An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping

Chao Qin and Miguel Á. Carreira-Perpiñán

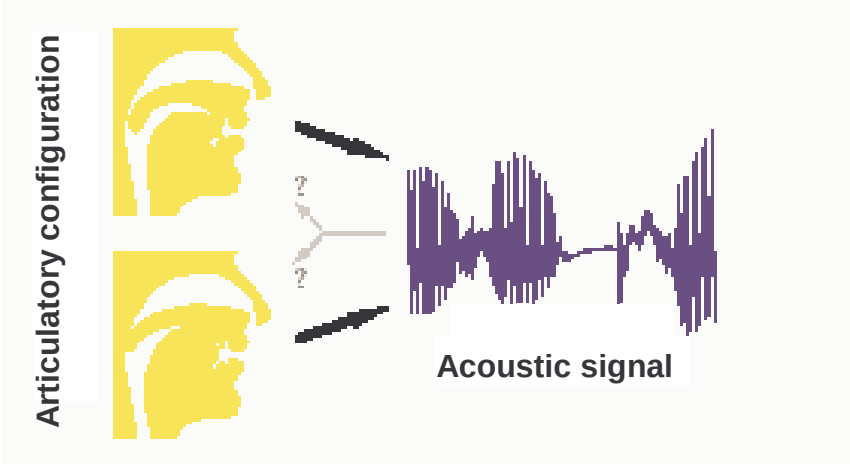
Dept. of Computer Science & Electrical Engineering, OGI/OHSU

(from July 2007 at University of California, Merced)

http://www.csee.ogi.edu/~cqin

Interspeech 2007, Antwerp

Introduction



- Articulatory-to-acoustic (forward) mapping
  - Nonlinear and uni-valued mappings
- Acoustic-to-articulatory (inverse) mapping
  - Multi-valued mappings or nonuniqueness
- Applications
  - Improve speech recognition, synthesis, and coding
  - Provide visual aid for language learning

Standard arguments for nonuniqueness

- Ventriloquist
- Webster's horn equation

$$\frac{d}{dx}\left(A(x)\cdot\frac{dP(x,s)}{dx}\right)-\frac{s^2}{c^2}A(x)\cdot P(x,s)=0$$

$A(x)$: area function

$P(x,s)$: pressure

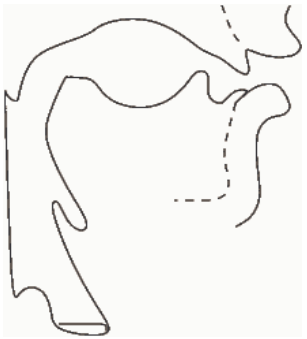
c : velocity of sound

s: complex freq. variable

- Bite block and compensation of articulators
- Realization of the consonant /r/ in American English (Espy-Wilson *et al* '00)

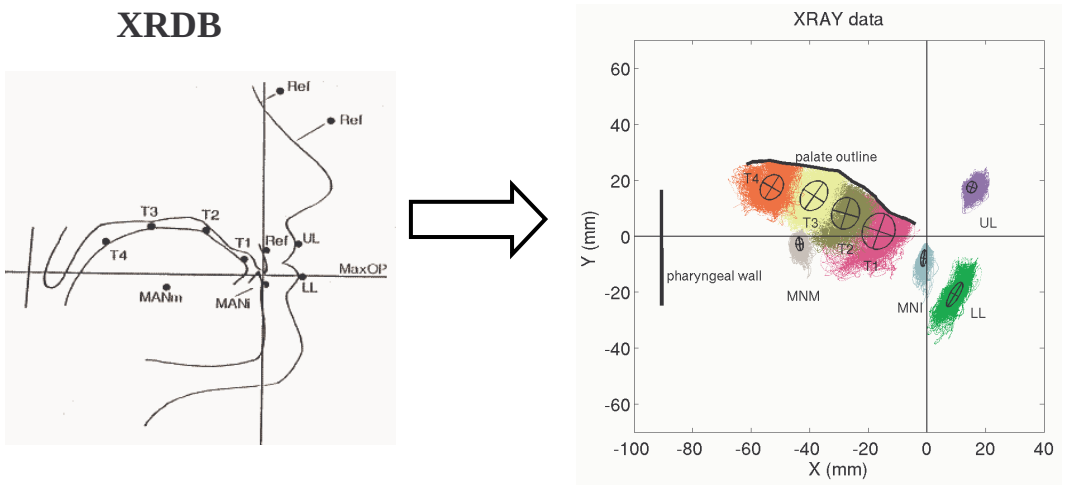
Retroflex

Bunched



Empirical investigation of nonuniqueness

- Evidence for nonuniqueness came from
  - Atypical speech
  - Theoretical study
  - Very specific sound
- How does nonuniqueness happen in normal speech?
- Wisconsin X-ray microbeam database *(*Westbury'94*)*
  - Simultaneous audio + pellet movements



Methodological setup

- Dataset $\{x_n, y_n\}_{n=1}^{N}$
  - one male speaker *jw11* from XRDB
- Articulatory feature x
  - horizontal-vertical coordinates of pellets
- Acoustic feature y
  - 20-order LPC
- Idea
  - Given $\{y_m\}$ forming a uni-modal dist.,

    if $\{x_m\}$ form a multi-modal dist.,

    instantaneous inverse mapping is nonunique/multi-valued

    else

    instantaneous inverse mapping is unique/uni-valued

    end

Systematic search for multimodality

- Algorithm

$k \leftarrow 0$ # of nonuniqueness in db

for $n \in \{1, ..., N\}$

  search in the db $\{x_m\}$ that map approximately to y_n Inversion

  determine uni/multi- modality of $\{x_m\}$ by # of modes Clustering

  if multi-modal $k \leftarrow k+1$; end Update

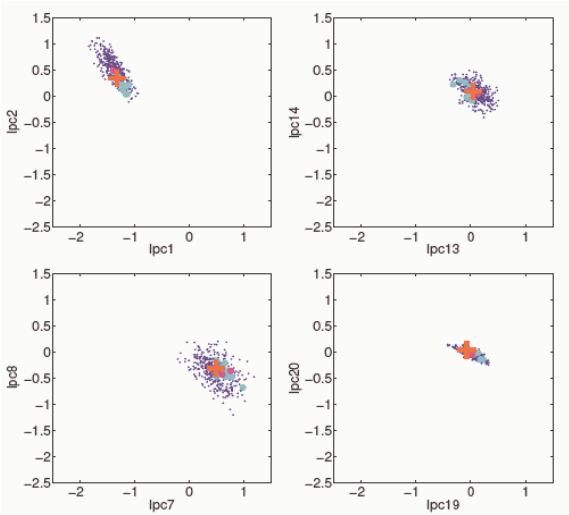
end

k / N % nonuniqueness in db

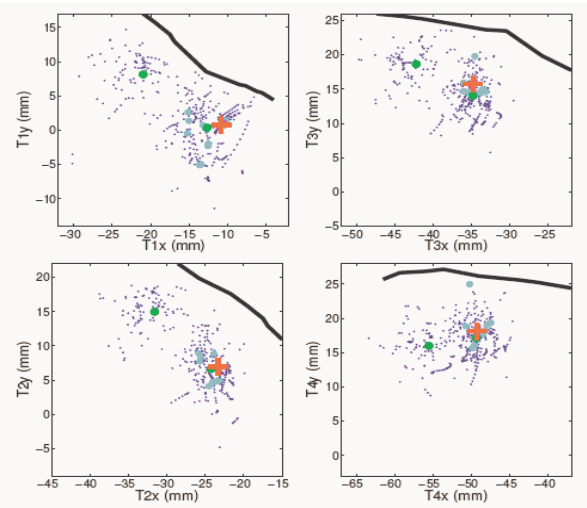
- Results
  - 5% acoustic vectors in the database yield multimodality, i.e., nonuniqueness occurs infrequently

Nonuniqueness in instantaneous inverse mapping

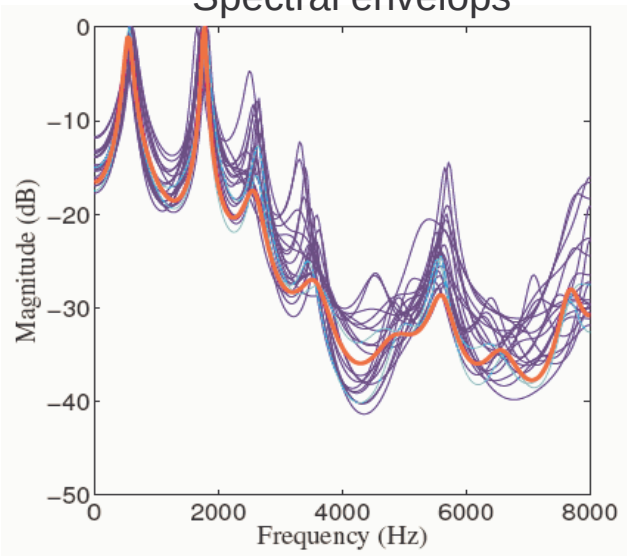
Acoustic space



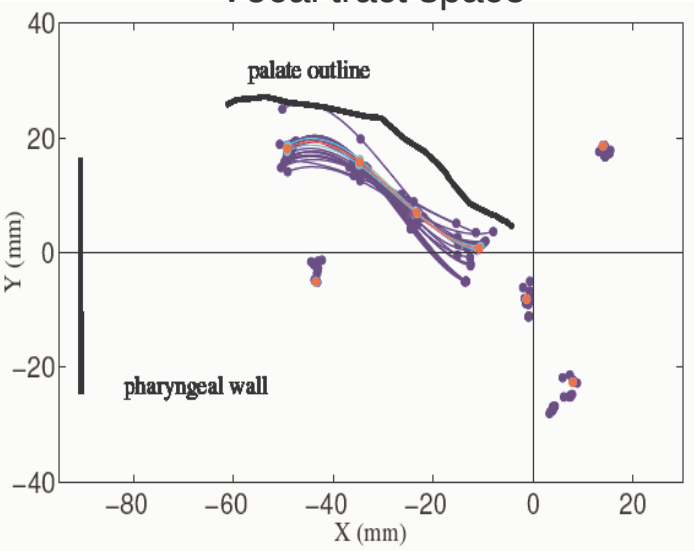
Articulatory space



Spectral envelops

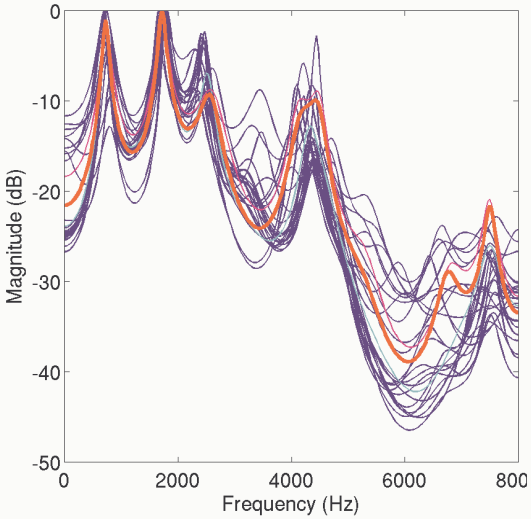


Vocal tract space

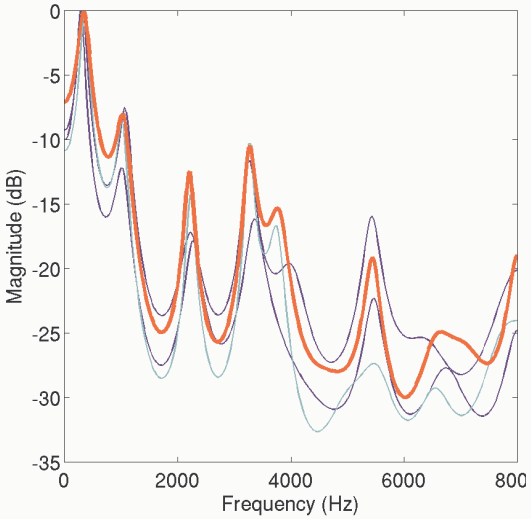


Sounds which do not show multimodality

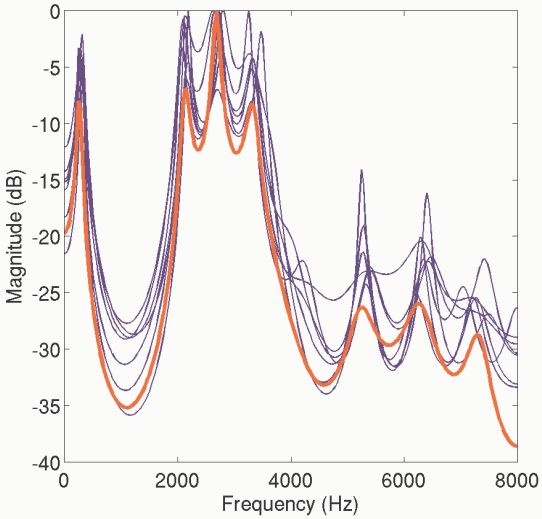
/ae/ in has

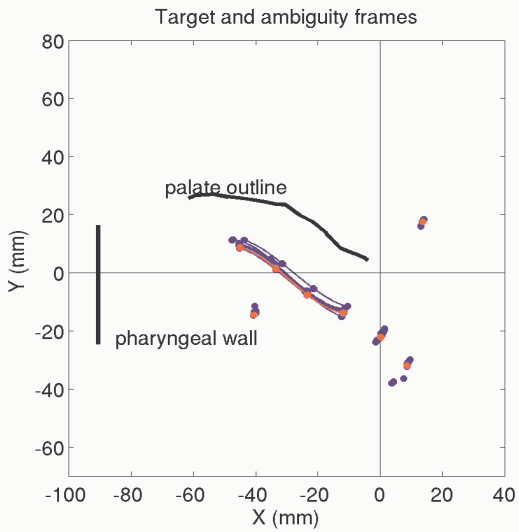


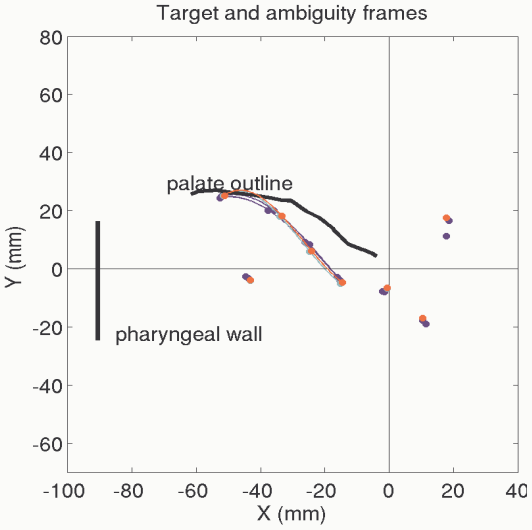
/uw/ in school

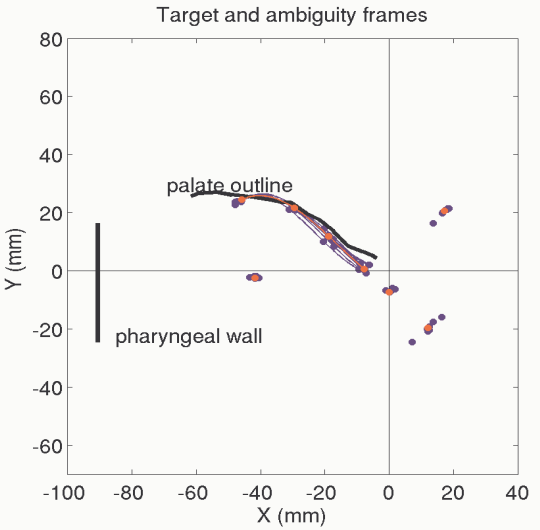


/y/ in you



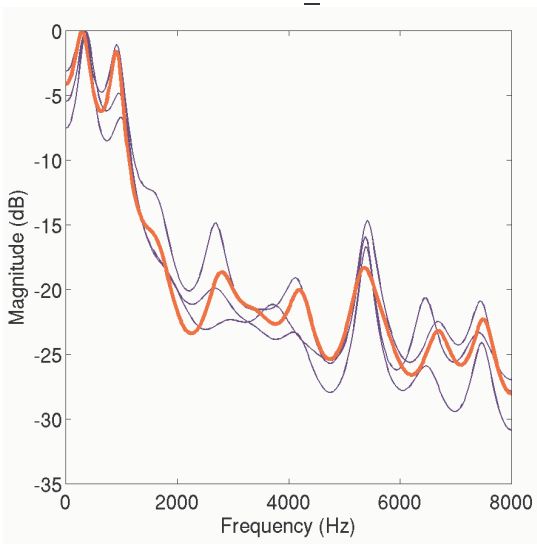


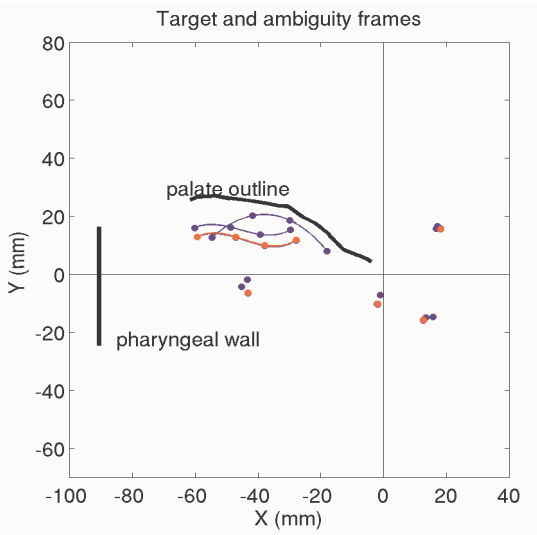




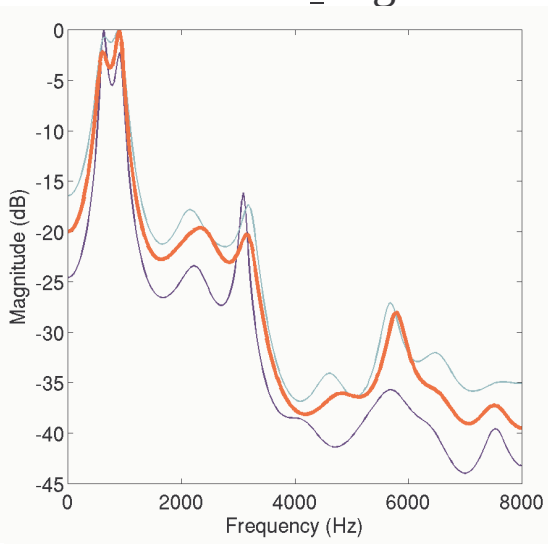
Sounds which show multimodality

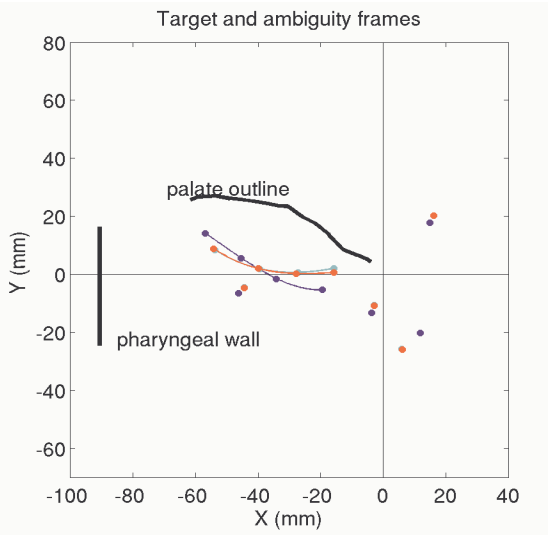
/r/ in row



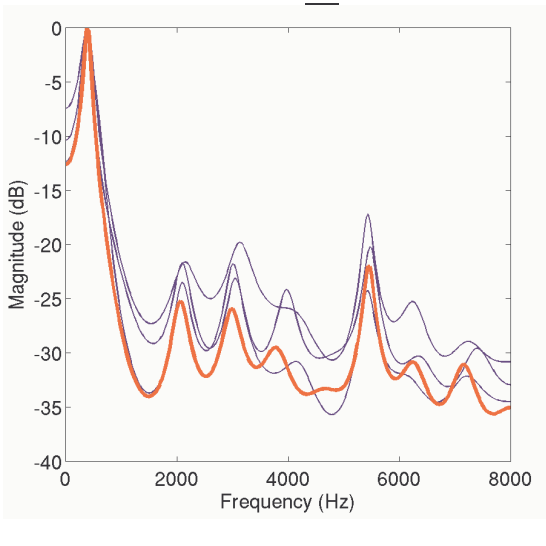


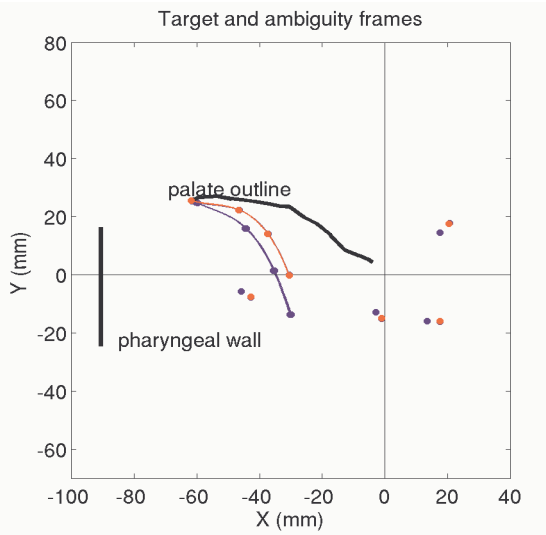
/l/ in long





/w/ in work





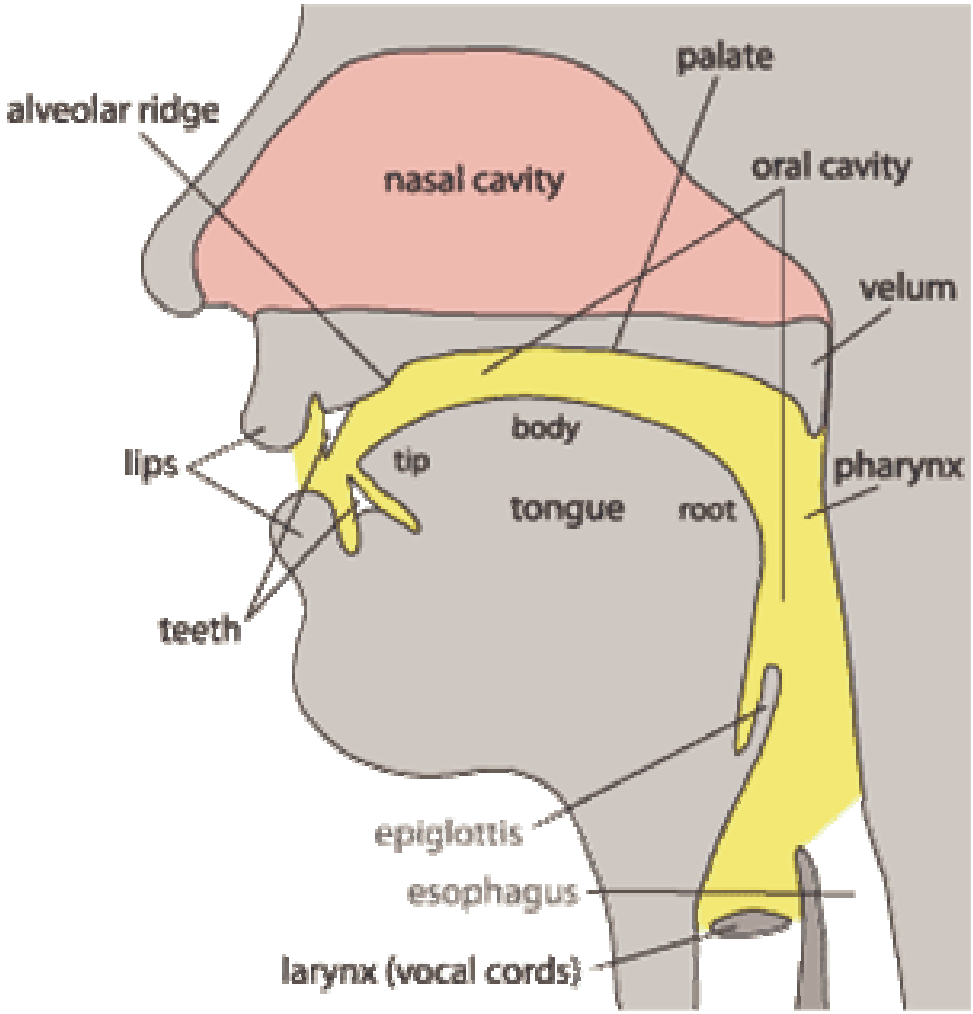
Conclusions

- Nonuniqueness does exist in instantaneous inverse mapping (in normal speech) but happens infrequently.
- Sound /r/, /l/, /w/ can be pronounced in multiple ways
- Limitations
  - Considered only a single speaker from X-ray
  - Did not consider the acoustic context
  - Incomplete (sparse) representation of the vocal tract

Acknowledgement

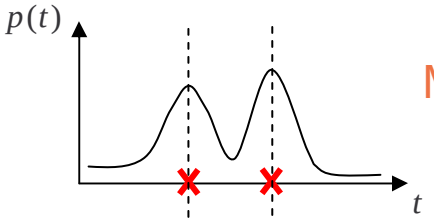
- MACP and CQIN thank Korin Richmond for valuable discussions
- MACP and CQIN thank John Westbury for XRDB
- Supported by NSF CAREER award IIS-0546857

Mid-sagittal view of vocal tract



Mode-finding for Gaussian mixture

- Given a Gaussian mixture $p(t) = \sum_{m=1}^{M} p(m) \cdot p(t \mid m)$ with $\{\pi_m, \mu_m, \Sigma_m\}_{m=1}^{M}$

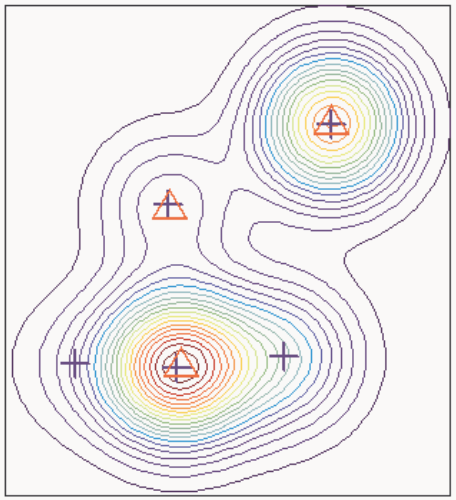


Modes satisfy: $\begin{cases} \nabla p(t) = 0 \\ (\nabla \nabla^T) p(t) < 0 \end{cases}$

- Let $\nabla p(t) = 0$, we can obtain the fixed-point iteration $t^{(\tau+1)} = f(t^{(\tau)})$

$$t^{(\tau+1)} = \sum_{m=1}^{M} p(m \mid t^{(\tau)}) \cdot \mu_m, \text{ if } \Sigma_m = \Sigma \text{ for } \forall m$$

- Start iterations from all component centers $\{\mu_m\}_{m=1}^{M}$



Carreira-Perpinan'99