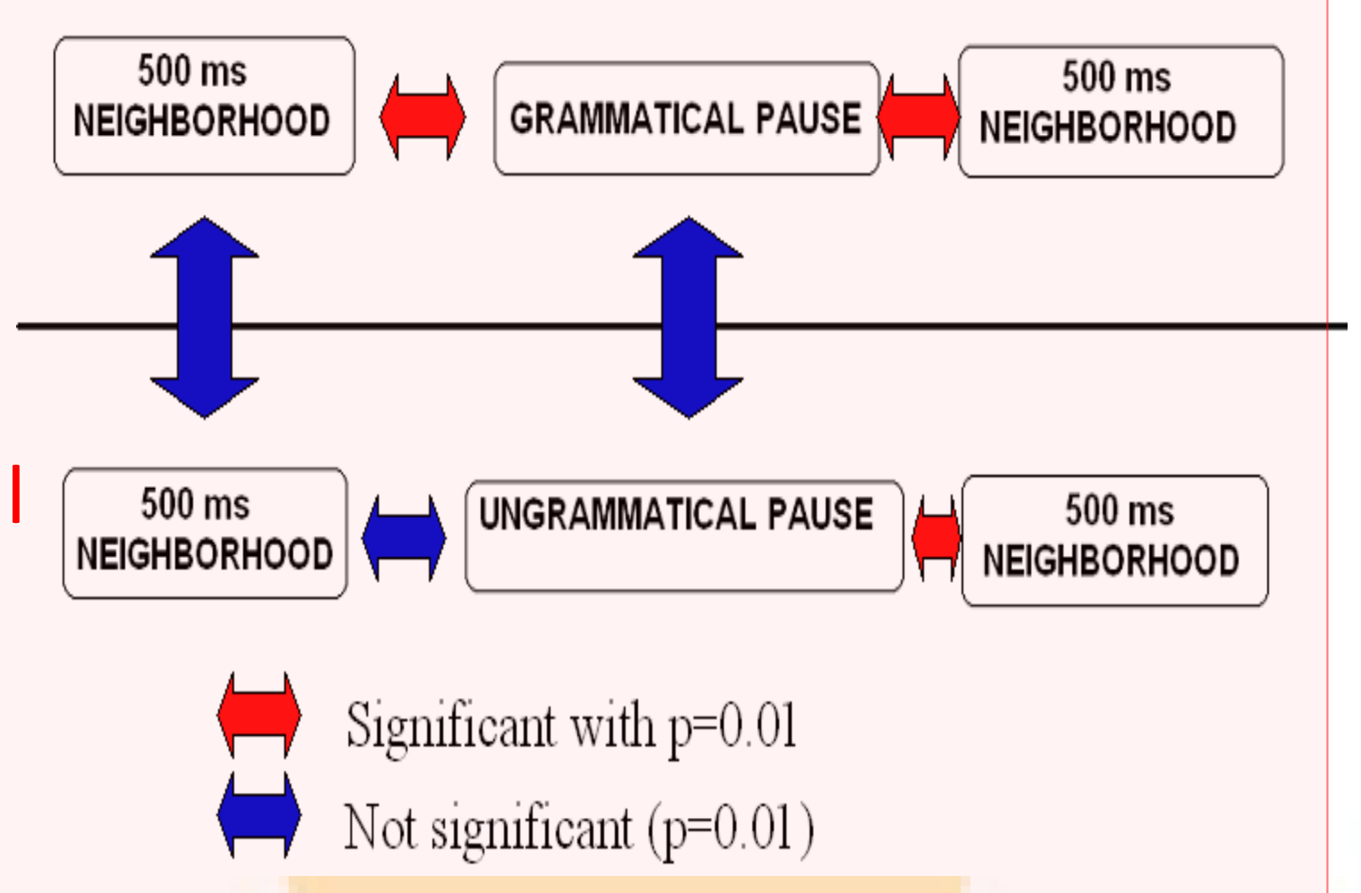
STEP 2: To find TTCD for a NON-CRITICAL frame, simply find the minimum distance from the mean of this TT point cloud to the tongue contour .



Analysis of pausing behavior in spontaneous speech

<u>**Data**</u>: 7 subjects – 20-30s spon speech responses to questions like "tell me more about your family"…

Grammatical and ungrammatical pauses classified manually and verified by a linguist.



<u>**Results:**</u>

- Grammatical case: the speed of the articulators drops significantly at the pause (p~0) and then increases again to the pre-pause level
- Ungrammatical case: only a slight drop into the pause, followed by a sudden post-pausal jump (back into 'grammaticality
- Variance for ungrammatical pauses (& neighborhoods) higher (20%)

Analysis of spectral reduction in VNV sequences, extracted from read and spontaneous speech

<u>Data:</u> Parallel MRI/audio corpus of TIMIT shibboleth sentences and spontaneous responses to questions (e.g., "tell me about your favorite cuisine", etc.) from one American English (female) speaker

Totally 53 read v/s 117 spontaneous VNV samples

SHAPE PROPERTY	MEASURE	HIGHER?
Area between lips and tongue tip	VTAD A1	SPONTANEOUS
Area between tongue and hard palate	VTAD A2	READ
Area of pharyngeal region	A3 + A4	READ

KINEMATIC PROPERTY	MEASURE	HIGHER
Maximum and average width of velum opening	VEL (Tract variable)	NEITHER
Rate of change of areas of the vocal tract	All ΔVTADs	NEITHER
Maximum and average speed of velum opening	Gradient Frame Energy	NEITHER
VARIABLITY PROPERTY	**MEASURE**	**HIGHER**
Variance of vocal tract area over course of VNV	All VTADs	SPONTANEOUS
Blurring in constriction location	Palate point cloud variance	SPONTANEOUS
Variance in articulator speeds	Gradient frame energy	SPONTANEOUS

FUTURE DIRECTIONS

- Develop further models of a central 'cognitive planner' and how recovery from perturbation of utterance structure happens
- Use these results to postulate models of speech planning and execution in different speaking styles & manners of articulation
- Apply these ideas/models to text-to-speech and dialog manager systems.

References

Browman, C. P., & Goldstein, L., (1992) "Articulatory phonology: an overview," *Phonetica*, 49, 155-180.

D. Byrd & E. Saltzman. (2003) "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," *Journal of Phonetics*, 31, 149-180

Saltzman, E. L., & Munhall, K. G. (1989). "A dynamical approach to gestural patterning in speech production," Ecological Psychology, 1, 333–382.