Measuring and Using Speech Production Information

...: Some New Opportunities

Shrikanth (Shri) Narayanan
SAIL: Signal Analysis and Interpretation Laboratory
http://sail.usc.edu

Prof. Ken Stevens, 1924-2013
To whom we owe a lot...

SPASR, Aug 2013

USC
School of Engineering

University of Southern California

Speech Production and Articulation kNowledge Group

http://sail.usc.edu/span

Diverse Corpora

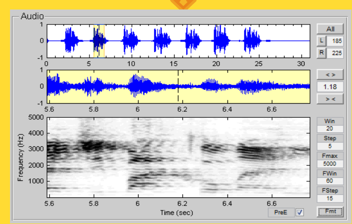
- Multilingual Ling. Material
- MOCHA-TIMIT
- Audio Books
- North Wind
- Spontaneous Speech

Multimodal Phonetic Data Acquisition

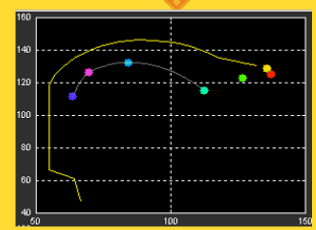
RT-MRI

3d MRI

Audio

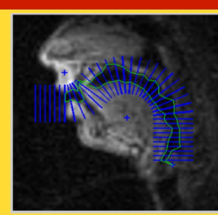


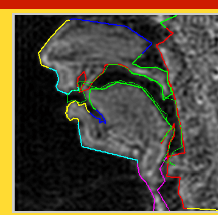
EMA

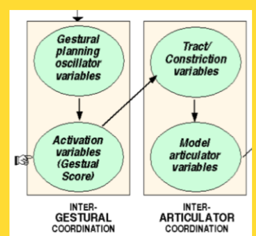


Multimodal Analysis & Modeling

- direct image analysis
- forced alignment
- articulator tracking
- acoustic feature extraction
- cross-modal registration
- airway segmentation
- morphological characterization
- task-dynamic modeling
- dynamic 3d vocaltract modeling
- HM Modelling of articulatory states







New Insights Into

- dynamics of production
- 3d vocal tract shaping
- articulatory coordination
- source-filter interaction
- encoding of emotive factors
- realization of prosody
- speaker-specific phonetics

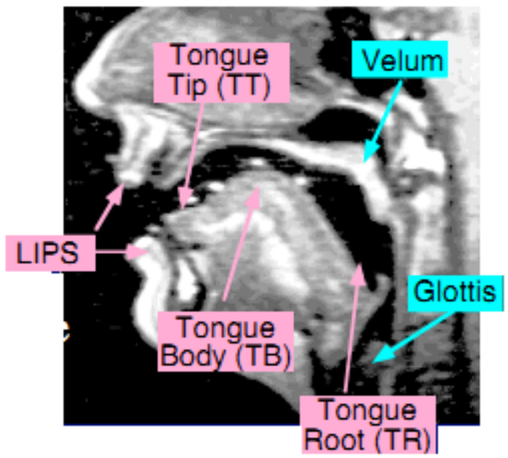
TECHNOLOGY APPLICATIONS

ARTICULATORY PHONOLOGY-BASED "GESTURES"

ARTICULATORY PHONOLOGY-BASED "GESTURES"

Gestural hypothesis:
Act of speaking can be decomposed into atomic units of action, or **gestures**.

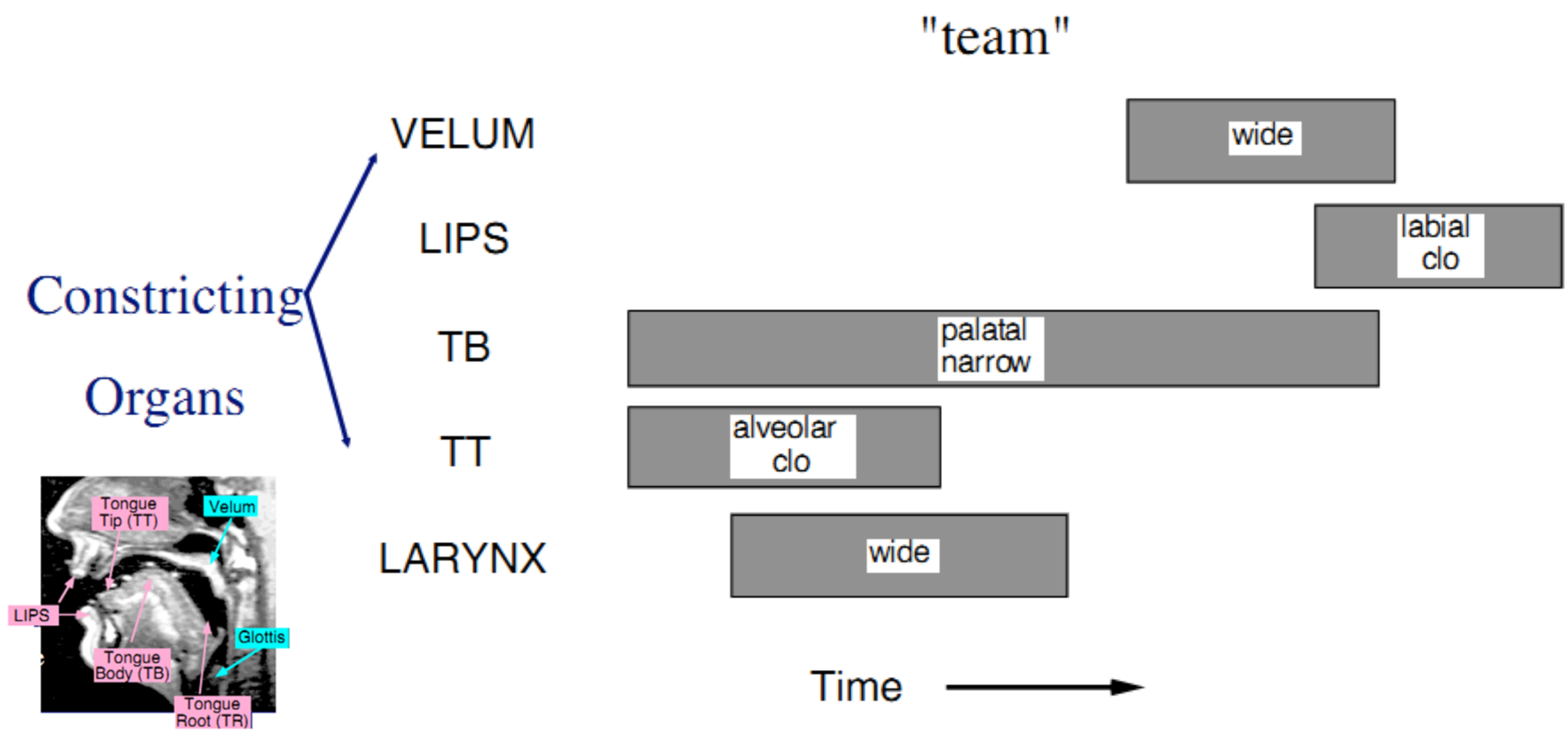
Gestures are dynamically-controlled constriction actions of distinct vocal tract organs. (e.g., lips, tongue tip, tongue body, velum, glottis)



C. Browman and L. Goldstein, "Dynamics and articulatory phonology," Mind as motion: Explorations in the dynamics of cognition, 1995.

ARTICULATORY PHONOLOGY-BASED "GESTURES"

ARTICULATORY PHONOLOGY-BASED "GESTURES"



Gestural scores (Browman and Goldstein, 1992, 1995) represent latent activation intervals for dynamical systems controlling constrictions.

Theoretical themes

- ***compositionality in time***:
  - diphthong production
  - nasal coordination
  - prosody of read/spontaneous speech
  - geminate vs. singleton consonants
- ***compositionality in space***:
  - 'complex consonant' production: liquids, coronals, fricatives
  - characterization of retroflexion
  - structure and realization of consonant clusters
- ***compositionality in cognition***:
  - speech errors
  - human beatboxing
  - velic coordination

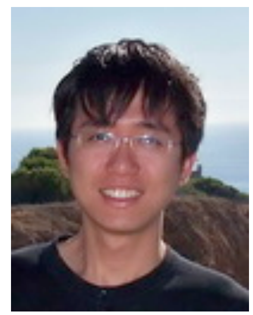
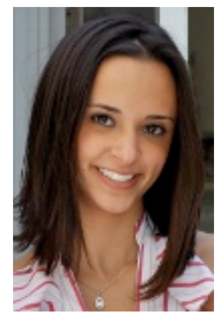
5

USC SPAN CoreTeam







Alums





Talk Premise & Layout

- **Understanding the system that produces speech is essential to improving the performance of speech technology systems**
  - Scientific studies: Empirical analyses, Direct system (forward) modeling
  - Technology studies: Feature engineering, Inverse modeling, Applications to ASR, Speaker Modeling, Synthesis, Clinical problems

✓ Measuring speech production
  - Multimodal approaches: EMA, Ultrasound, MRI,..

✓ Extracting features (representations)
  - Direct & Estimated (inversion)

✓ Modeling speech production
  - Theoretically inspired & Data-driven

✓ Applications
  - ASR, Speaker modeling

Speech Production Studies: Data Is Integral

- Observe, measure, visualize articulatory details during speech
- Long history of instrumentation and imaging applications
- Number of techniques, each with its own strengths and limitations
  - Spatial and temporal resolution
  - Subject safety
  - Flexibility, ease of use, portability
  - Data interpretability
  - Specific research and application needs

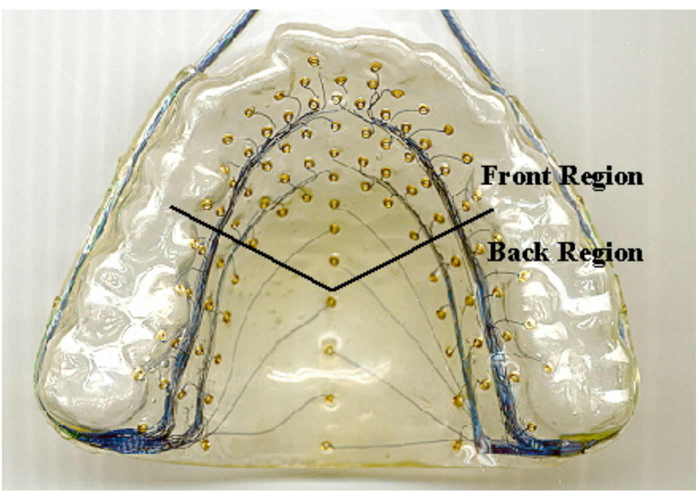
Classic Speech Production Data

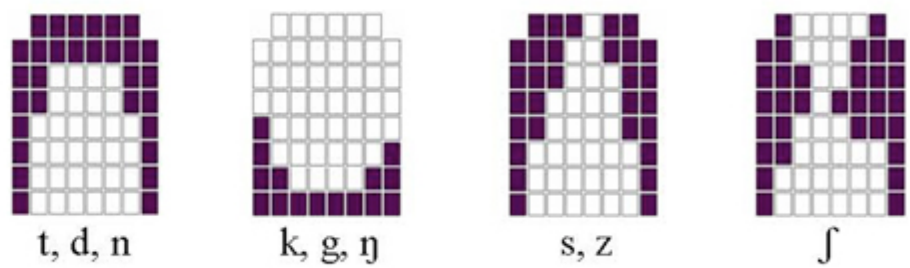
X-ray (Stevens, 1962)

http://psyc.queensu.ca/~munhallk/05_database.htm

Ultrasound (Stone, 1980)

http://www.speech.umaryland.edu

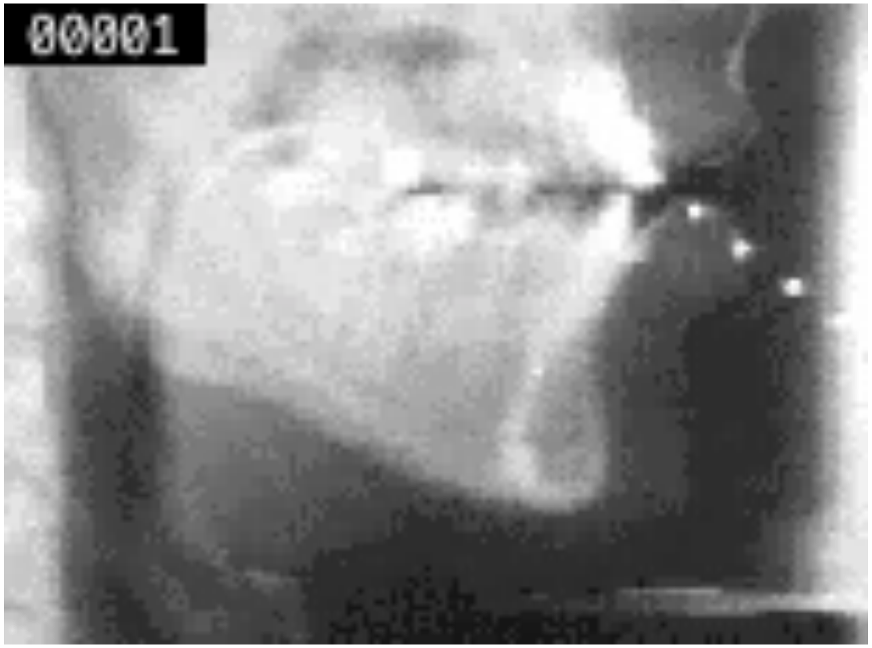




Electropalatography

(courtesy: UCLA Phonetics Lab)

Classic Speech Production Data

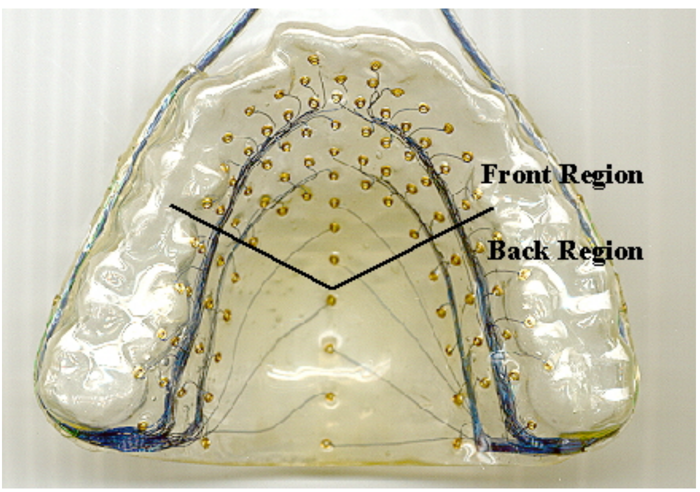


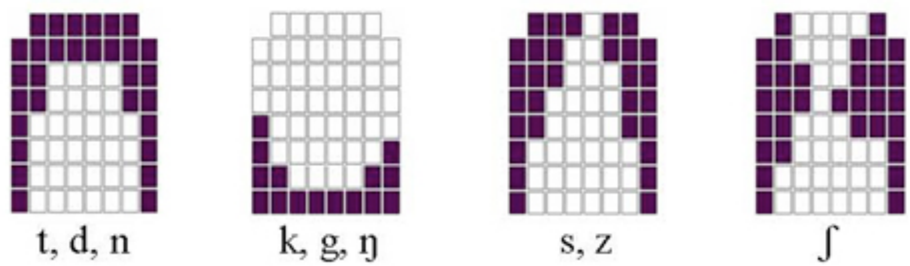
X-ray (Stevens, 1962)

http://psyc.queensu.ca/~munhallk/05_database.htm

Ultrasound (Stone, 1980)

http://www.speech.umaryland.edu

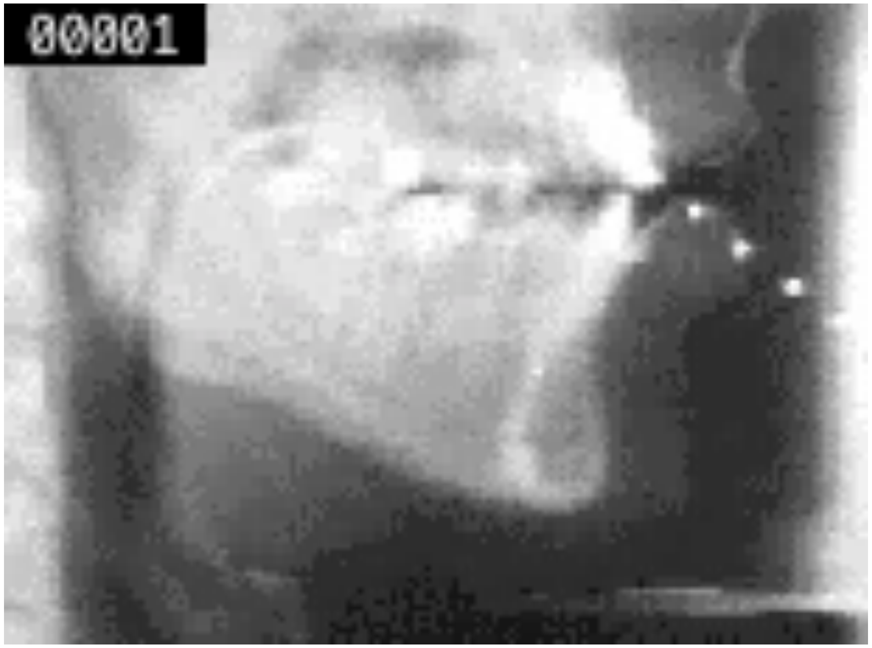




Electropalatography

(courtesy: UCLA Phonetics Lab)

Classic Speech Production Data

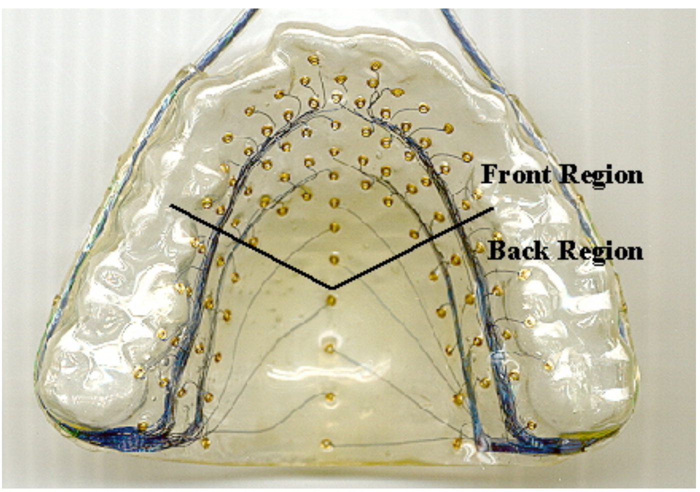


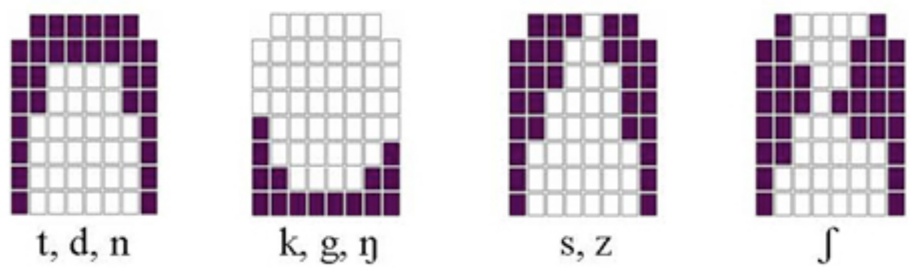
X-ray (Stevens, 1962)

http://psyc.queensu.ca/~munhallk/05_database.htm

Ultrasound (Stone, 1980)

http://www.speech.umaryland.edu

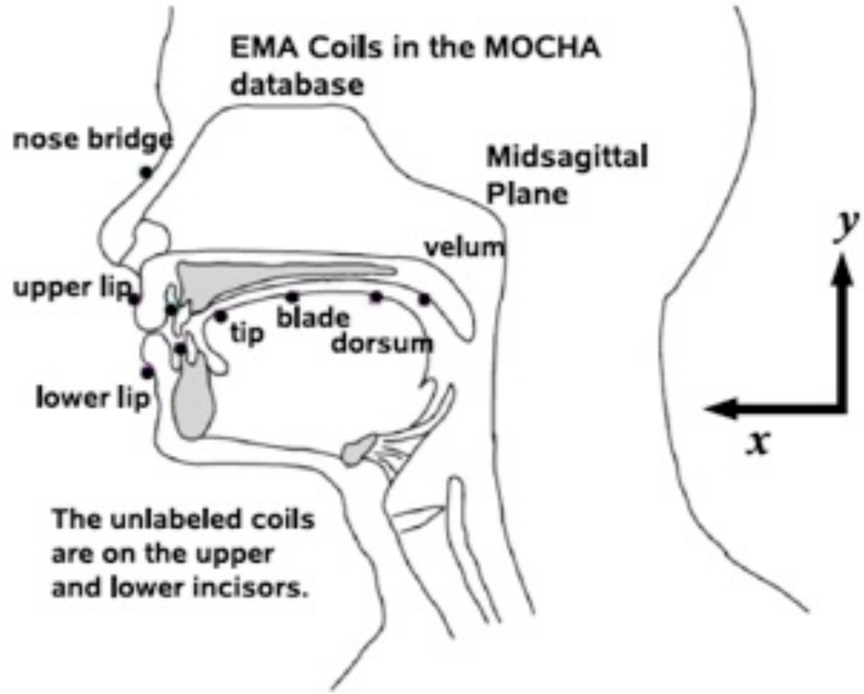




Electropalatography

(courtesy: UCLA Phonetics Lab)

ELECTROMAGNETIC ARTICULOGRAPHY (EMA)

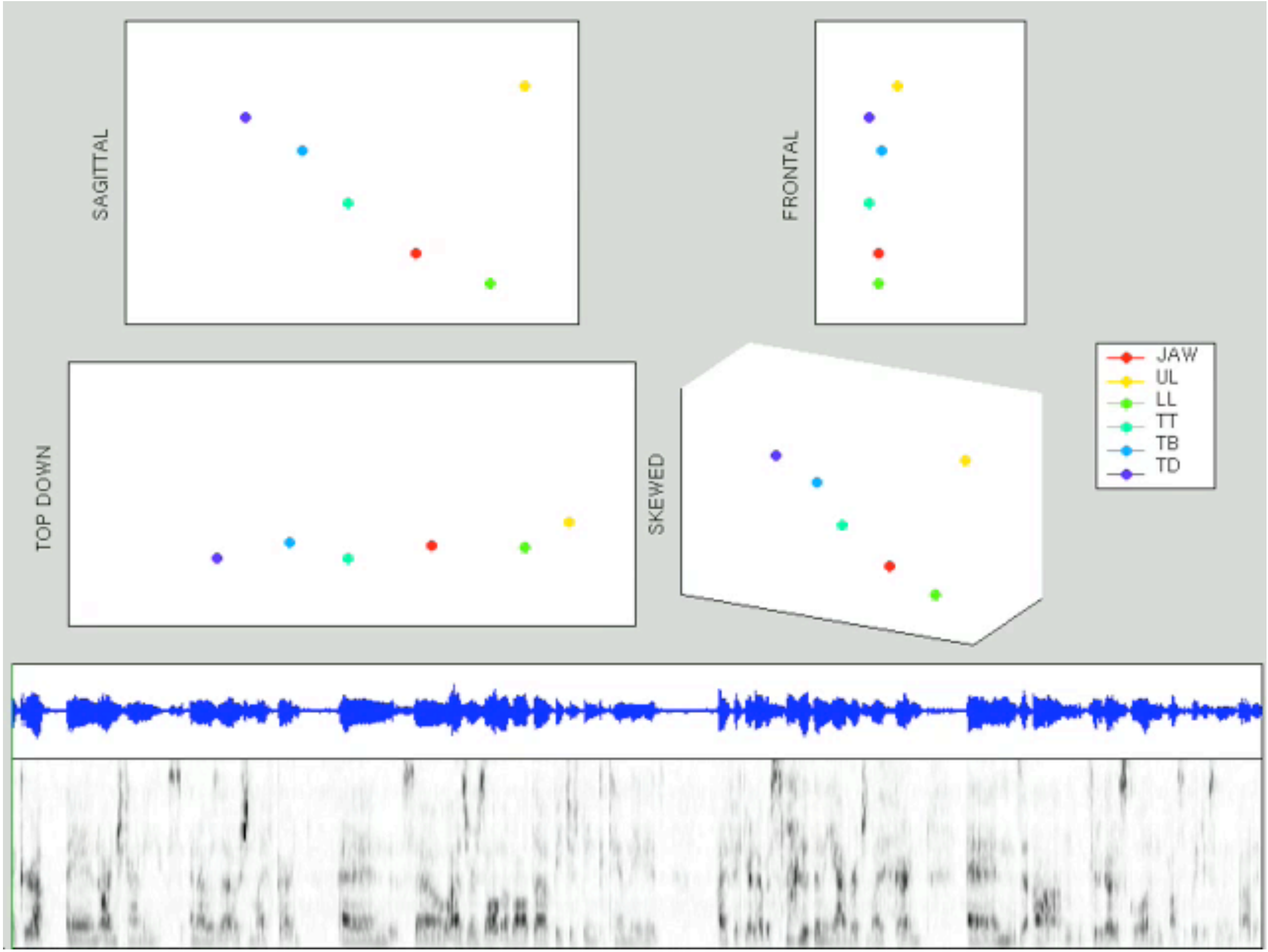


Wrench (2000)

A. Wrench, A multichannel articulatory database and its application for automatic speech recognition
Proceedings 5th Seminar of Speech Production, 2000

ELECTROMAGNETIC ARTICULOGRAPHY (EMA CORPUS, USC 2007)

ELECTROMAGNETIC ARTICULOGRAPHY (EMA CORPUS, USC 2007)



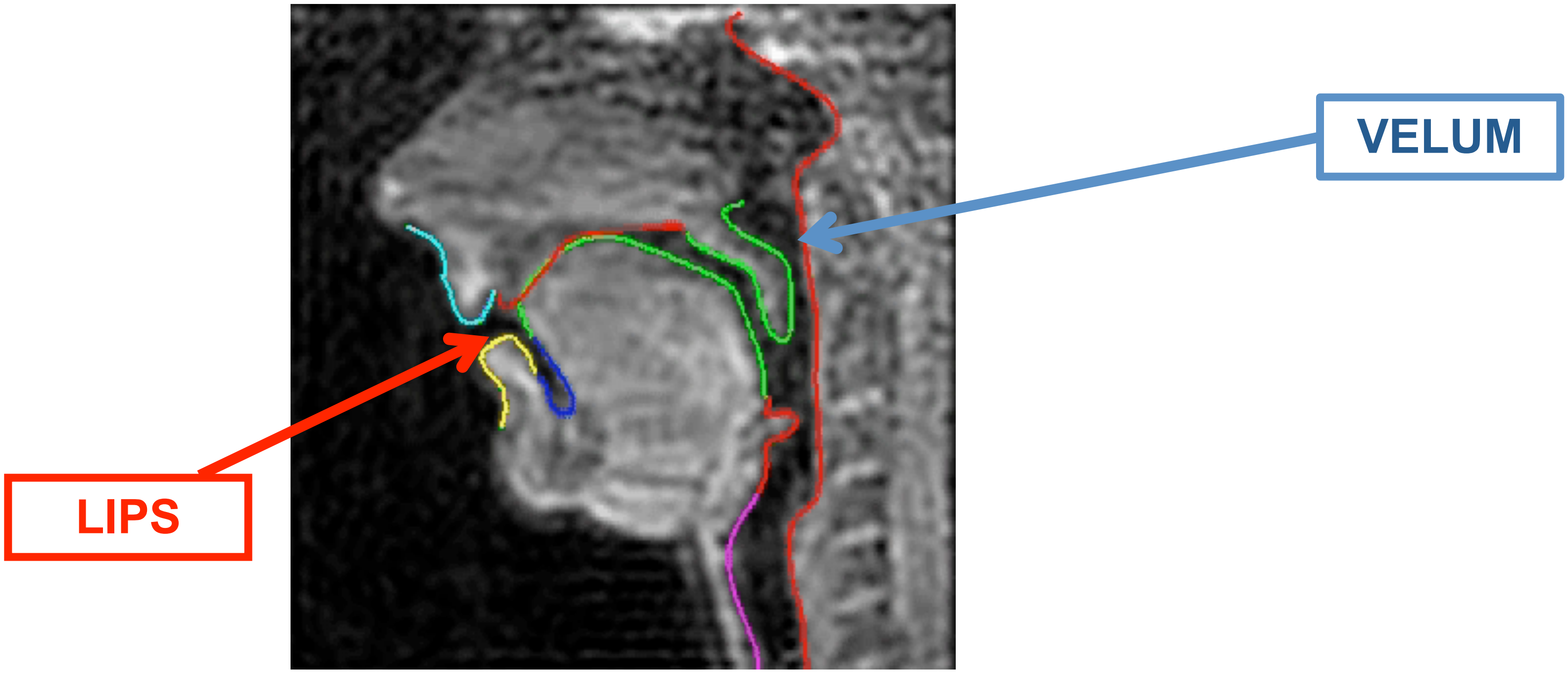
11

REAL-TIME MRI (rt-MRI)

Offers full midsaggital view of all supraglottal vocal tract articulators
(cf. to x-ray microbeam, EMA, ultrasound.)

S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," JASA, vol. 115, p. 1771, 2004.

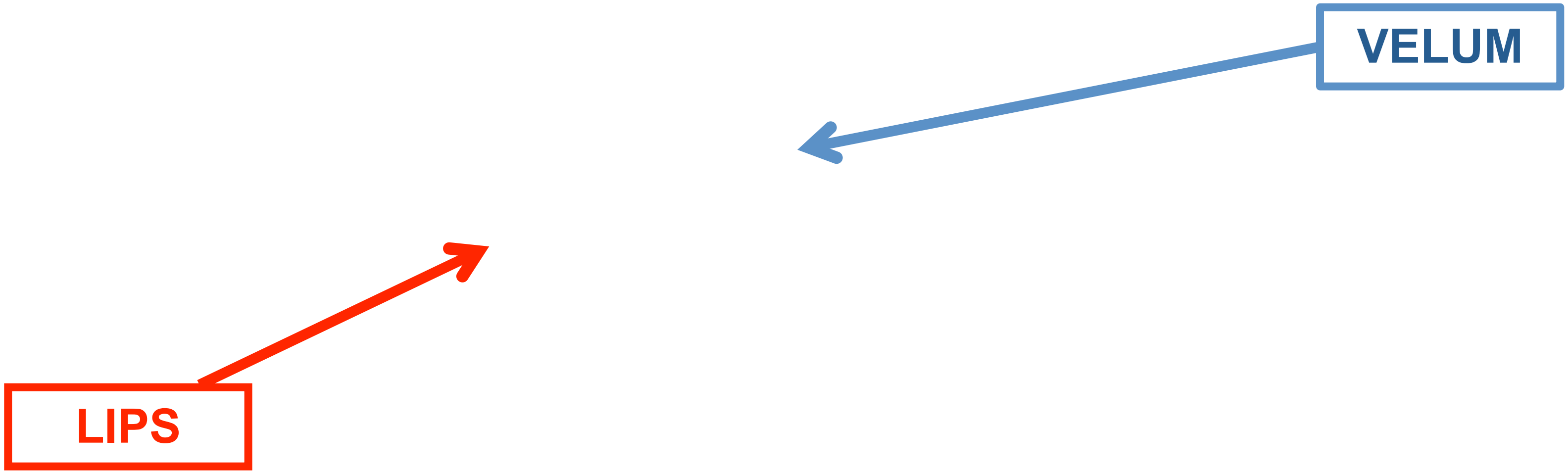
REAL-TIME MRI (rt-MRI)



Offers full midsaggital view of all supraglottal vocal tract articulators
(cf. to x-ray microbeam, EMA, ultrasound.)

S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," JASA, vol. 115, p. 1771, 2004.

REAL-TIME MRI (rt-MRI)

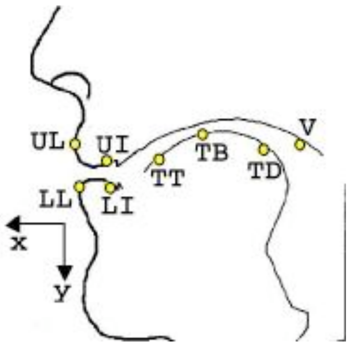


Offers full midsaggital view of all supraglottal vocal tract articulators
(cf. to x-ray microbeam, EMA, ultrasound.)

S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," JASA, vol. 115, p. 1771, 2004.

HOW DO DIFFERENT TECHNIQUES COMPARE?

HOW DO DIFFERENT TECHNIQUES COMPARE?



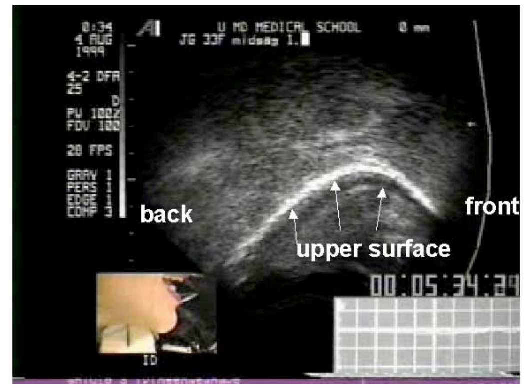
EMA (Wrench 2000)

X-Ray Microbeam, XRMB

(Westbury 1994)

(rt) Magnetic Resonance Imaging, MRI

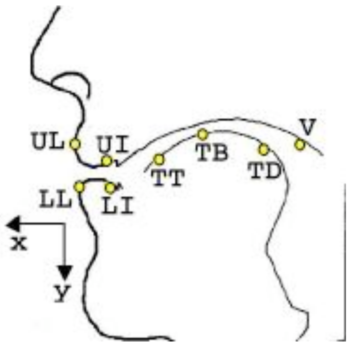
(Narayanan 2004)



Ultrasound

(Stone 1980; Whalen 2005)

HOW DO DIFFERENT TECHNIQUES COMPARE?



EMA (Wrench 2000)

X-Ray Microbeam, XRMB

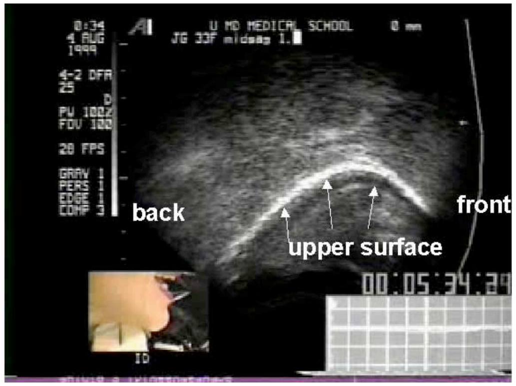
(Westbury 1994)

Fleshpoints

(rt) Magnetic Resonance Imaging, MRI

(Narayanan 2004)

Full mid-sagittal (or any section) view; 3D

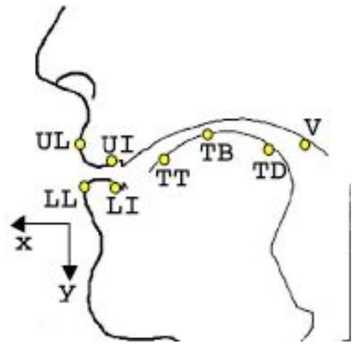


Ultrasound

(Stone 1980; Whalen 2005)

Tongue (partial, surface view)

HOW DO DIFFERENT TECHNIQUES COMPARE?



EMA (Wrench 2000)

X-Ray Microbeam, XRMB

(Westbury 1994)

Fleshpoints

Invasive

Cumbersome

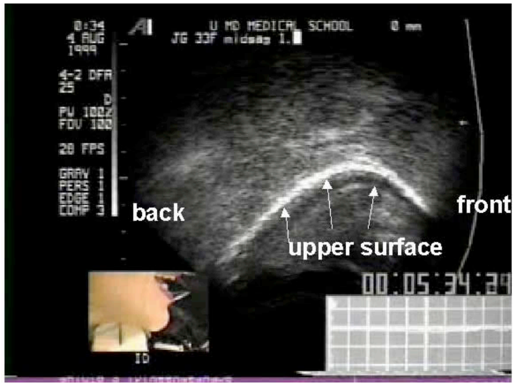
(rt) Magnetic Resonance Imaging, MRI

(Narayanan 2004)

Full mid-sagittal (or any section) view; 3D

Non-invasive

Cumbersome



Ultrasound

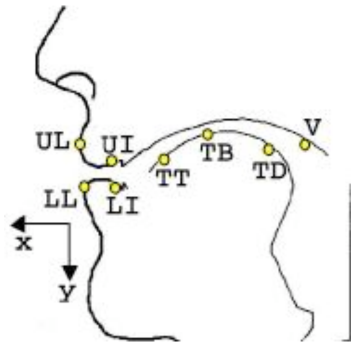
(Stone 1980; Whalen 2005)

Tongue (partial, surface view)

Minimally invasive

Portable, Easy

HOW DO DIFFERENT TECHNIQUES COMPARE?



EMA (Wrench 2000)

X-Ray Microbeam, XRMB

(Westbury 1994)

Fleshpoints

Invasive

Cumbersome

~100-500 Hz

(rt) Magnetic Resonance Imaging, MRI

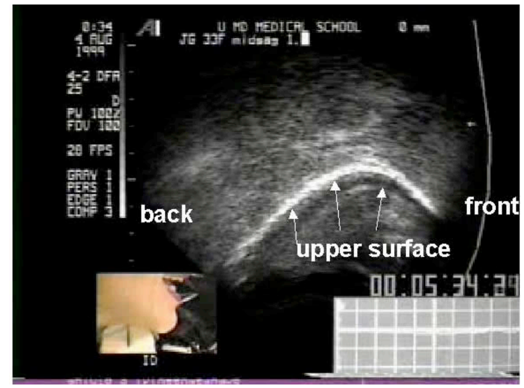
(Narayanan 2004)

Full mid-sagittal (or any section) view; 3D

Non-invasive

Cumbersome

~20-30 Hz



Ultrasound

(Stone 1980; Whalen 2005)

Tongue (partial, surface view)

Minimally invasive

Portable, Easy

~50-300 Hz

SOME DATA SUITABLE FOR "TECHNOLOGY" STUDIES

XRMB (Univ. of Wisconsin) ***[1]***

www.uni-jena.de/~x1siad/uwxrmbdb.html

- 32 F, 25 M; 118 different tasks incl. read sentences, paragraphs

MOCHA-TIMIT (Univ. of Edinburgh)*[2]*

http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

- One male and one female subject, each reading 460 TIMIT utterances
- Pre-processing of seven articulatory trajectories (500Hz)

EMA Database@MURI (Univ. of Southern California)*[3]*

http://sail.usc.edu/data.php

- One male American; Spontaneous conversations of 14 sessions (each ~5min)
- Pre-processing of six articulatory trajectories (200Hz)

[1] J. Westbury. X-RAY MICROBEAM SPEECH PRODUCTION DATABASE USER'S HANDBOOK, 1994.

[2] A.A. Wrench. A new resource for production modelling in speech technology. In Proc. Inst. of Acoust. (WISP), Stratford-upon-Avon, UK, volume 23 (3), pages 207-217, 2001.

[3] Jorge Silva, Vivek Rangarajan, Viktor Rozgic and Shrikanth S. Narayanan, Information theoretic analysis of direct articulatory measurements for phonetic discrimination, in: Proceedings ICASSP, pages 457-460, 2007

The mngu0 database

http://www.mngu0.org

- **EMA, MRI, Dental Casts Audio (from Edinburgh, LMU, Saarland)**
  - EMA: Articulators: Upper and lower lips, jaw, and three tongue points; 1,300 utterances
  - MRI: 3D volume 13 vowels, 16 consonants & Midsagittal “dynamic” scans CVCs, (C=16,V=3)

USC-TIMIT: A MULTIMODAL ARTICULATORY DATA CORPUS FOR SPEECH RESEARCH

USC UNIVERSITY OF SOUTHERN CALIFORNIA

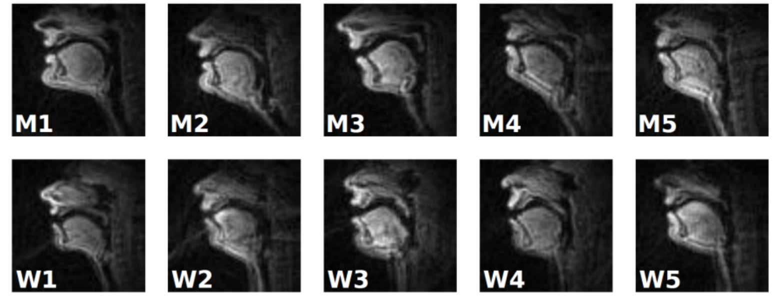
SAIL

NATIONAL INSTITUTES OF HEALTH

LAC+USC HEALTHCARE NETWORK

NSF

- 10 American English talkers (5M, 5F).
- Real time MRI (5 speakers also with EMA) and synchronized audio.
- 460 sentences each (>20 minutes)
- Freely available for speech research.

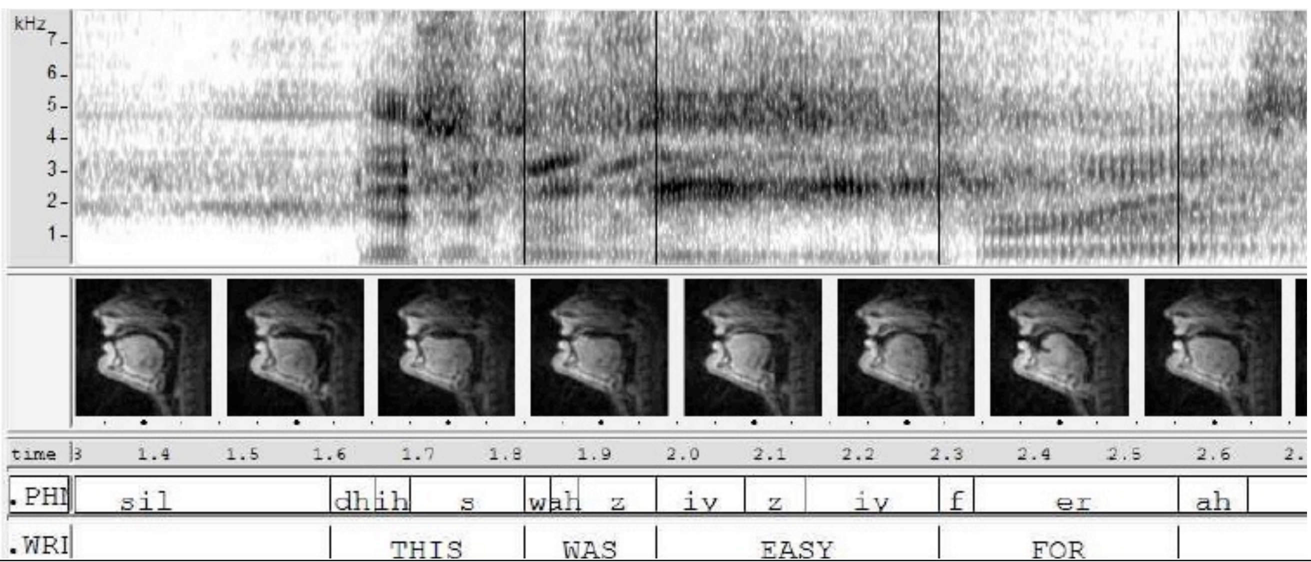


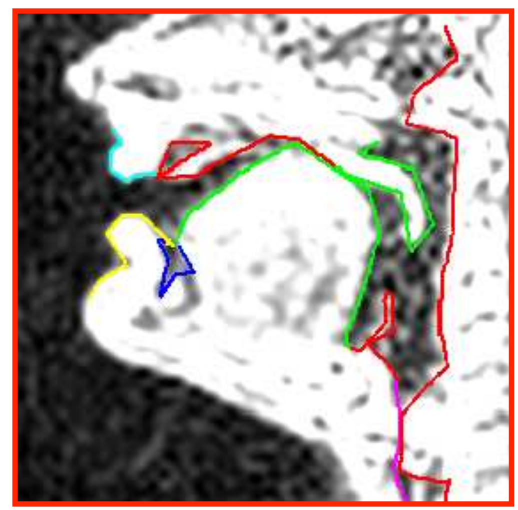
WEB-LINK (with download info):

http://sail.usc.edu/span/usc-timit/

SAIL homepage: http://sail.usc.edu

Narayanan et al. (2011). A Multimodal Real-Time MRI Articulatory Corpus for Speech Research. InterSpeech.



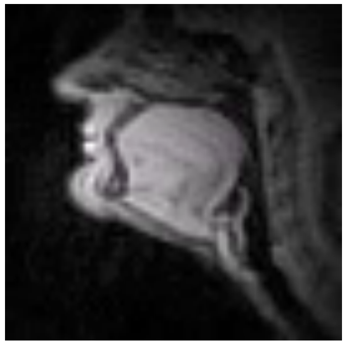


Some USC-TIMIT examples

M1 M2 F1 F2

Some USC-TIMIT examples

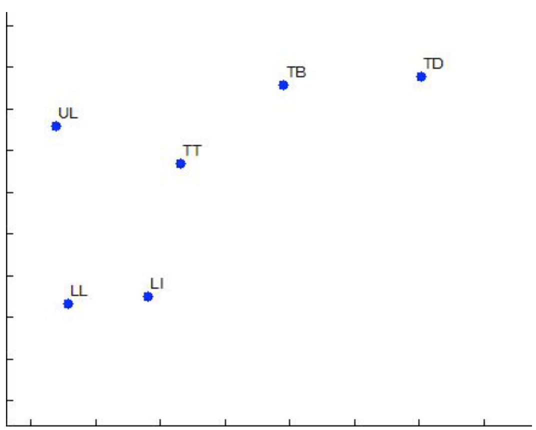
M1

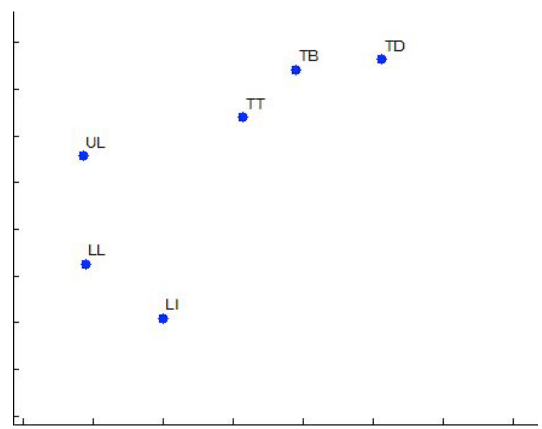


M2

F1

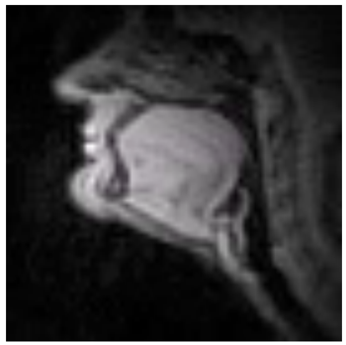
F2





Some USC-TIMIT examples

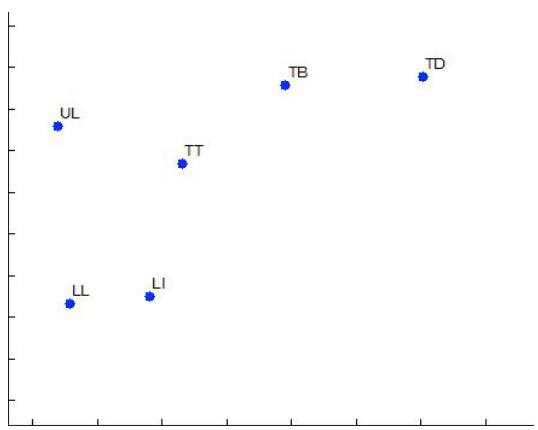
M1

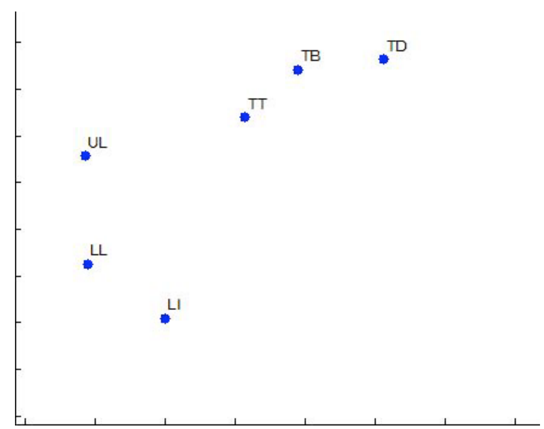


M2

F1

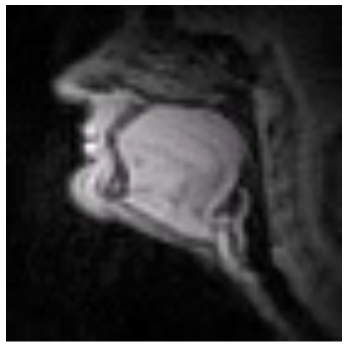
F2





Some USC-TIMIT examples

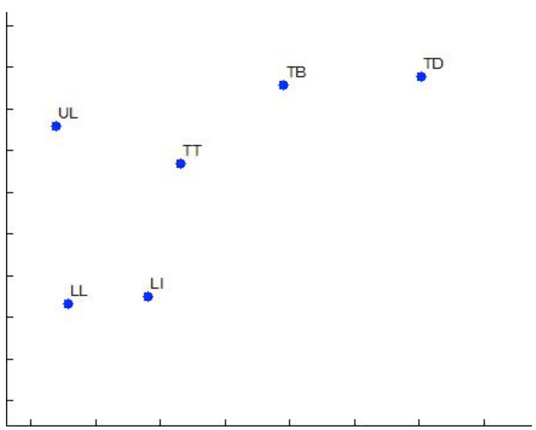
M1

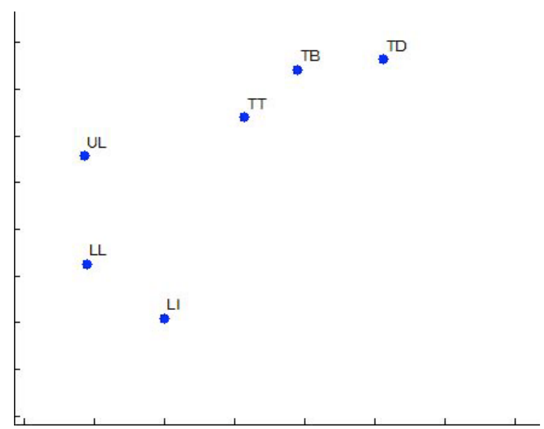


M2

F1

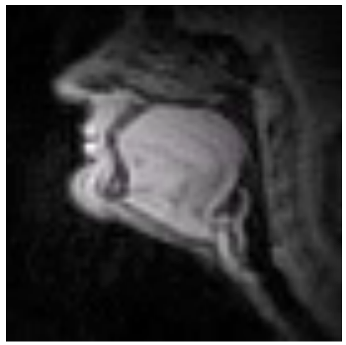
F2





Some USC-TIMIT examples

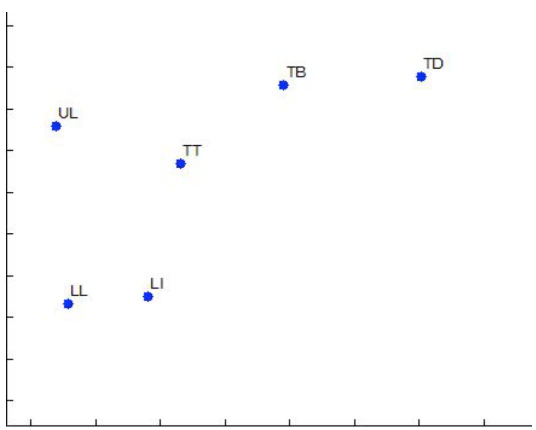
M1

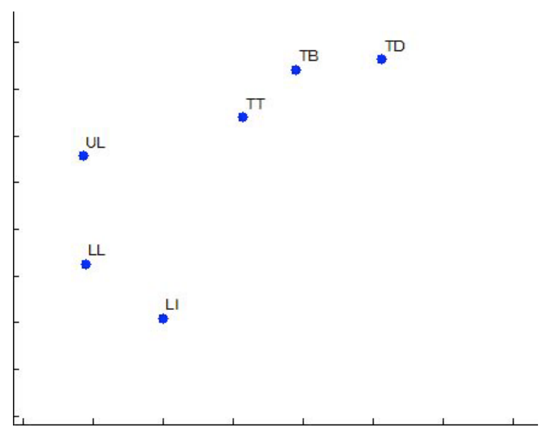


M2

F1

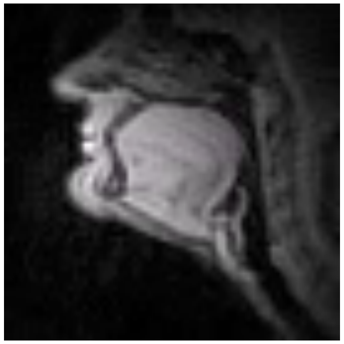
F2





Some USC-TIMIT examples

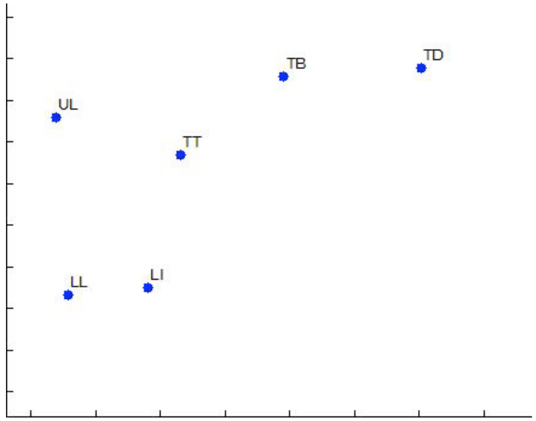
M1

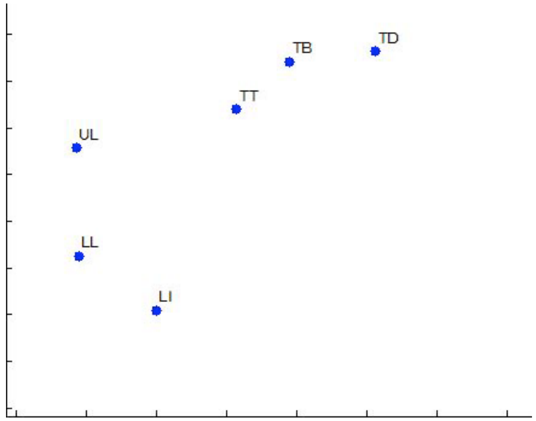


M2

F1

F2

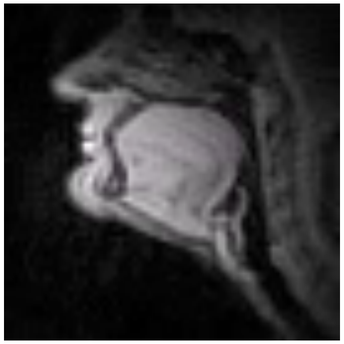




17

Some USC-TIMIT examples

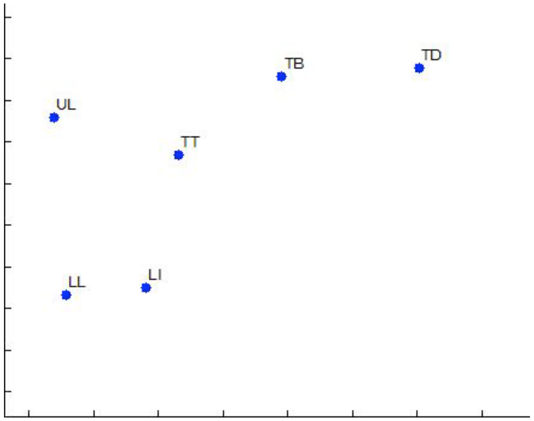
M1

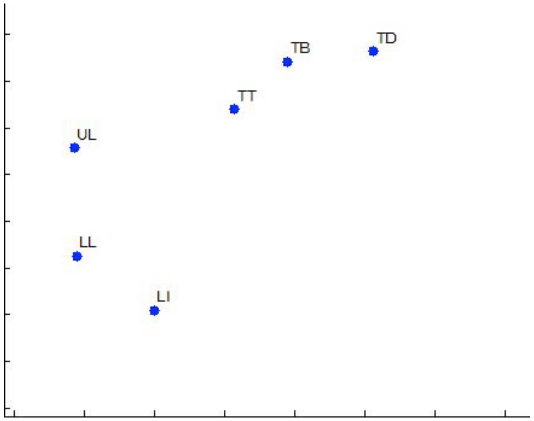


M2

F1

F2





17

Dynamic 3D visualization

Coronal movie

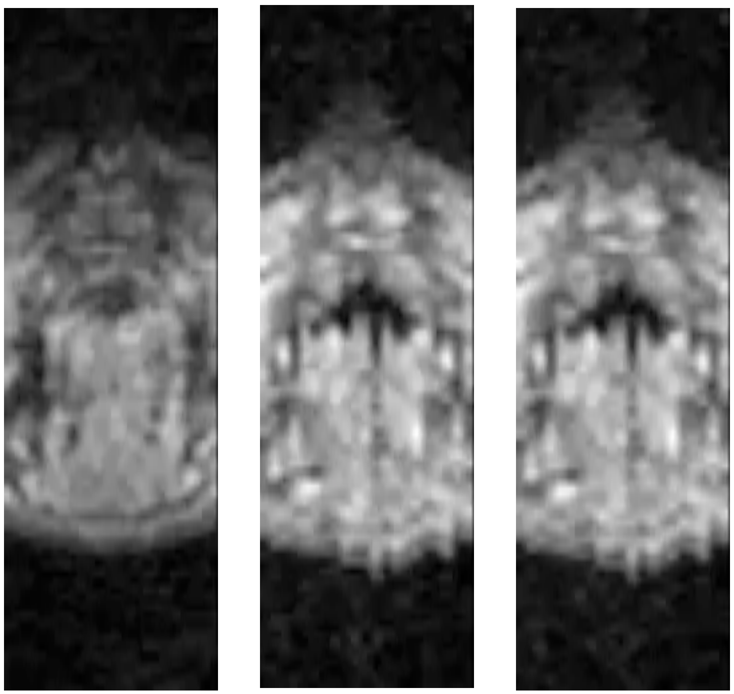
Aligned

Unaligned

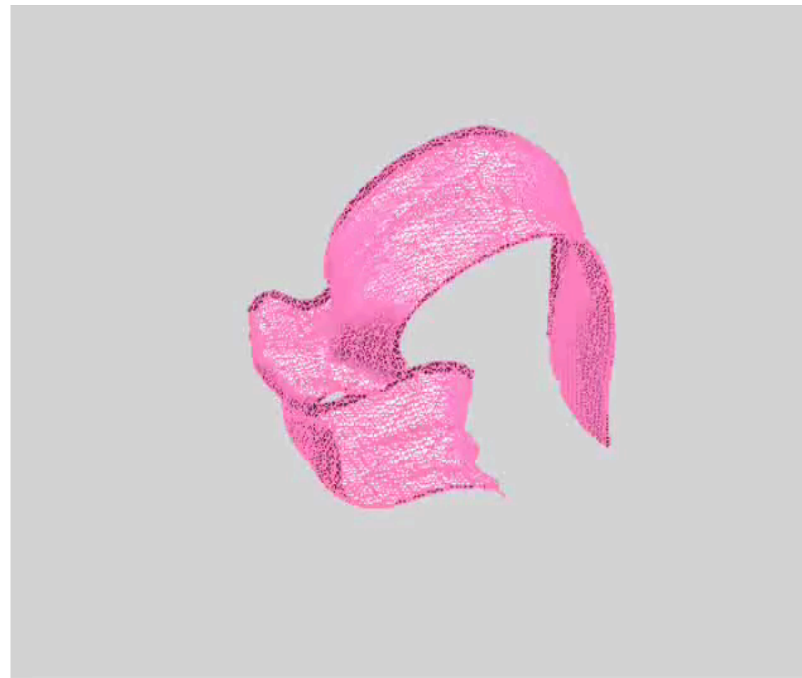
Yinghua Zhu, Yoon-Chul Kim, Michael Proctor, Shrikanth Narayanan, Krishna S. Nayak. Dynamic 3D Visualization of Vocal Tract Shaping during Speech. IEEE Transactions on Medical Imaging. 32(5): 838 - 848, May 2013.

18

Dynamic 3D visualization

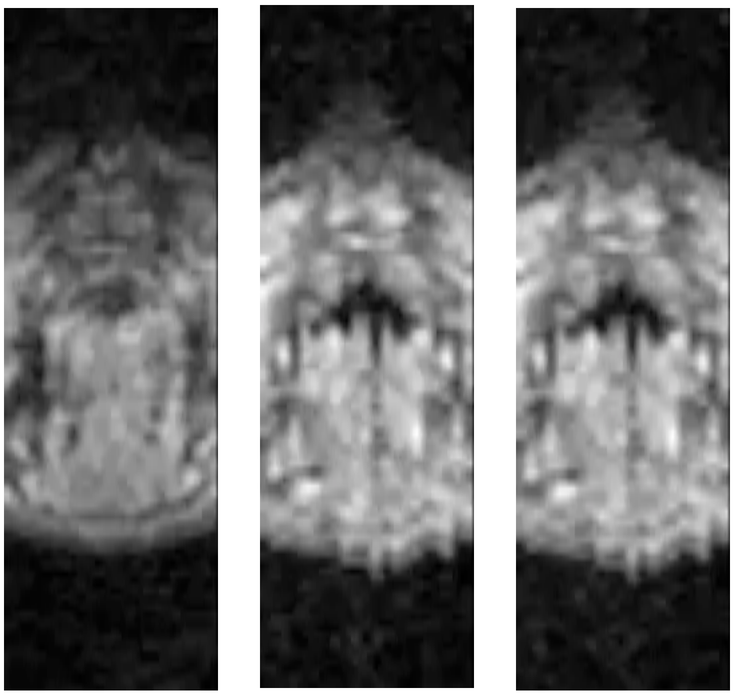


Coronal movie　　Aligned　　Unaligned

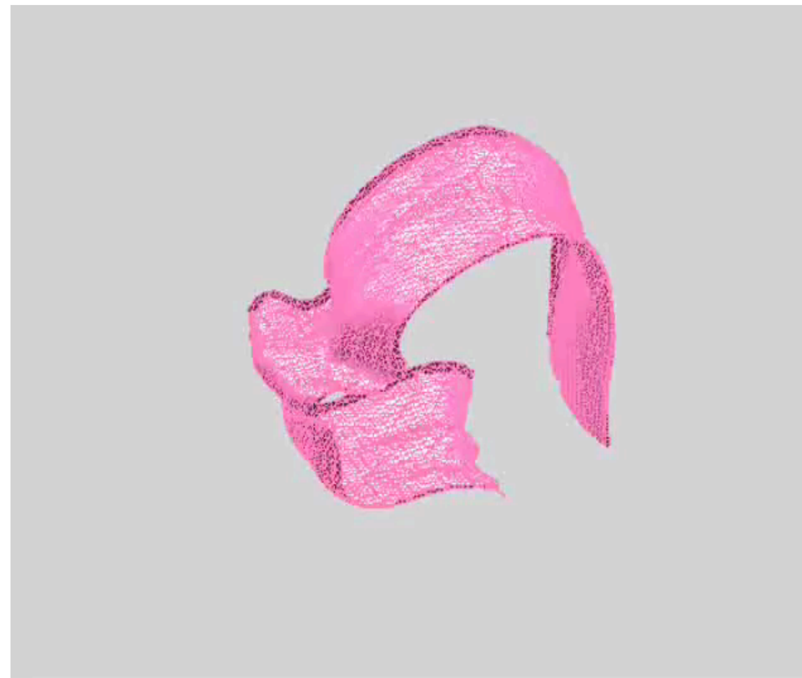


Yinghua Zhu, Yoon-Chul Kim, Michael Proctor, Shrikanth Narayanan, Krishna S. Nayak. Dynamic 3D Visualization of Vocal Tract Shaping during Speech. IEEE Transactions on Medical Imaging. 32(5): 838 - 848, May 2013.

Dynamic 3D visualization



Coronal movie — Aligned — Unaligned



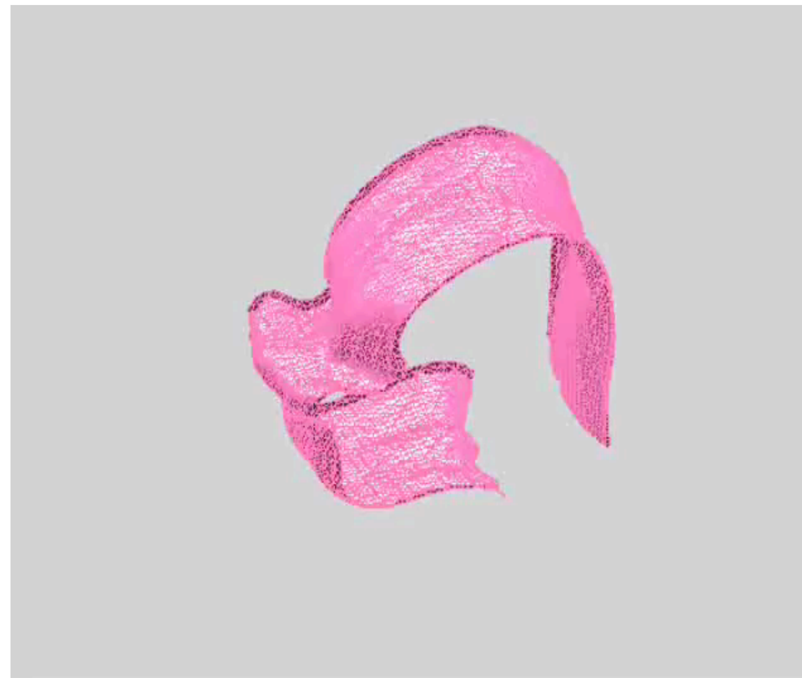
Yinghua Zhu, Yoon-Chul Kim, Michael Proctor, Shrikanth Narayanan, Krishna S. Nayak. Dynamic 3D Visualization of Vocal Tract Shaping during Speech. IEEE Transactions on Medical Imaging. 32(5): 838 - 848, May 2013.

Dynamic 3D visualization

Coronal movie

Aligned

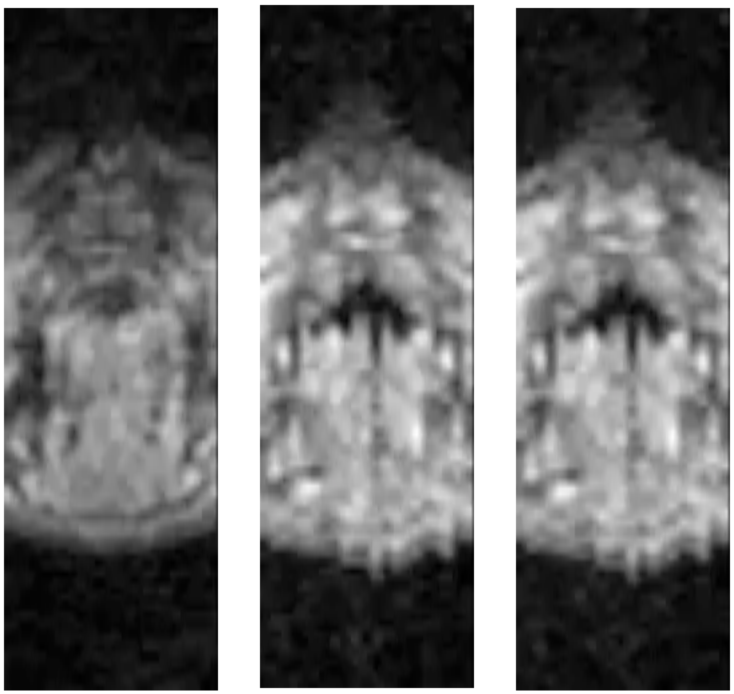
Unaligned



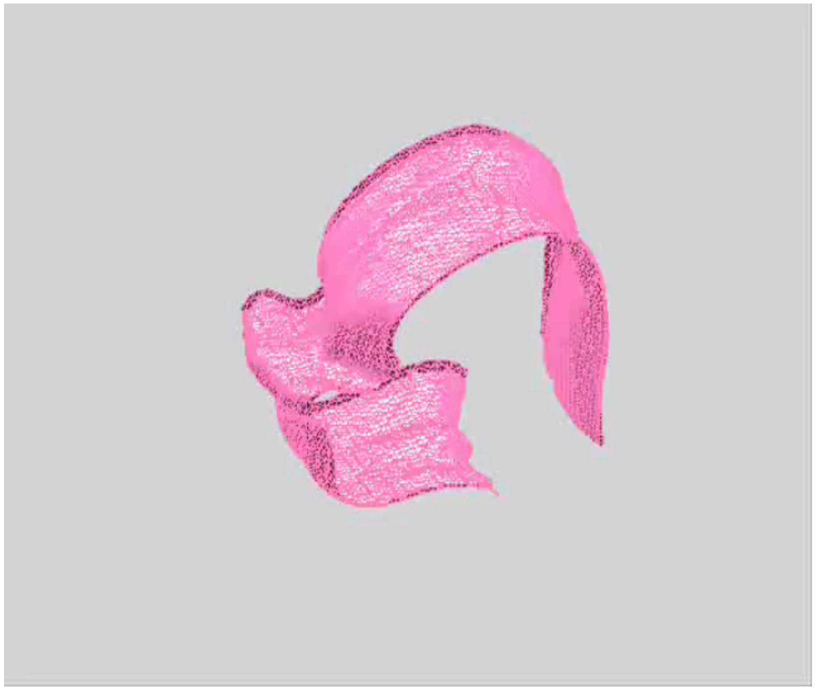
Yinghua Zhu, Yoon-Chul Kim, Michael Proctor, Shrikanth Narayanan, Krishna S. Nayak. Dynamic 3D Visualization of Vocal Tract Shaping during Speech. IEEE Transactions on Medical Imaging. 32(5): 838 - 848, May 2013.

18

Dynamic 3D visualization



Coronal movie   Aligned   Unaligned



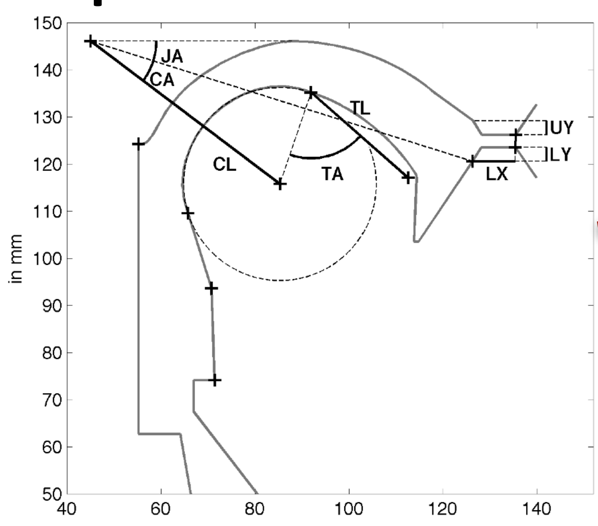
Yinghua Zhu, Yoon-Chul Kim, Michael Proctor, Shrikanth Narayanan, Krishna S. Nayak. Dynamic 3D Visualization of Vocal Tract Shaping during Speech. IEEE Transactions on Medical Imaging. 32(5): 838 - 848, May 2013.

TADA-TIMIT: A SIMULATED ARTICULATORY CORPUS

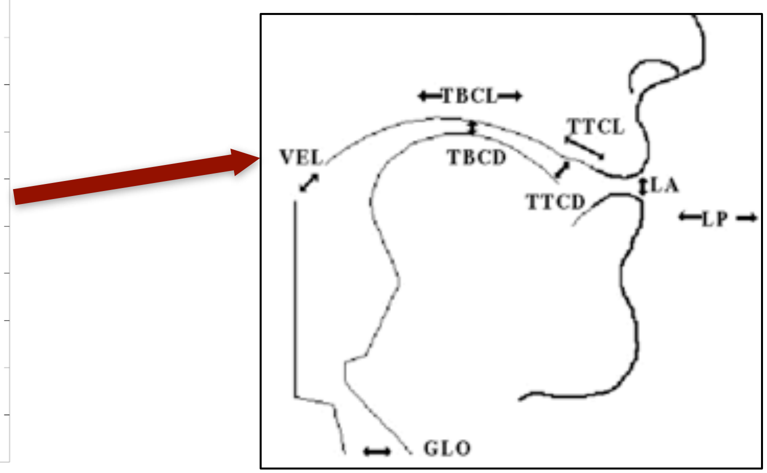
Contains: Morphological and kinematical descriptions, dynamical systems parameters, control signals

Useful for: Validation, development, comparison

postural variables



constriction tasks



H. Nam, L. Goldstein, E. Saltzman, and D. Byrd. TADA: An enhanced, portable task dynamics model in Matlab. Journal of the Acoustical Society of America, 115(5):2430–2430, 2004.

What can we do with all these data?

Seek confirmatory/deeper/newer insights into well known questions in linguistics, speech science with traditional methods

Segmental speech characteristics

Tongue shaping of English sibilant fricatives /s/ and /sh/ in various vowel contexts

Sagittal

Coronal

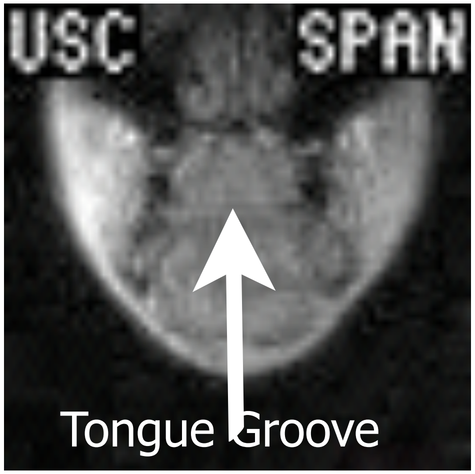
"Go pasop ok. Go pashop ok."

There are more stimuli in corpus: "paseep", "peesop", "peeseep" etc.

Segmental speech characteristics

Tongue shaping of English sibilant fricatives /s/ and /sh/ in various vowel contexts

Sagittal



Coronal

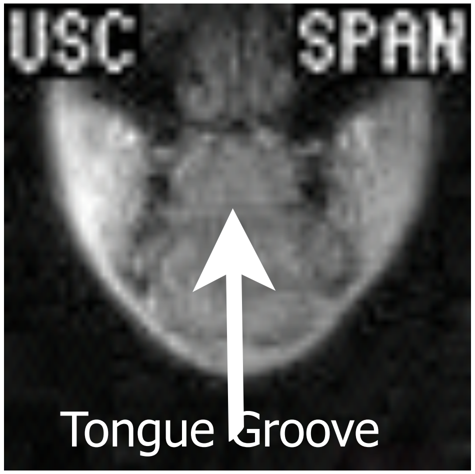
"Go pasop ok. Go pashop ok."

There are more stimuli in corpus: "paseep", "peesop", "peeseep" etc.

Segmental speech characteristics

Tongue shaping of English sibilant fricatives /s/ and /sh/ in various vowel contexts

Sagittal



Coronal

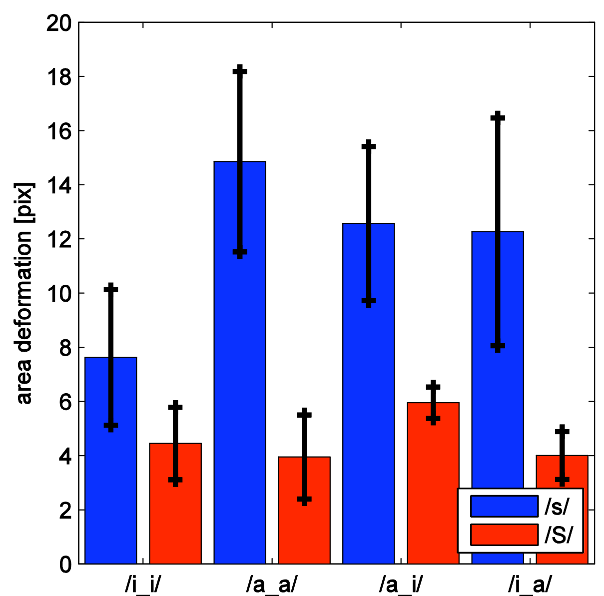
"Go pasop ok. Go pashop ok."

There are more stimuli in corpus: "paseep", "peesop", "peeseep" etc.

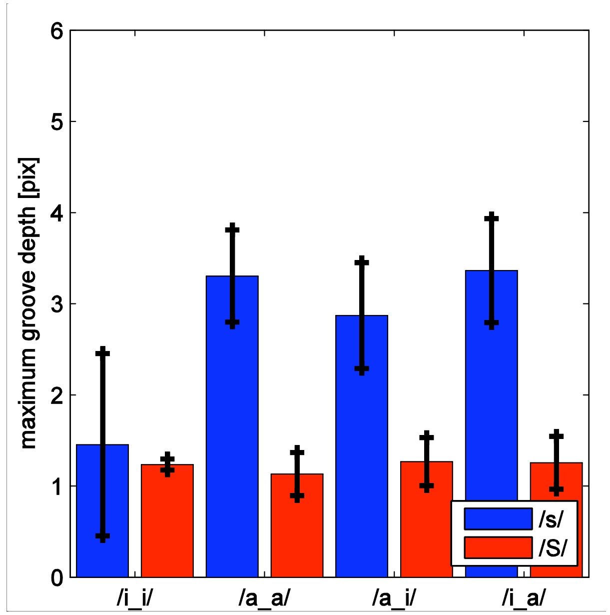
Research studies – quasi-static

Tongue shaping of English sibilant fricatives /s/ and /sh/ in various vowel contexts

Some findings:



Tongue surface for /sh/ is more parallel to the palate than for /s/.



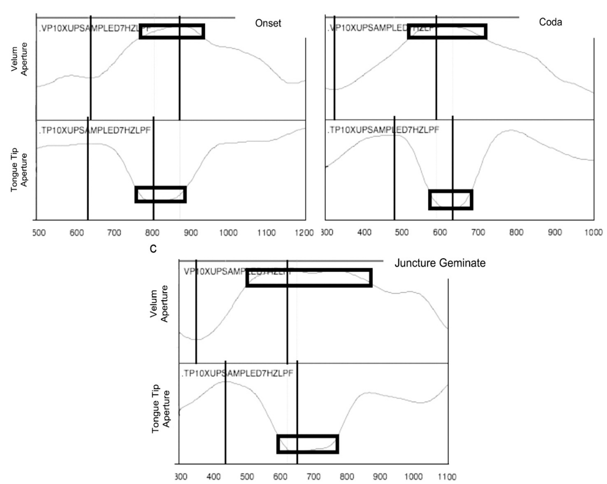
/s/ has a deeper tongue groove than /sh/.

Erik Bresch, Daylen Riggs, Louis Goldstein, Dani Byrd, Sungbok Lee, Shrikanth Narayanan. An analysis of vocal tract shaping in English sibilant fricatives using real-time magnetic resonance imaging. Proceedings of Interspeech 2008.

Dynamic characteristics

coordination between adjacent segments and linguistic structure..?

Velum-oral coordination of English nasals



Systematic timing differences between tongue and velum constriction forming events?

Direct observation of tongue and velum tract variables TTCD, VEL

Nasal position

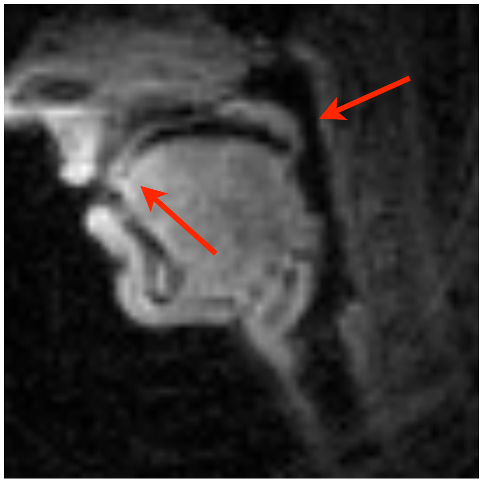
- Onset: /bow-know/, /toe-node/
- Coda: /bone-oh/, /tone-ode/
- Juncture geminate: /bone-know/, /tone-node/

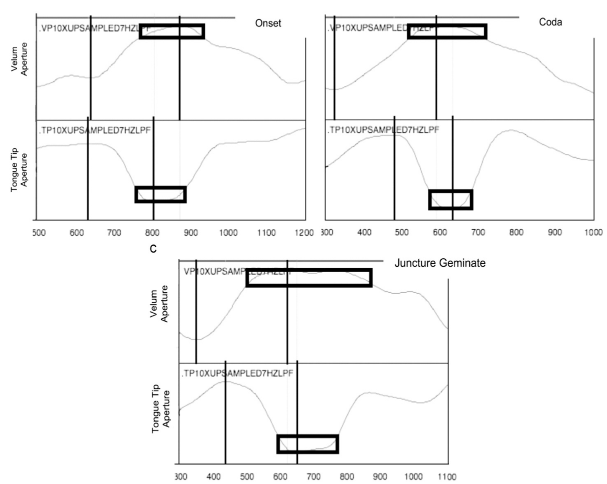
- ❖ Data processing
  - Segment stimuli from carrier
  - Trace vocal tract
  - Measure VEL, TTCD constriction degree time series
- ❖ Define timing criteria
  - Time lag, e.g., w.r.t. 95% threshold
- ❖ Evaluate statistical significance of lag measures

Dynamic characteristics

coordination between adjacent segments and linguistic structure..?

Velum-oral coordination of English nasals





Systematic timing differences between tongue and velum constriction forming events?

Direct observation of tongue and velum tract variables TTCD, VEL

Nasal position

Onset: /bow-know/, /toe-node/

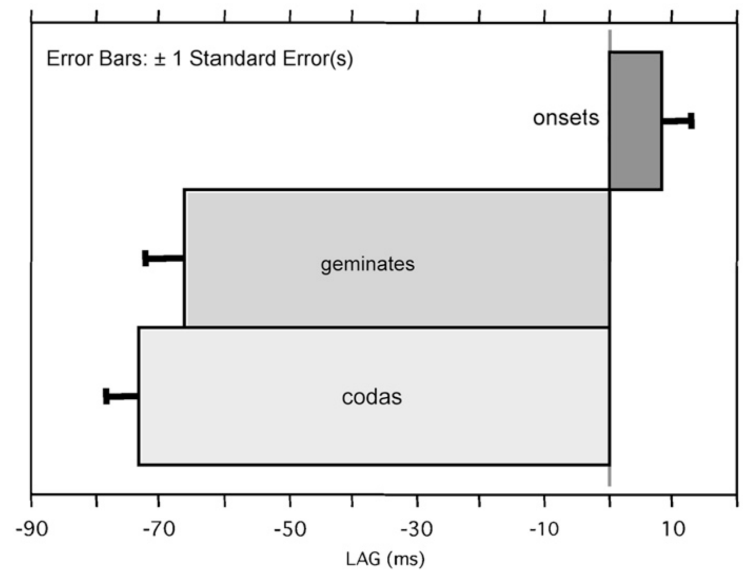
Coda: /bone-oh/, /tone-ode/

Juncture geminate: /bone-know/, /tone-node/

- Data processing
  - Segment stimuli from carrier
  - Trace vocal tract
  - Measure VEL, TTCD constriction degree time series
- Define timing criteria
  - Time lag, e.g., w.r.t. 95% threshold
- Evaluate statistical significance of lag measures

Results

Velum-oral coordination of English nasals



The velum opening lags behind tongue tip closure if the nasal is in onset position.

Intergestural timing patterns sensitive to local stress context ==>
Underlying timing specification that can yield flexibly

D. Byrd, S. Tobin, E. Bresch, and S. Narayanan. Timing effects of syllable structure and stress on nasals: a real-time MRI examination. J. Phonetics. 37: 97–110, 2009.

25

Allows exploration of novel data-driven and hybrid knowledge-inspired approaches & models

Rest of the talk

Deriving articulatory representations

Direct methods

Raw measures

Derived task measures

Inverse methods

Some case studies

Vocal tract morphology

Articulatory setting

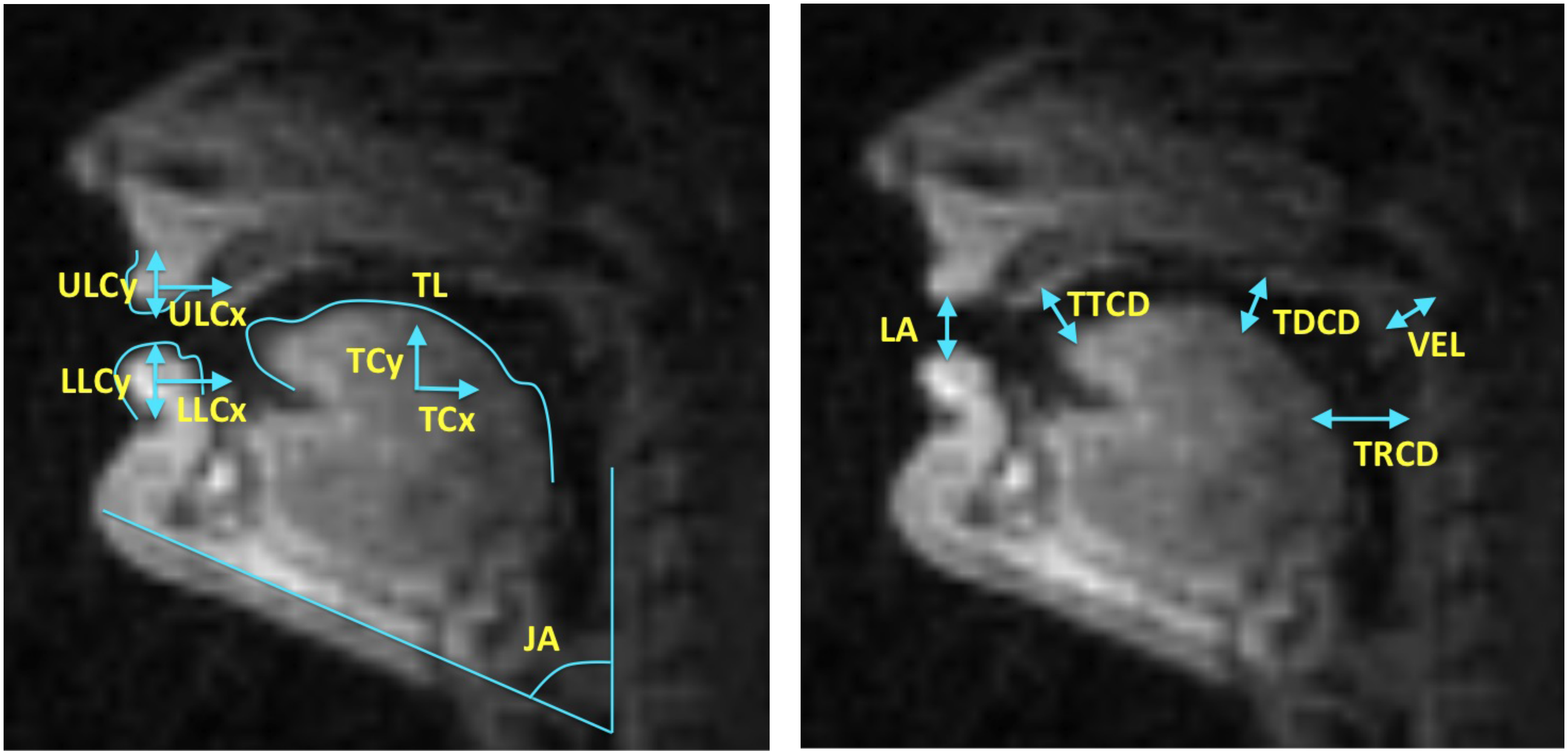
Relation between articulatory & acoustic representations

ASR and Speaker Verification

Back to basics: learning from data

DIRECT MEASURES FROM DATA

ARTICULATORY POSTURE & CONSTRICTION TASK VARIABLES

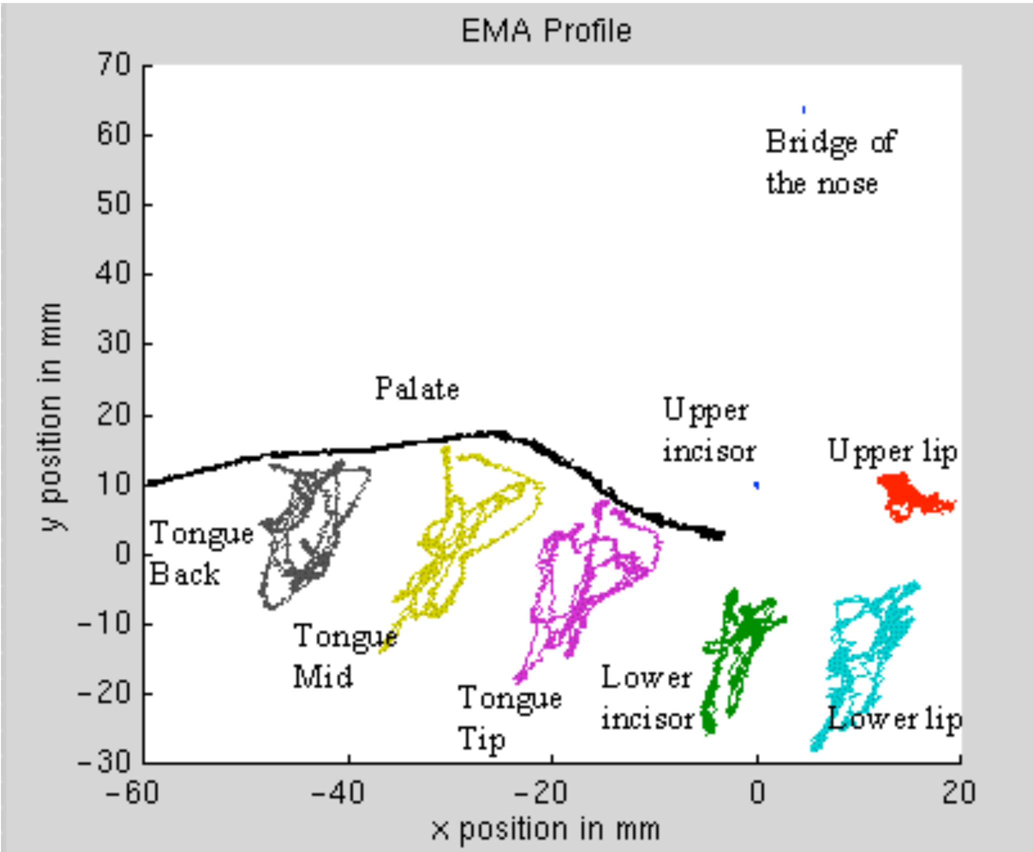


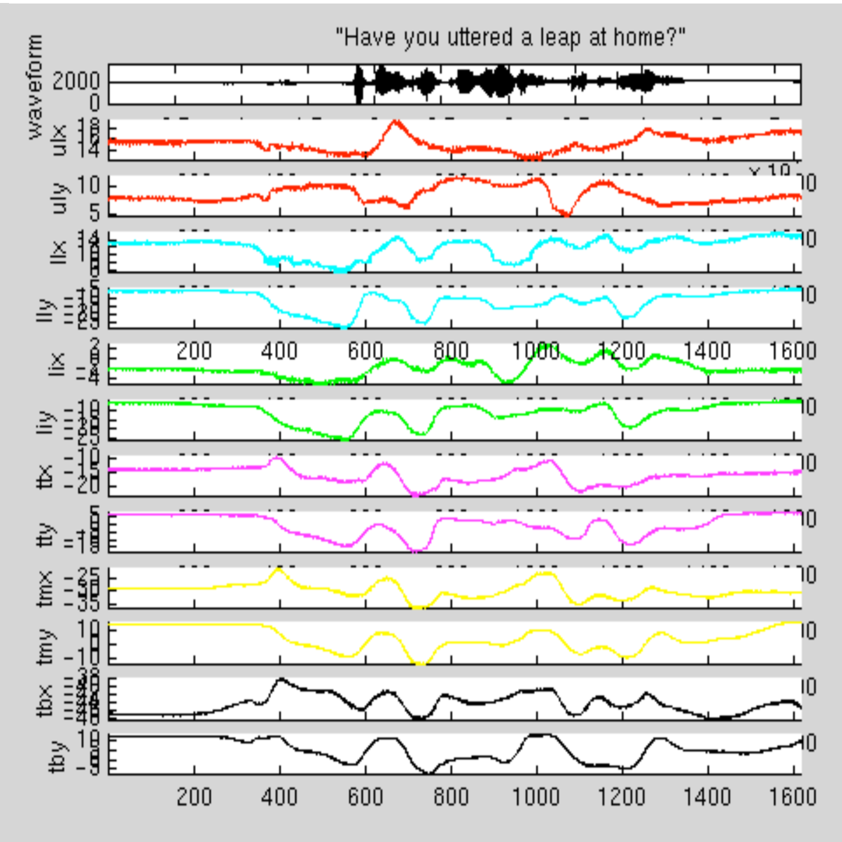
These feature sets are useful for modeling speech production dynamics

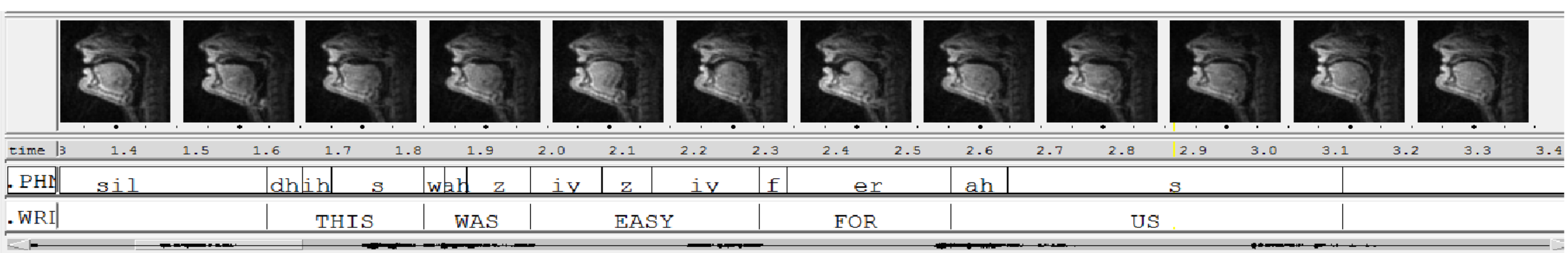
Adam Lammert, Louis Goldstein, Shrikanth Narayanan and Khalil Iskarous. Statistical Methods for Estimation of Direct and Differential Kinematics of the Vocal Tract. Speech Communication. 55: 147–161, 2013.

Vikram Ramanarayanan, Adam Lammert, Louis Goldstein and Shrikanth Narayanan. Articulatory settings facilitate mechanically advantageous motor control of vocal tract articulators. In Proceedings of Interspeech, 2013

RAW MEASUREMENT FEATURES







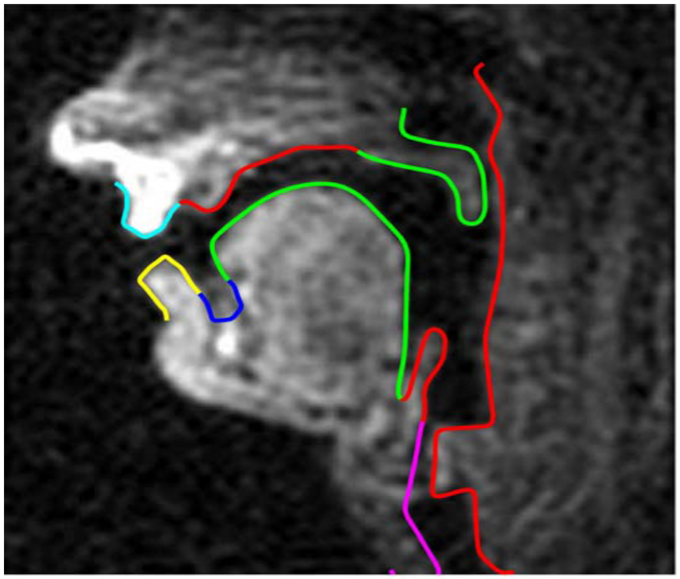
VOCAL TRACT CONTOURS

MODEL-BASED IMAGE SEGMENTATION IN THE FOURIER DOMAIN

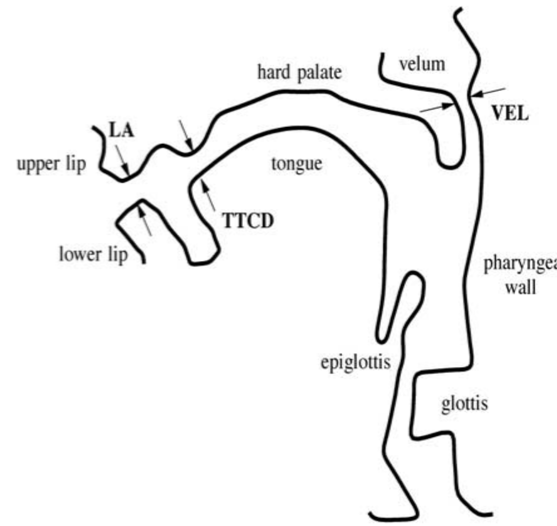
Erik Bresch and Shrikanth Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Transactions on Medical Imaging. 28(3): 323--338, March 2009.

VOCAL TRACT CONTOURS

MODEL-BASED IMAGE SEGMENTATION IN THE FOURIER DOMAIN



(a)



(b)

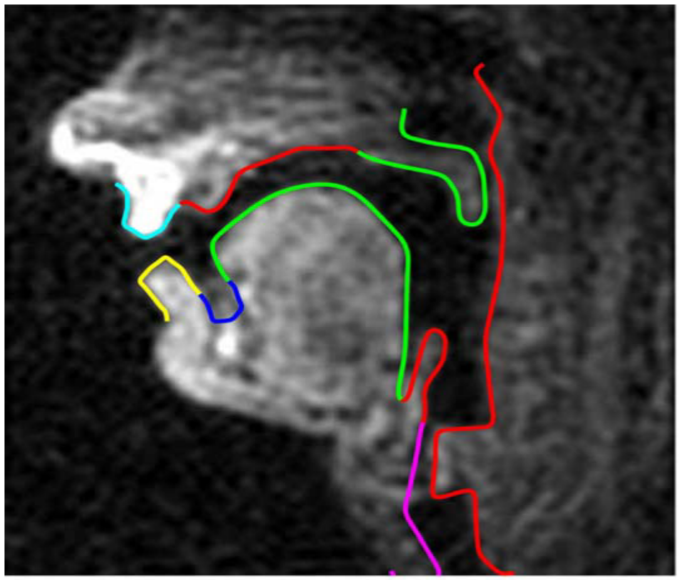
First define a contour model segmentation *manually* : each articulator in a different color

Erik Bresch and Shrikanth Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Transactions on Medical Imaging. 28(3): 323--338, March 2009.

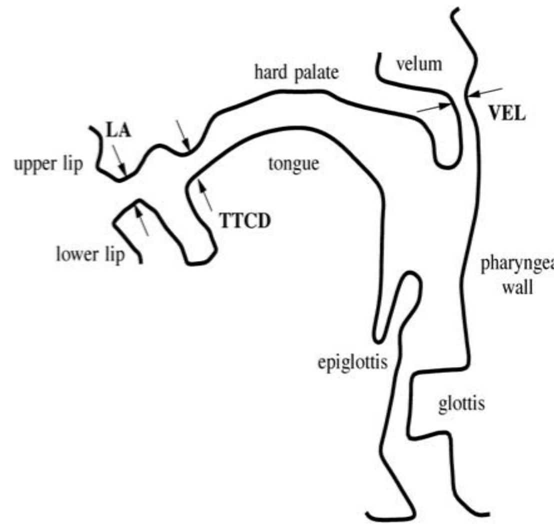
31

VOCAL TRACT CONTOURS

MODEL-BASED IMAGE SEGMENTATION IN THE FOURIER DOMAIN



(a)



(b)

First define a contour model segmentation *manually* : each articulator in a different color

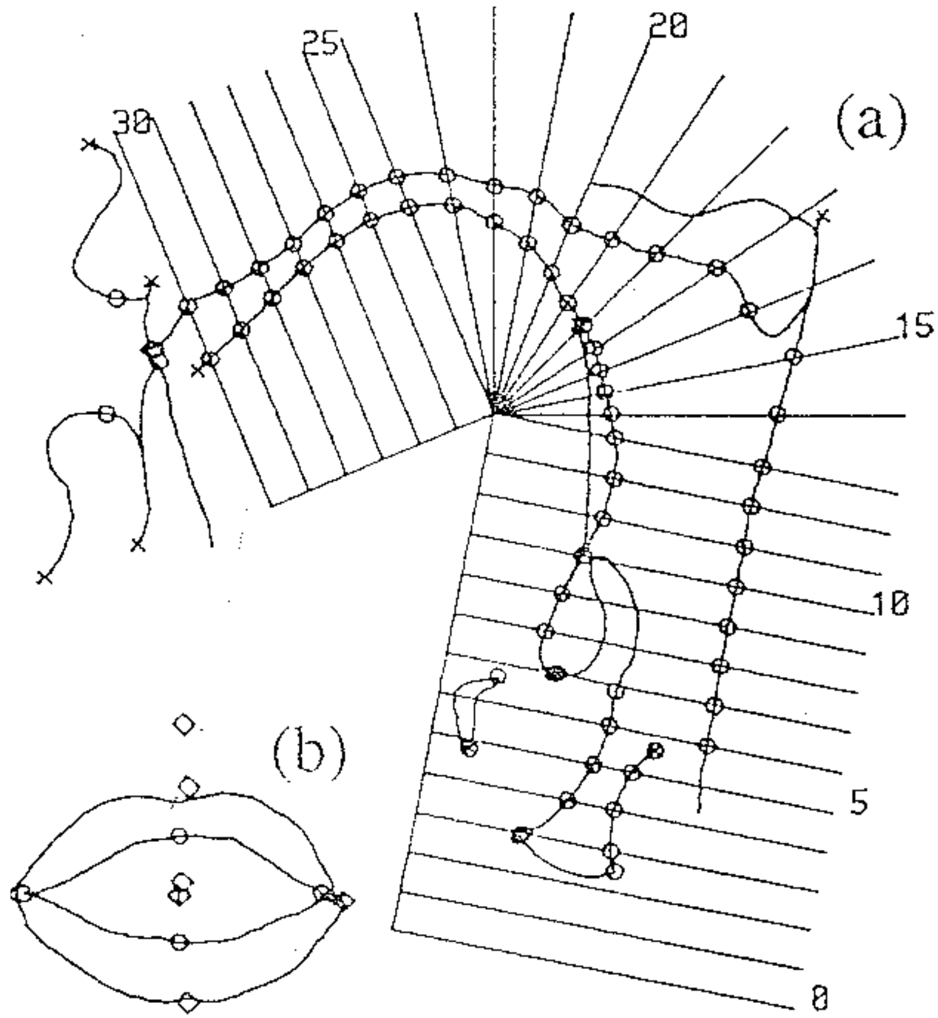
Now **hierarchically optimize** the model fit to the image in the Fourier domain using **gradient descent**!

Erik Bresch and Shrikanth Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Transactions on Medical Imaging. 28(3): 323--338, March 2009.

PARAMETRIZATION: WHAT HAS BEEN DONE?

PARAMETRIZATION: WHAT HAS BEEN DONE?

Ohman (1966), Mermelstein (1973) and Maeda (1990) proposed the use of semi-polar grids superimposed on the vocal tract

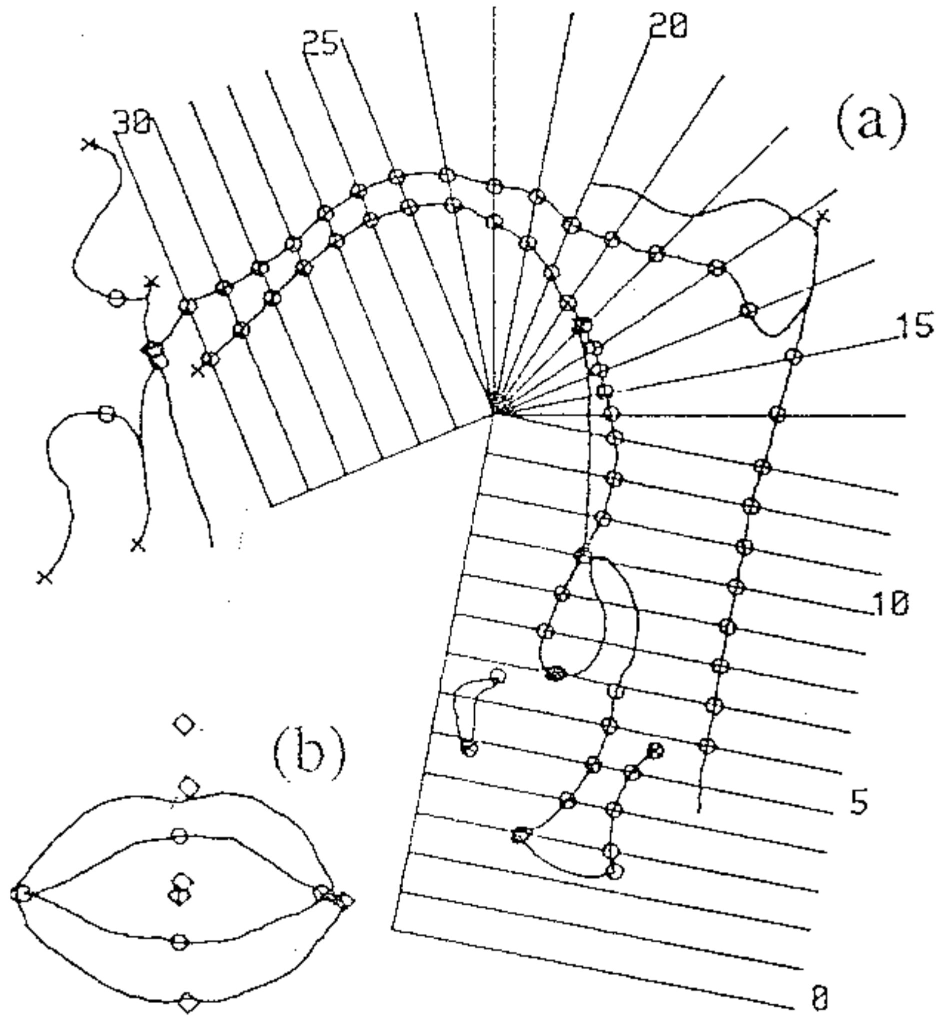


32

PARAMETRIZATION: WHAT HAS BEEN DONE?

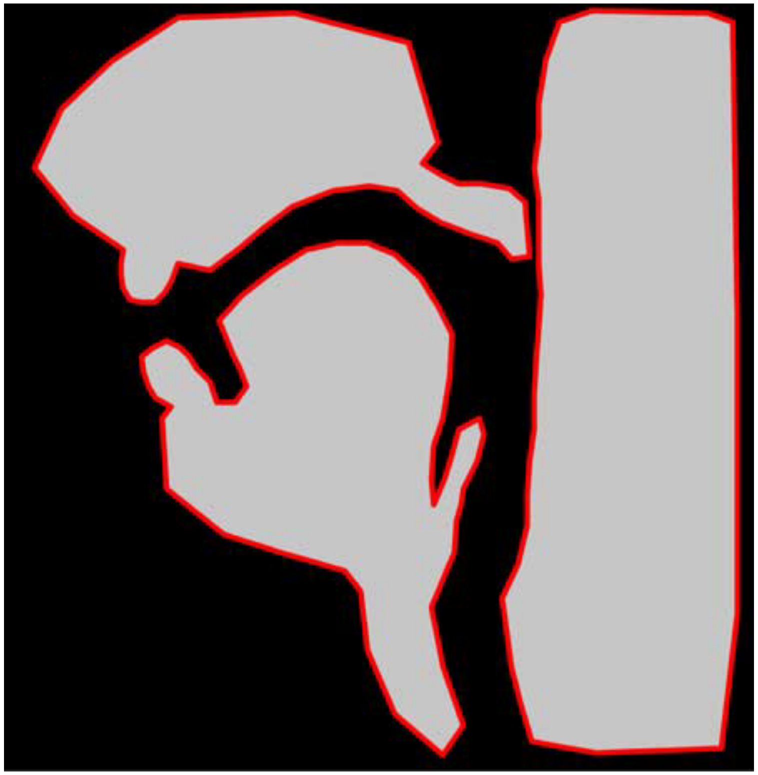
Ohman (1966), Mermelstein (1973) and Maeda (1990) proposed the use of semi-polar grids superimposed on the vocal tract

But these require manual intervention and are not comparable across subjects.



WHAT ARE DESIRABLE PROPERTIES?

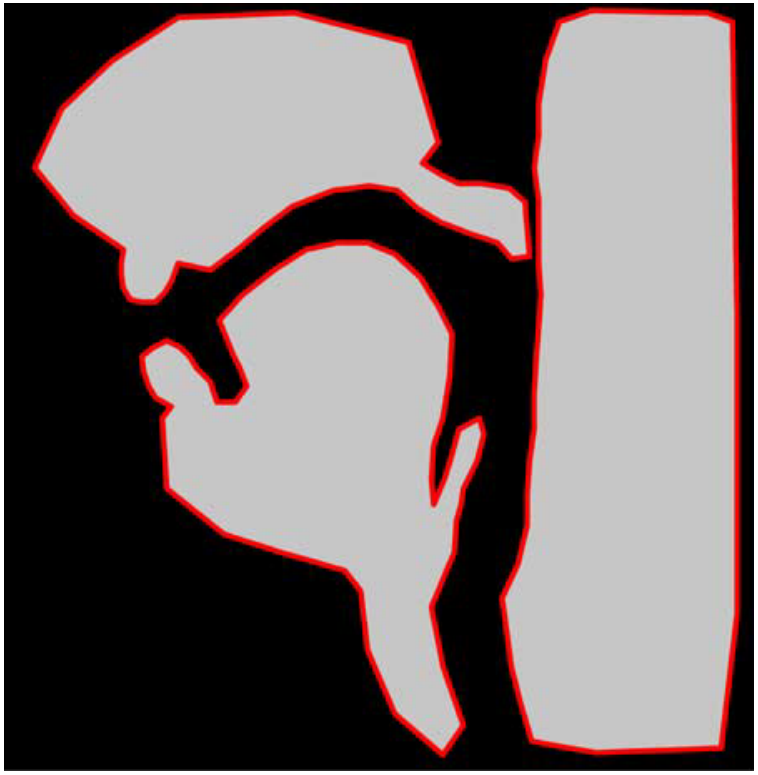
WHAT ARE DESIRABLE PROPERTIES?



33

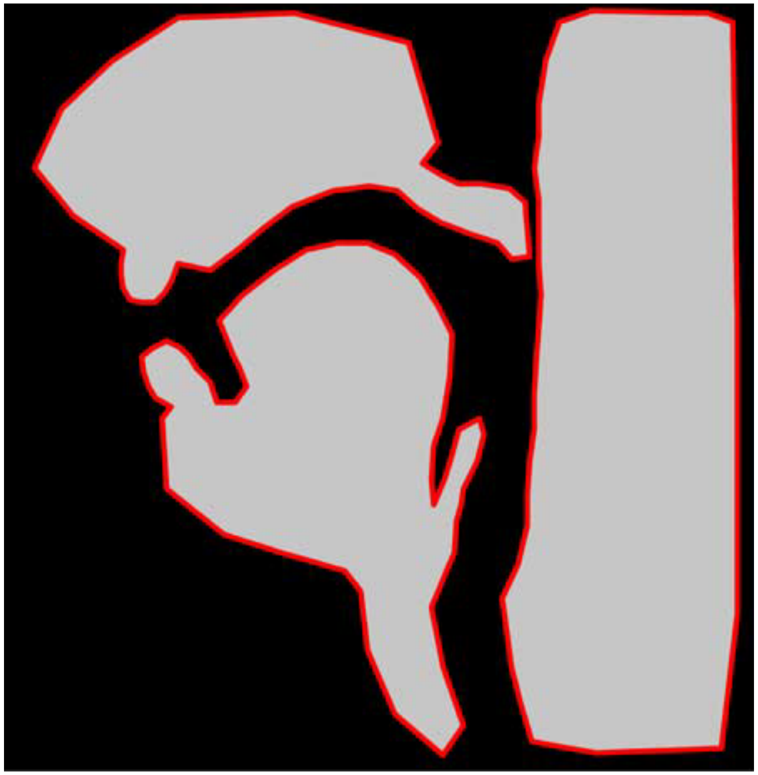
WHAT ARE DESIRABLE PROPERTIES?

1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure



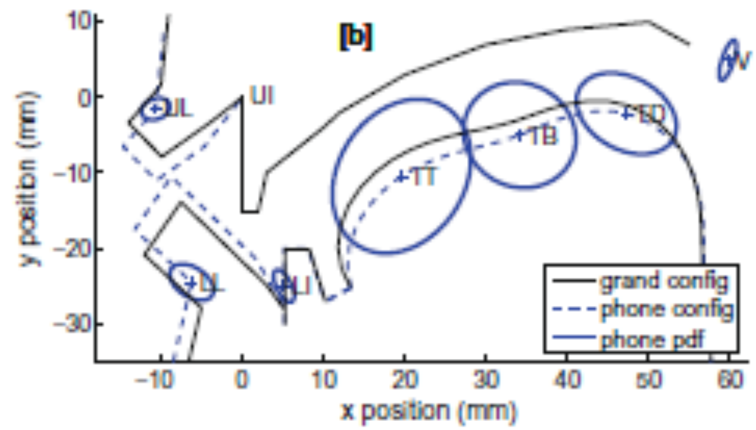
WHAT ARE DESIRABLE PROPERTIES?

1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure
2. they should sufficiently characterize vocal tract postures



WHAT ARE DESIRABLE PROPERTIES?

1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure
2. they should sufficiently characterize vocal tract postures



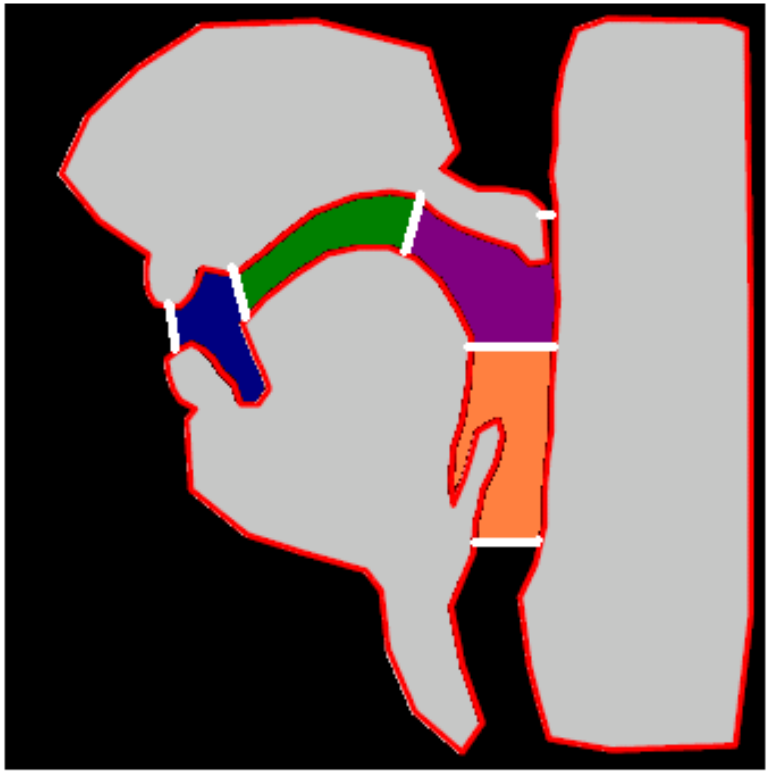
Critical articulator behavior – constrained

Dependent and *redundant* articulators – NOT constrained!

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.

WHAT ARE DESIRABLE PROPERTIES?

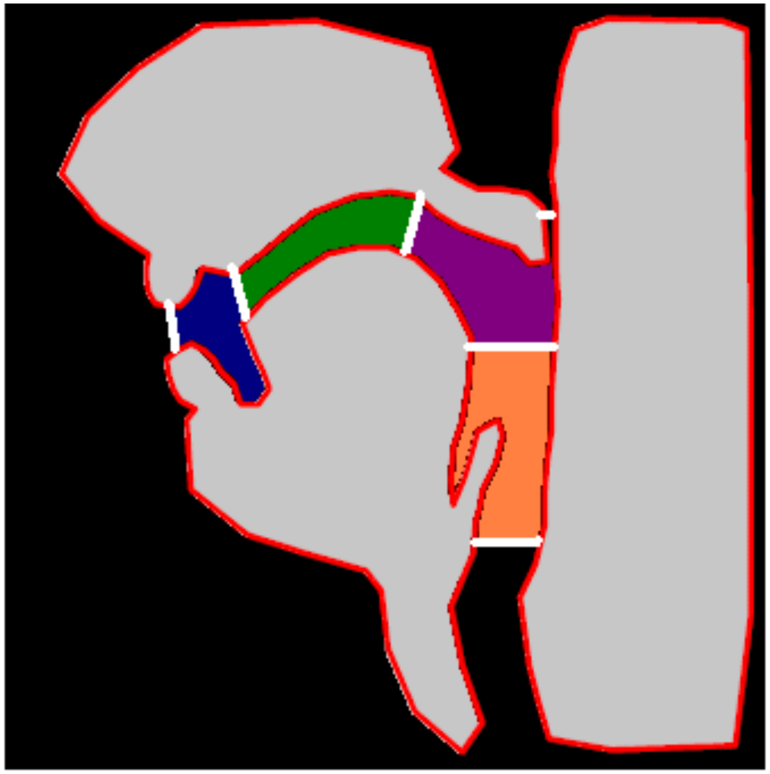
1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure
2. they should sufficiently characterize vocal tract postures



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

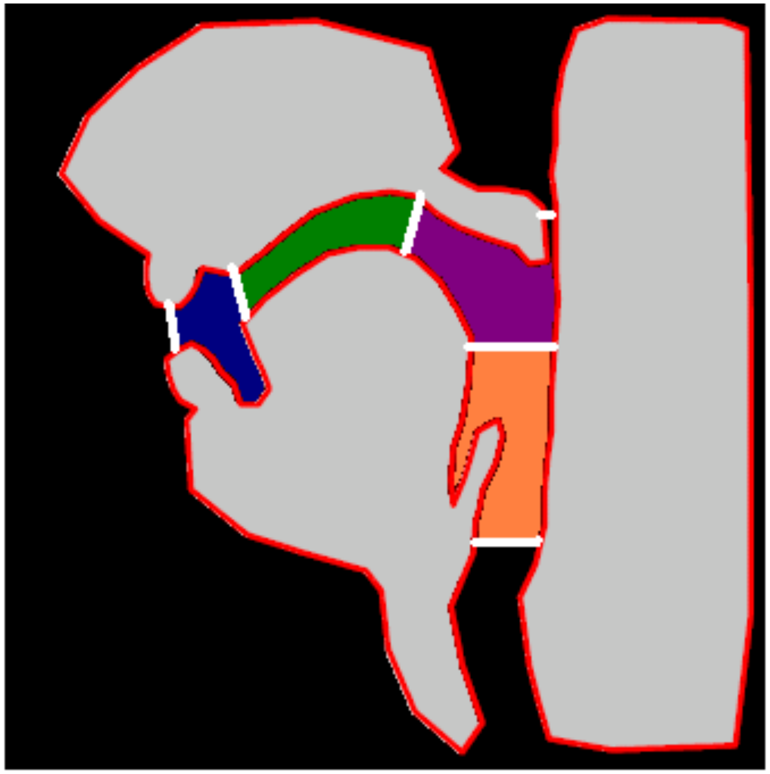
WHAT ARE DESIRABLE PROPERTIES?

1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure
2. they should sufficiently characterize vocal tract postures
3. should allow for meaningful comparison across speakers



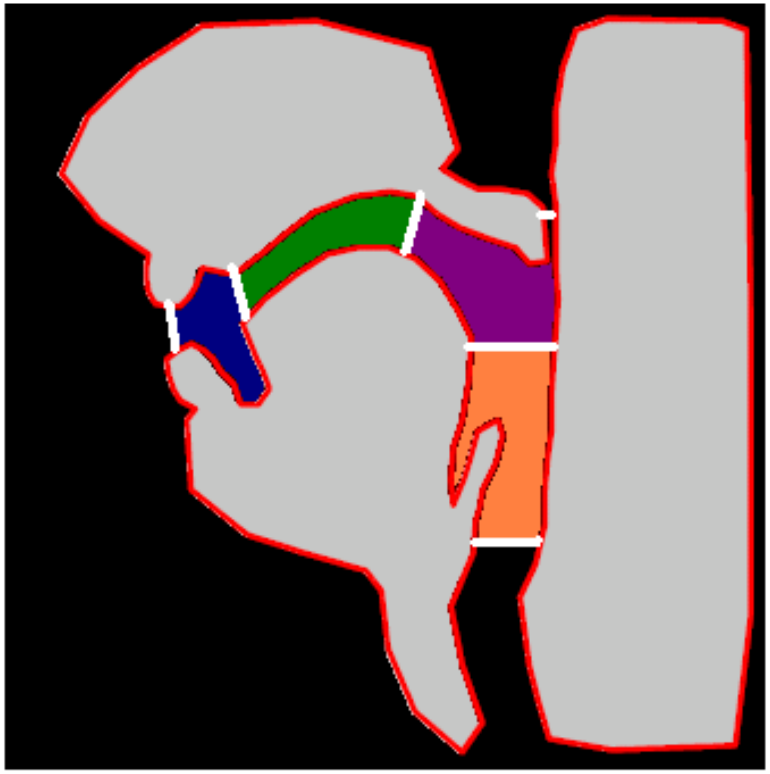
WHAT ARE DESIRABLE PROPERTIES?

1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure
2. they should sufficiently characterize vocal tract postures
3. should allow for meaningful comparison across speakers
4. they should involve as little manual intervention as possible



WHAT ARE DESIRABLE PROPERTIES?

1. robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure
2. they should sufficiently characterize vocal tract postures
3. should allow for meaningful comparison across speakers
4. they should involve as little manual intervention as possible



Idea: First compute meaningful cross-distances, then the areas bounded by them!