Reconstructing the Full Tongue Contour from EMA/X-Ray Microbeam

Chao Qin

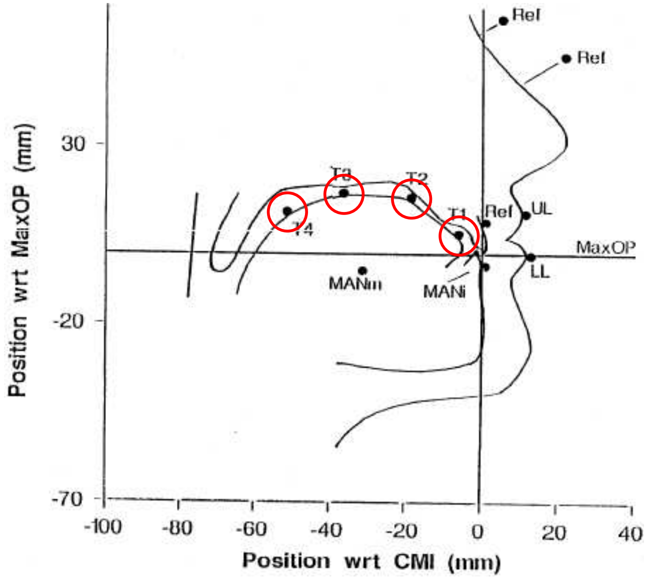
EECS, School of Engineering, UC Merced, USA

March 2010

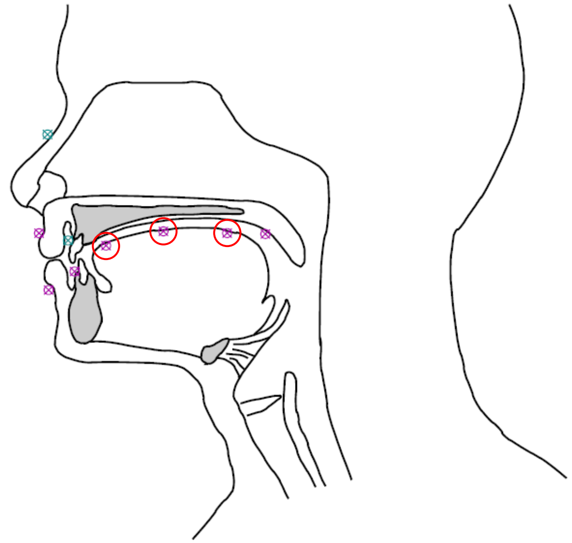
Introduction

- Tongue is the most important speech production articulator
- Articulatory datasets only provide sparse representation of tongue.

Wisconsin X-ray microbeam (XRMB)



MOCHA-TIMIT (MOCHA)



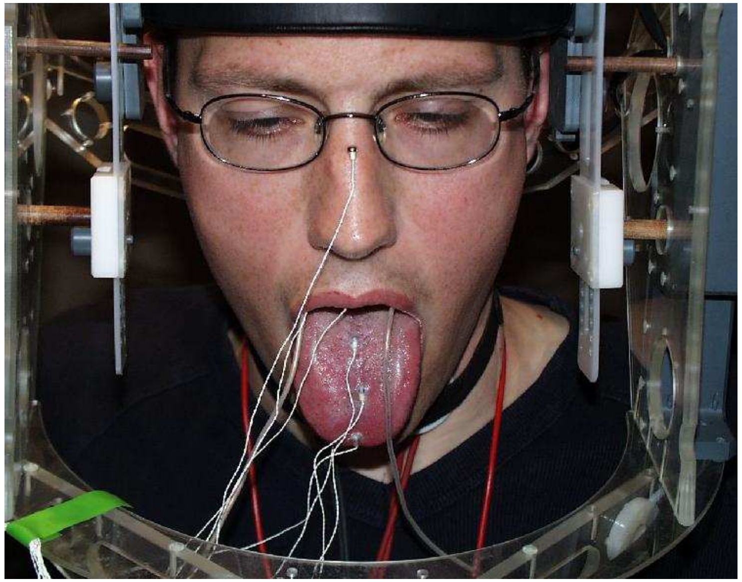
- Questions
  1. Can we reconstruct the realistic tongue shape from 3 or 4 pellets for an unknown speaker?
  2. Applications: synthesis and inversion

Multimodal fusion

Ultrasound

+

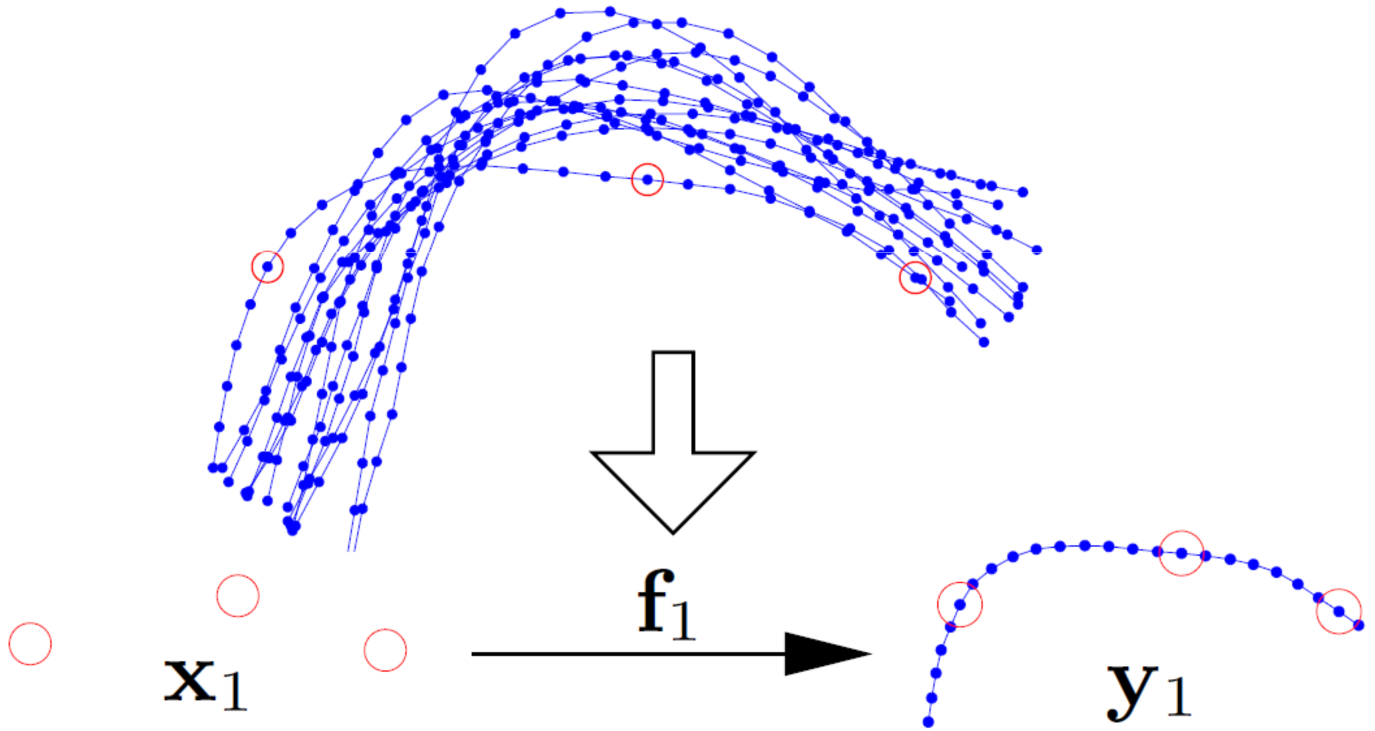
EMA



- Can we take advantage of both ultrasound and EMA recorded from different speakers and different sessions?
- Challenges
  - Speaker variability, eg. vocal tract length, tongue shape and length, etc

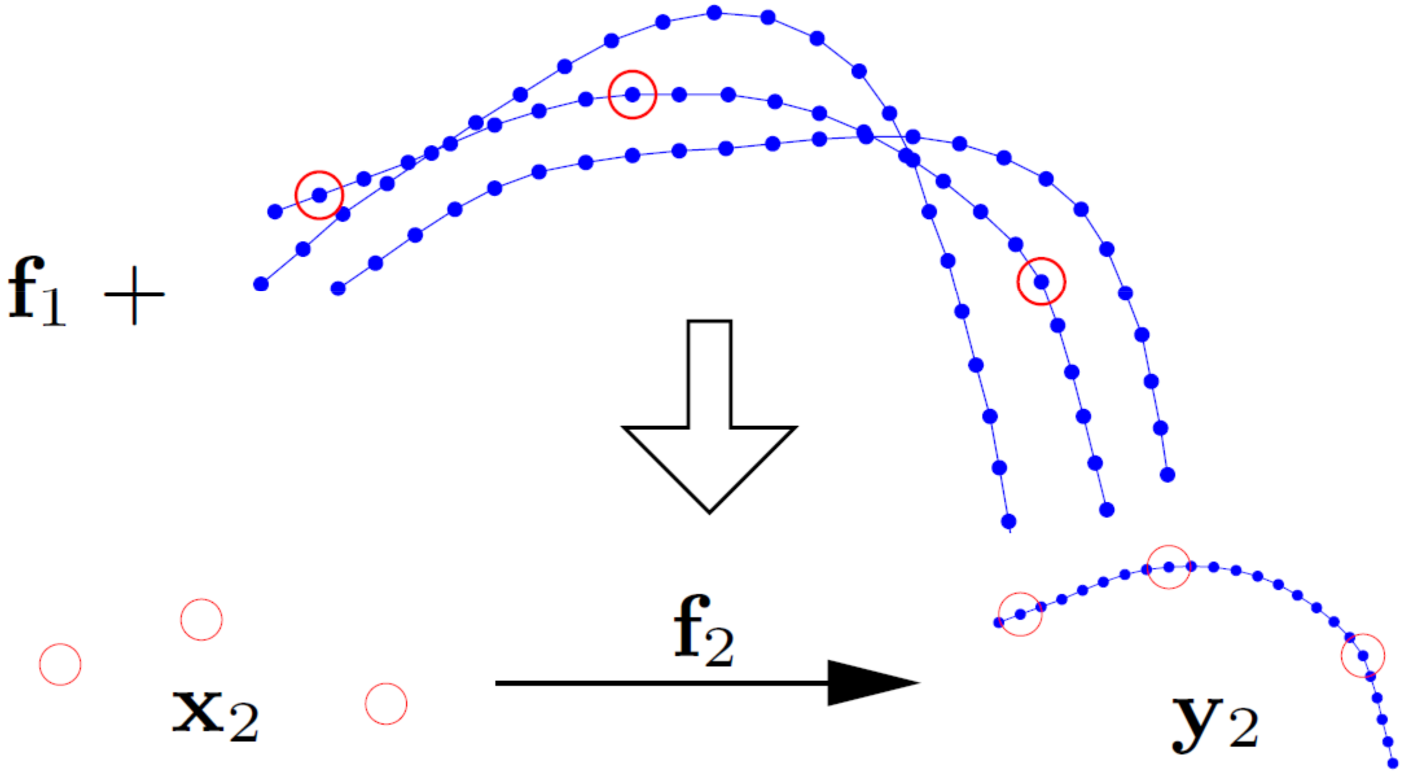
Data-driven approaches

P1: Training a predictive model $\mathbf{f}_1$ for speaker 1 with many full contours



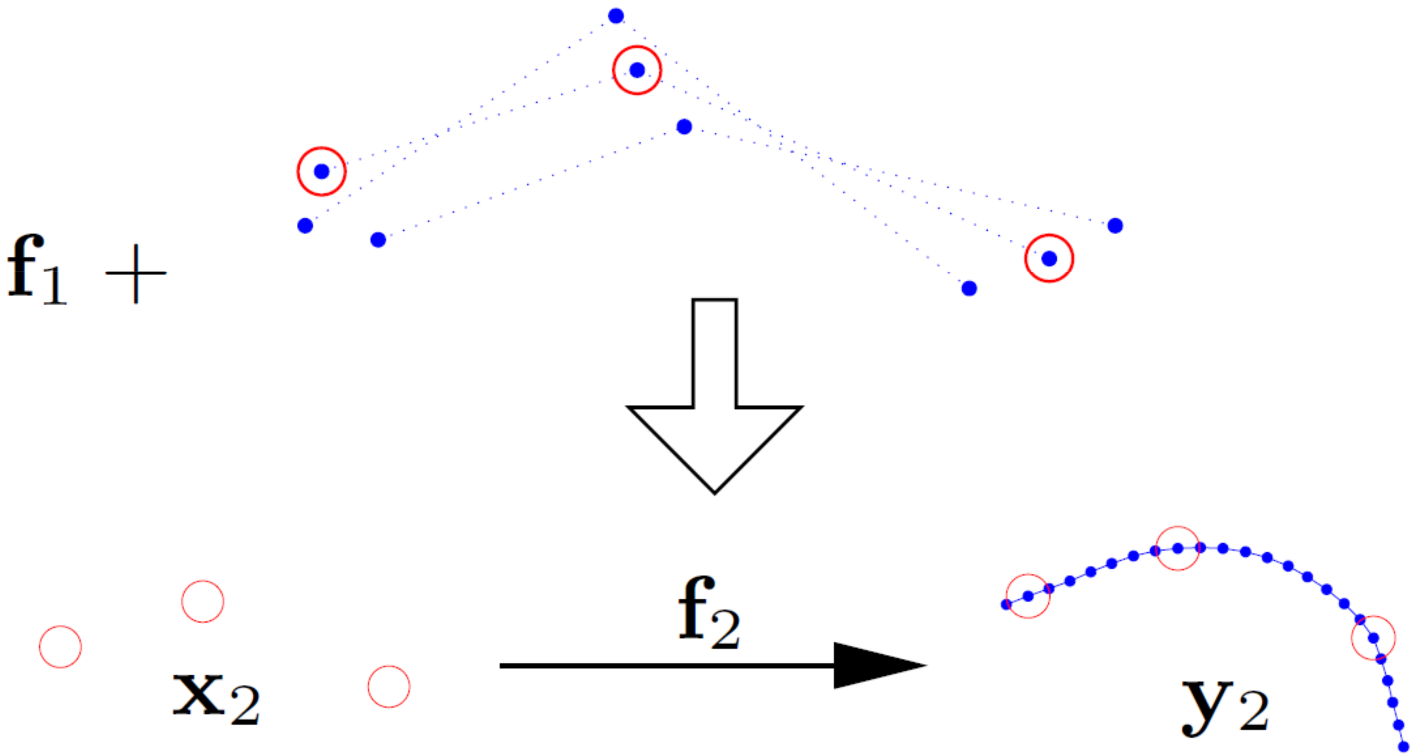
Data-driven approaches

P2: Adapting $\mathbf{f}_1$ to speaker 2 given a few full contours



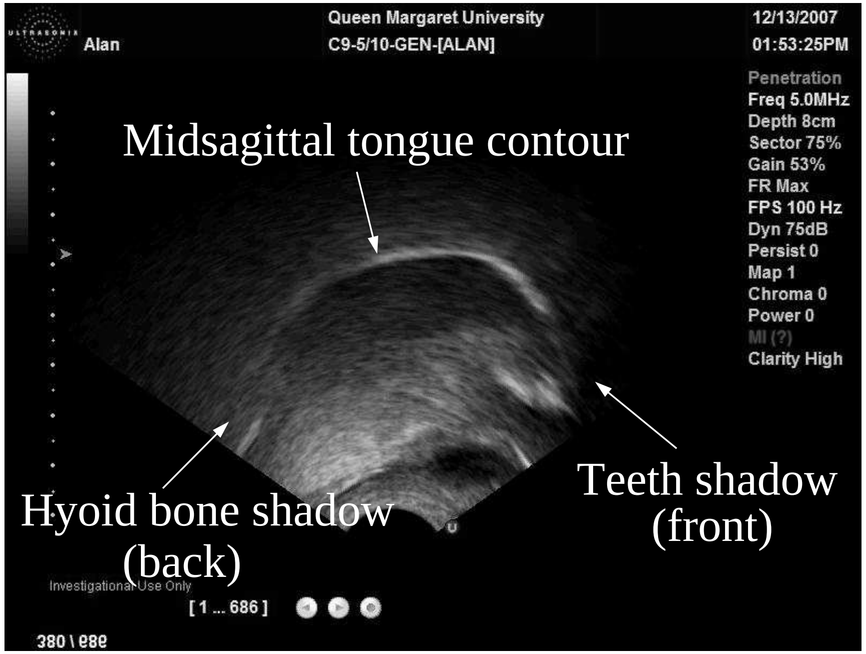
Data-driven approaches

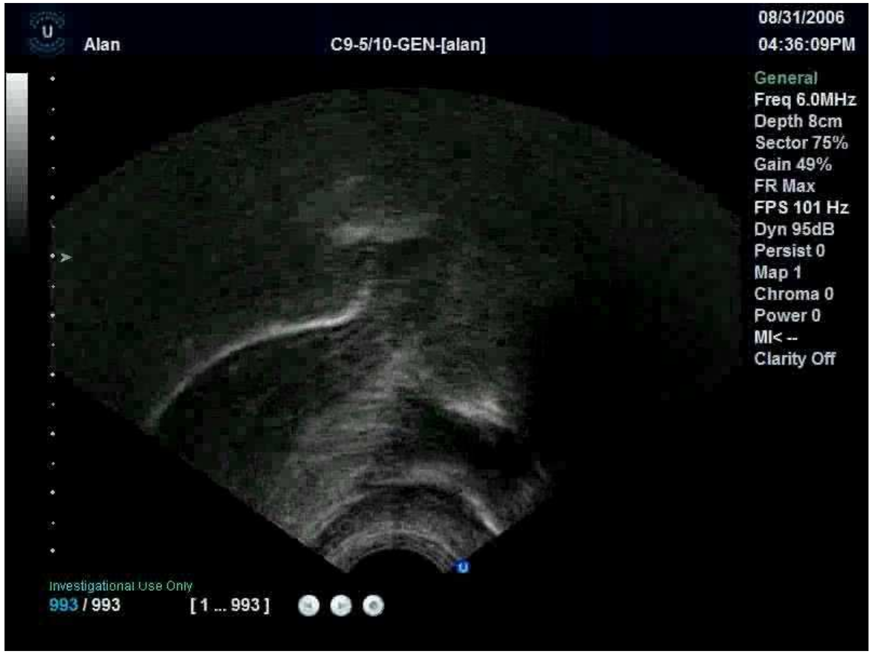
P3: Adapting $\mathbf{f}_1$ to speaker 2 given partial contours containing only the landmark positions



Data collection

- Ultrasound data of tongue movement



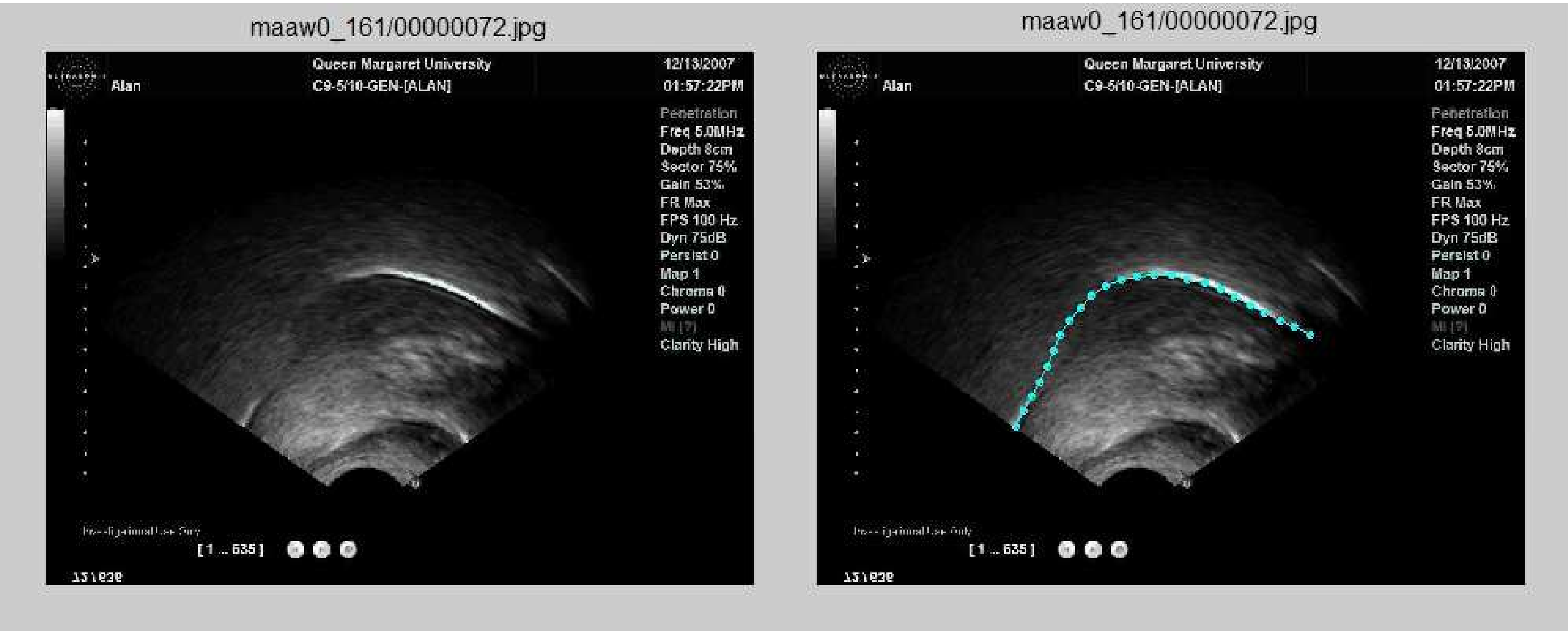


Data collection

- Ultrasound machine and head stabilization device (QMU, Edinburgh)

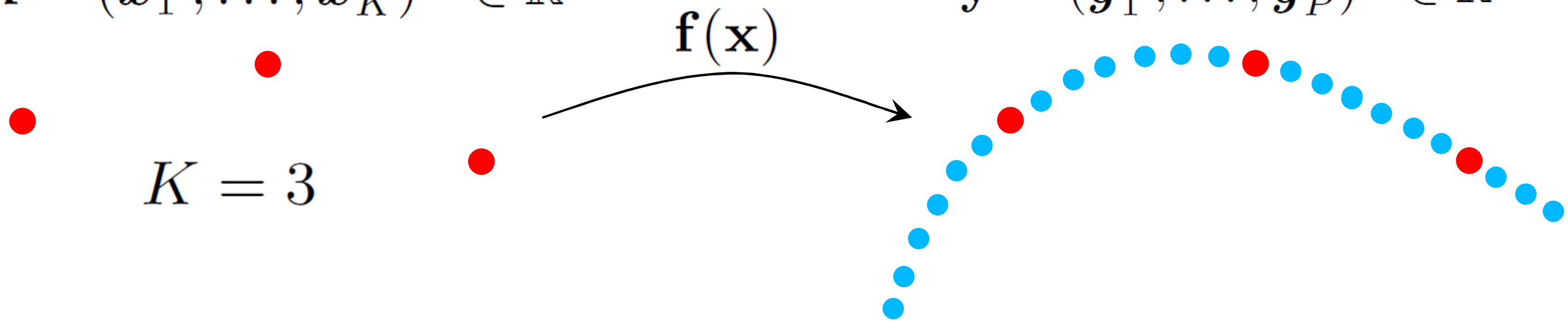
Data collection

- Tongue contour tracking
  - A difficult task due to noisy ultrasound images
  - Tongue parts are invisible from time to time
  - Solution: automatic track by EdgeTrak (Li et al' 05) + manual correction
- Tongue contour dataset
  - 22 read TIMIT sentences from a native Scottish English speaker
  - tongue contours and audios recorded in 2 sessions



P1: learn a predictive model of tongue shapes for a given speaker

$\mathbf{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_K^T)^T \in \mathbb{R}^{2K}$ $\mathbf{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_P^T)^T \in \mathbb{R}^{2P}$



- Assume midsaggital contours, but extendable to 3D tongue surfaces
- Given a training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ of tongue contours (ground truth)
- Predict a test contour $\mathbf{y}$ from the location $\mathbf{x}$ of K pellets (Qin et al'08)

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\Phi(\mathbf{x}) + \mathbf{w} \;, \; \phi_m(\mathbf{x}) = \exp\left(-\tfrac{1}{2} \left\|(\mathbf{x} - \boldsymbol{\mu}_m)/\sigma\right\|^2\right).$$

- Estimate the mapping $\mathbf{f}$ from the training set by the least-square

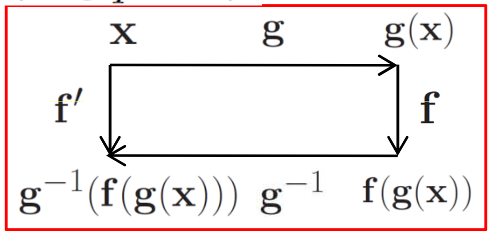
$$E(\mathbf{f}) = \sum_{n=1}^N \left\|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\right\|^2$$

P2: adapt the predictive model given full contours

- Model adaptation is very hard
- Adaptation based on feature normalization (Qin&Carreira-Perpiñán '09)
  - Key aspect: apply the same transformation to each 2D point of an $\mathbf{x}-$ or $\mathbf{y}-$ contour

$$\tilde{\mathbf{x}} = \mathbf{g}_{\mathbf{x}}(\mathbf{x}) = \begin{pmatrix} \mathbf{A}\boldsymbol{x}_1 + \mathbf{b} \\ \cdots \\ \mathbf{A}\boldsymbol{x}_K + \mathbf{b} \end{pmatrix} \qquad \tilde{\mathbf{y}} = \mathbf{g}_{\mathbf{y}}(\mathbf{y}) = \begin{pmatrix} \mathbf{A}\boldsymbol{y}_1 + \mathbf{b} \\ \cdots \\ \mathbf{A}\boldsymbol{y}_P + \mathbf{b} \end{pmatrix}$$

  - The adapted predictive mapping is given by $\mathbf{g}_{\mathbf{y}}^{-1} \circ \mathbf{f} \circ \mathbf{g}_{\mathbf{x}}$



  - Advantage of 2D-wise alignment mapping
    - Easily invertible and 6 parameters to estimate $\mathbf{A}_{2\times 2}$ and $\mathbf{b}_{2\times 1}$
    - Requires very little adaptation data with no need of correspondence
  - To estimate $\mathbf{g}$, we minimize the error function

$$\min_{\mathbf{A},\mathbf{b}} F(\mathbf{A}, \mathbf{b}) = \sum_{n=1}^{N} \|\mathbf{g}_{\mathbf{y}}(\mathbf{y}_n) - \mathbf{f}(\mathbf{g}_{\mathbf{x}}(\mathbf{x}_n))\|^2$$

  - Using only10~20 adaptation contours and 1 sec cputime, reconstruction errors are comparable to those with retraining with abundant dataset.

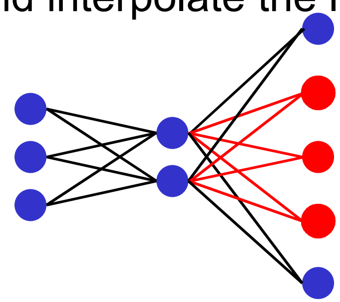
P3: adapt the predictive model given partial contours

- Given the partial, K-landmark contours from MOCHA/XRMB as adaptation data and no full contours, how to reconstruct the full tongue shape?
- Solution (Qin&Carreira-Perpiñán '10) → "this paper"
  - Consider the pellets coordinates as input $\mathbf{x}$ and also as output $\mathbf{y} = \mathbf{x}$
  - Minimize the new error function

$$\min_{\mathbf{A},\mathbf{b}} F_{\mathbf{x}}(\mathbf{A},\mathbf{b}) = \sum_{n=1}^{N} \left\| \mathbf{g}_{\mathbf{x}}(\mathbf{x}_n) - \mathbf{f}_{\mathbf{x}}(\mathbf{g}_{\mathbf{x}}(\mathbf{x}_n)) \right\|^2$$

Equivalent to seek $\{\mathbf{A},\mathbf{b}\}$ such that the adapted model $\mathbf{g}_{\mathbf{x}}^{-1} \circ \mathbf{f}_{\mathbf{x}} \circ \mathbf{g}_{\mathbf{x}}$ best approximate the identity mapping and interpolate the landmarks

$$\boldsymbol{f} \rightarrow \boldsymbol{f_x}$$



  - Apply $\{\mathbf{A},\mathbf{b}\}$ to reconstruct the entire contour as $\mathbf{g}_{\mathbf{y}}^{-1} \circ \mathbf{f} \circ \mathbf{g}_{\mathbf{x}}$

P3: solution

- Problems:
  - Tongue compresses and stretches from time to time
  - Our training contours show mostly equidistant contour points
  - Small % of frames show distances between pellets differ by 30%
  - Including unusual frames results in bad results
- Solution:
  - Discarding unusual frames wastes useful data
  - Instead, regularize $F_{\mathbf{x}}(\mathbf{A},\mathbf{b})$ to encourage $\mathbf{A}$ to have a low condition number

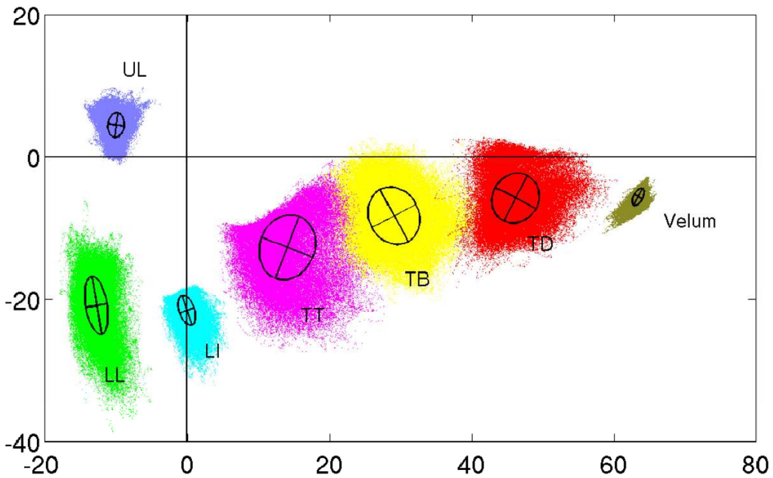
$$\min_{\mathbf{A},\mathbf{b}} F_{\mathbf{x}}(\mathbf{A},\mathbf{b}) + \lambda C(\mathbf{A}), \quad \lambda \geq 0.$$

    where $C(\mathbf{A}) = \operatorname{tr}\left(\mathbf{A}^T\mathbf{A}\right) - D\det\left(\mathbf{A}^T\mathbf{A}\right)^{1/D}$ for $\mathbf{A}_{D\times D}$
  - We choose C(A) since it is easier to minimize than $\operatorname{cond}(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$
  - We use BFGS to find the optimal A, b; converges in ~10 iterations, each costs O(N.M.K)
- Advantage of regularization: make the algorithm robust to landmarks misspecification

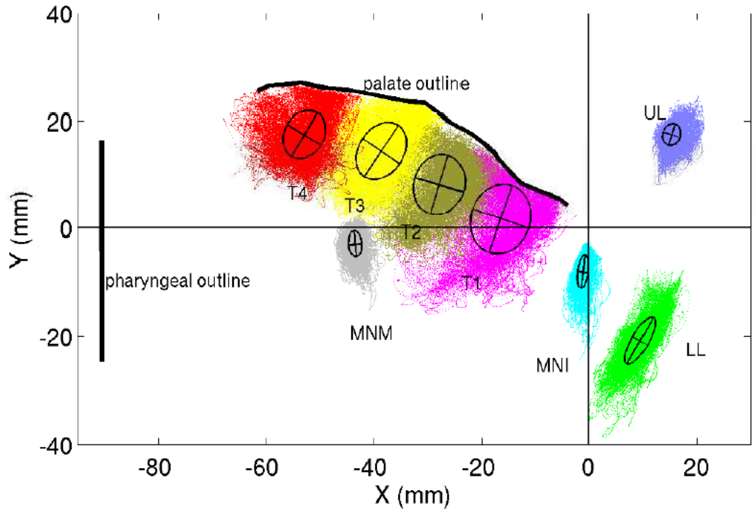
P3: solution continued

- Determine the landmark location by hand

MOCHA ([4 9 14])



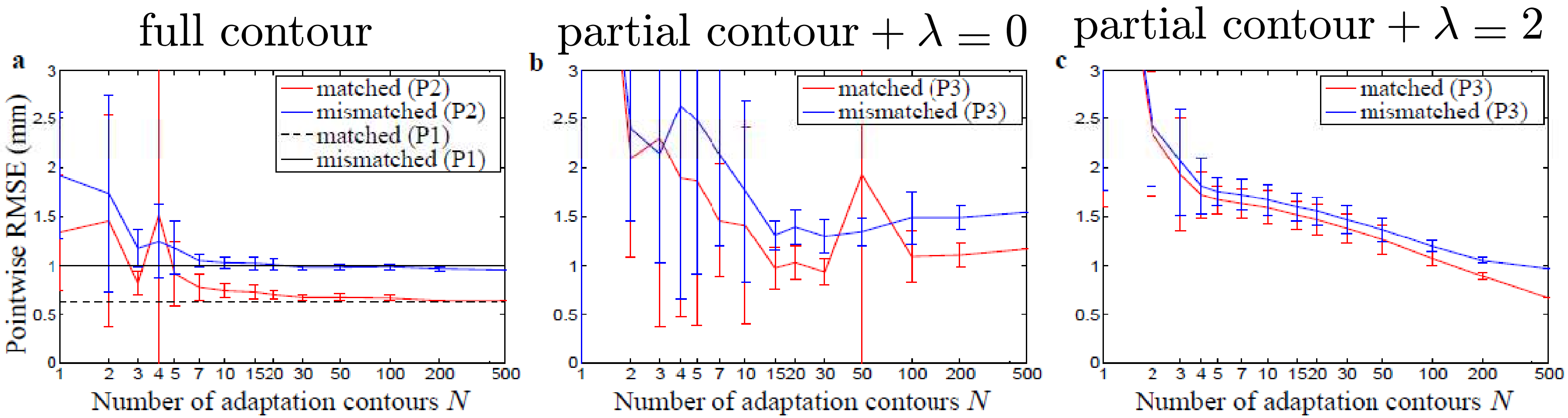
XRMB ([3 7 11 15])



- Computation complexity
  - $\mathcal{O}(NMK)$ per BFGS iteration, converges around 10 iterations
- Predictive model
  - RBF: $M = 500$ basis functions, width $\sigma = 55$ mm and regularization parameter 10^{-4} trained by cross-validation from dataset

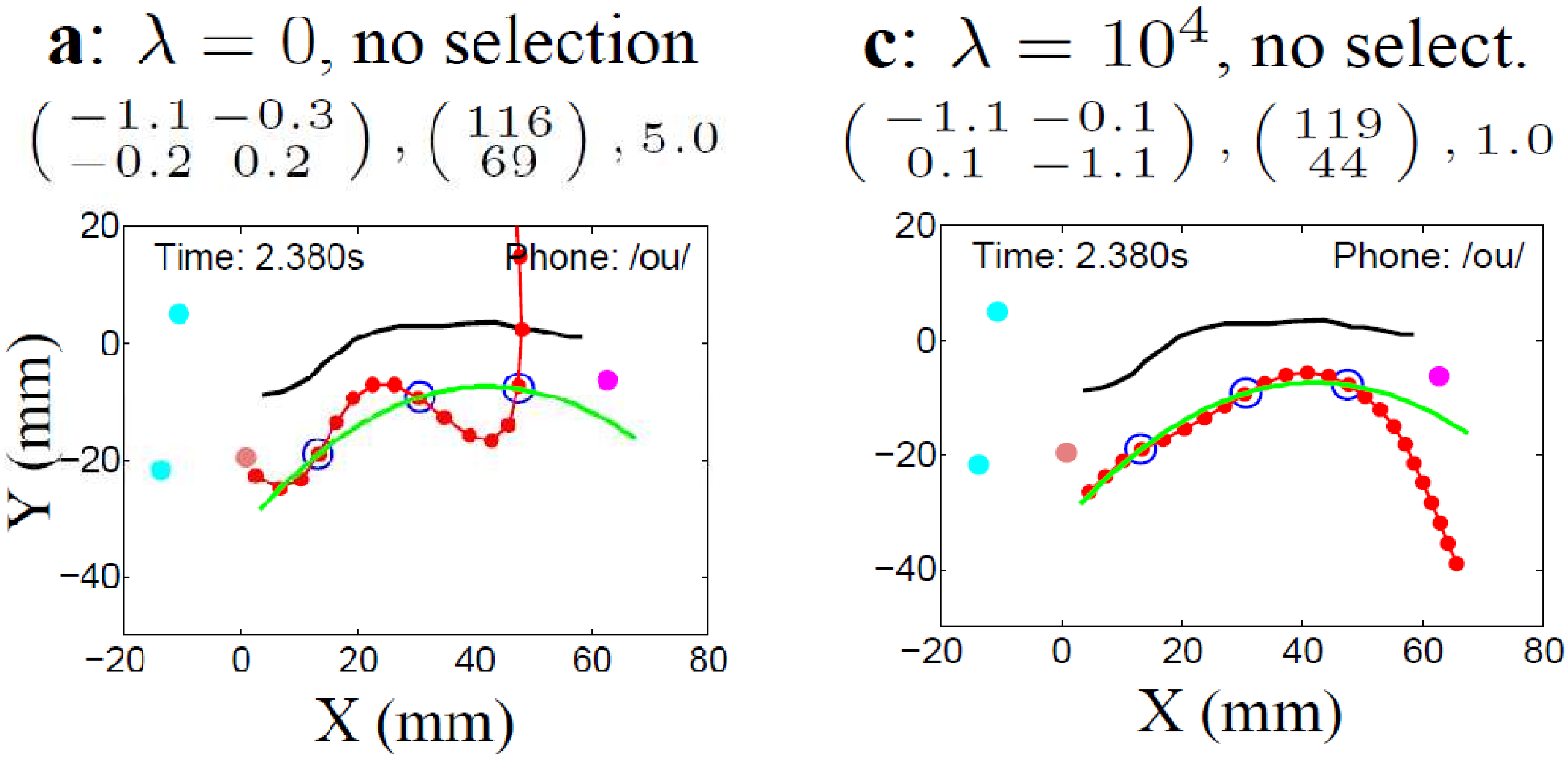
P3: reconstruction error with known ground truth

- Setup for experiment 1:
  - Use the tongue database
  - 991 contours for testing and up to 500 contours for use in adaptation
  - All contours transformed by $\mathbf{A} = \begin{pmatrix} 1.1 & -0.05 \\ -0.1 & 1.2 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 10 \\ -10 \end{pmatrix}$
  - Two choices of landmarks' placement:
    - Matched: [4 9 14] as in training
    - Mismatched: [4.2 9.2 14.2]

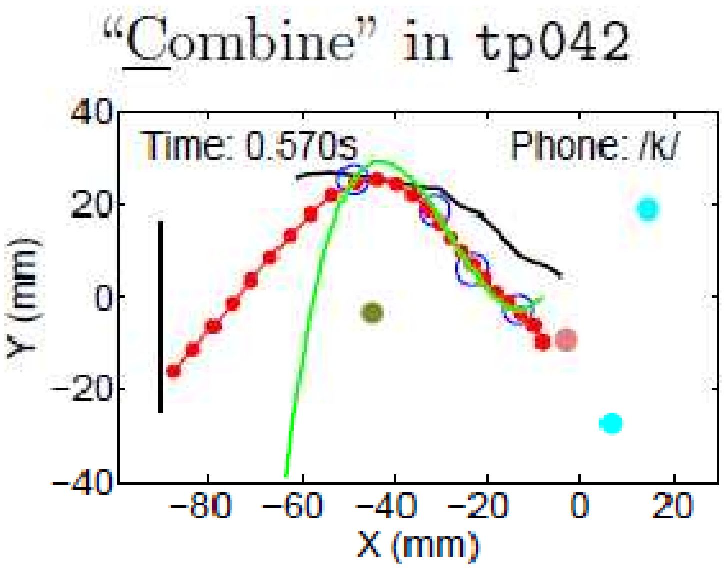


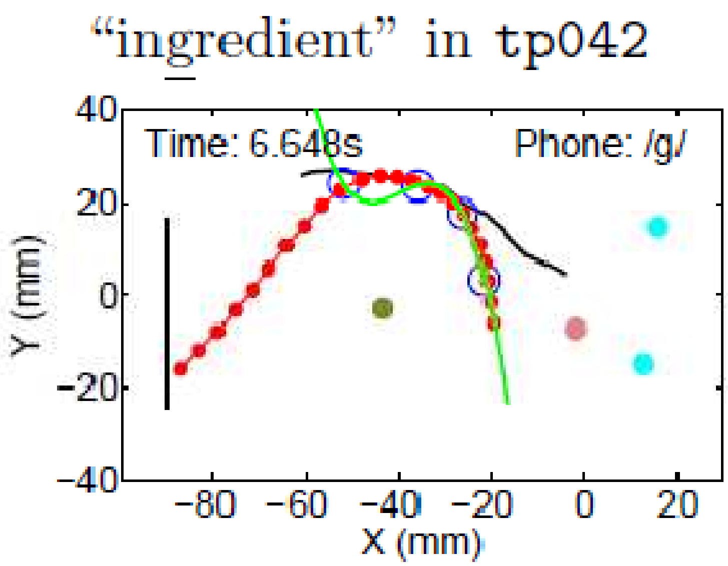
P3: effects of regularization on condition number

- Setup for experiment 2
  - Reconstruct full tongue contours for MOCHA/XRMB databases
  - Use $N = 3\,600$ partial contours from MOCHA for adaptation

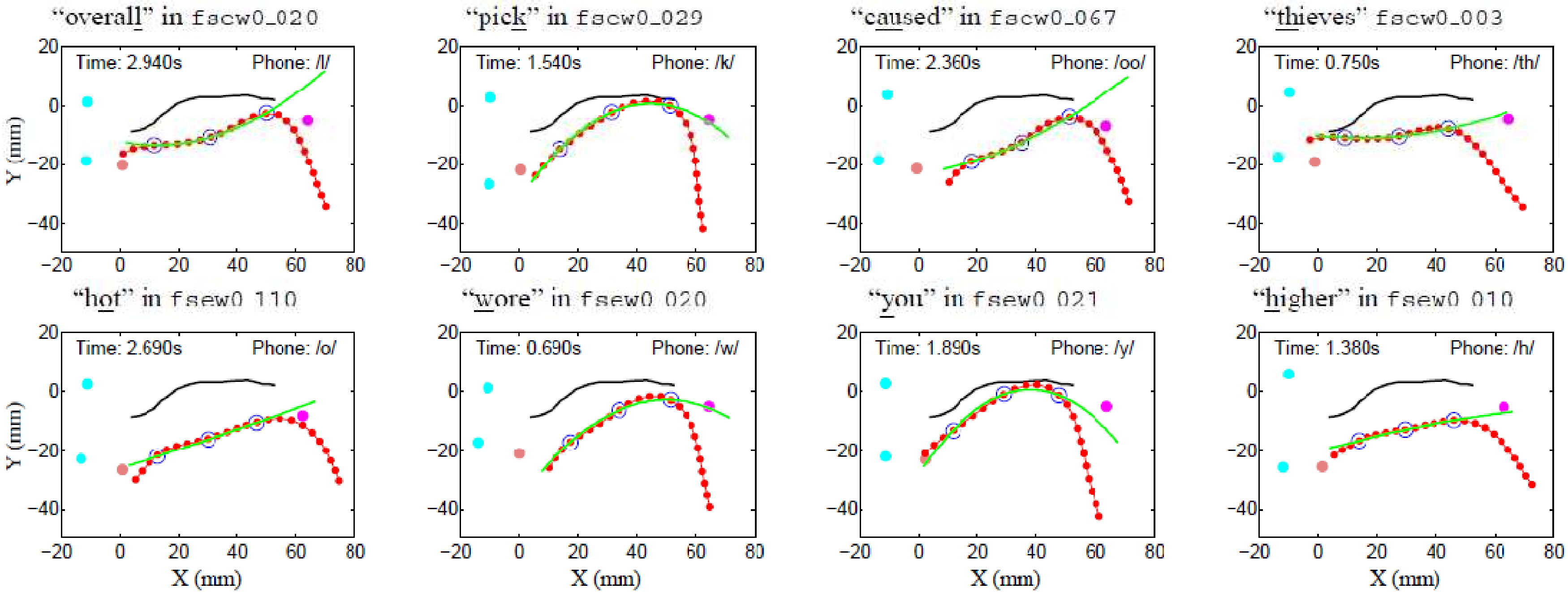


P3: reconstruction vs. spline interpolation

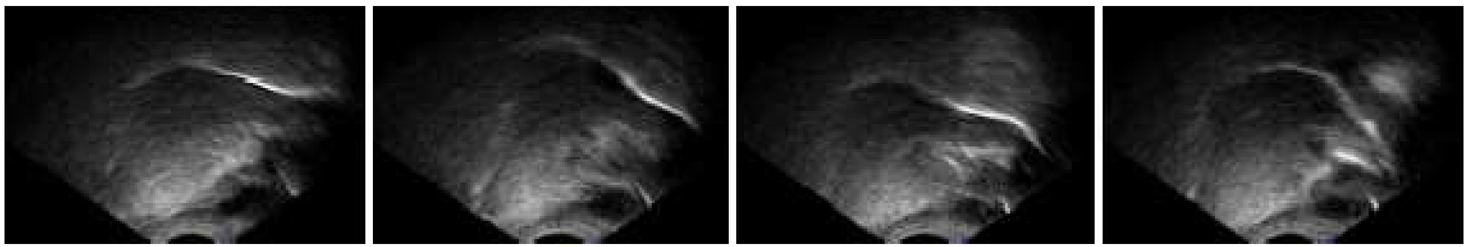




P3: realistic tongue reconstruction (MOCHA)

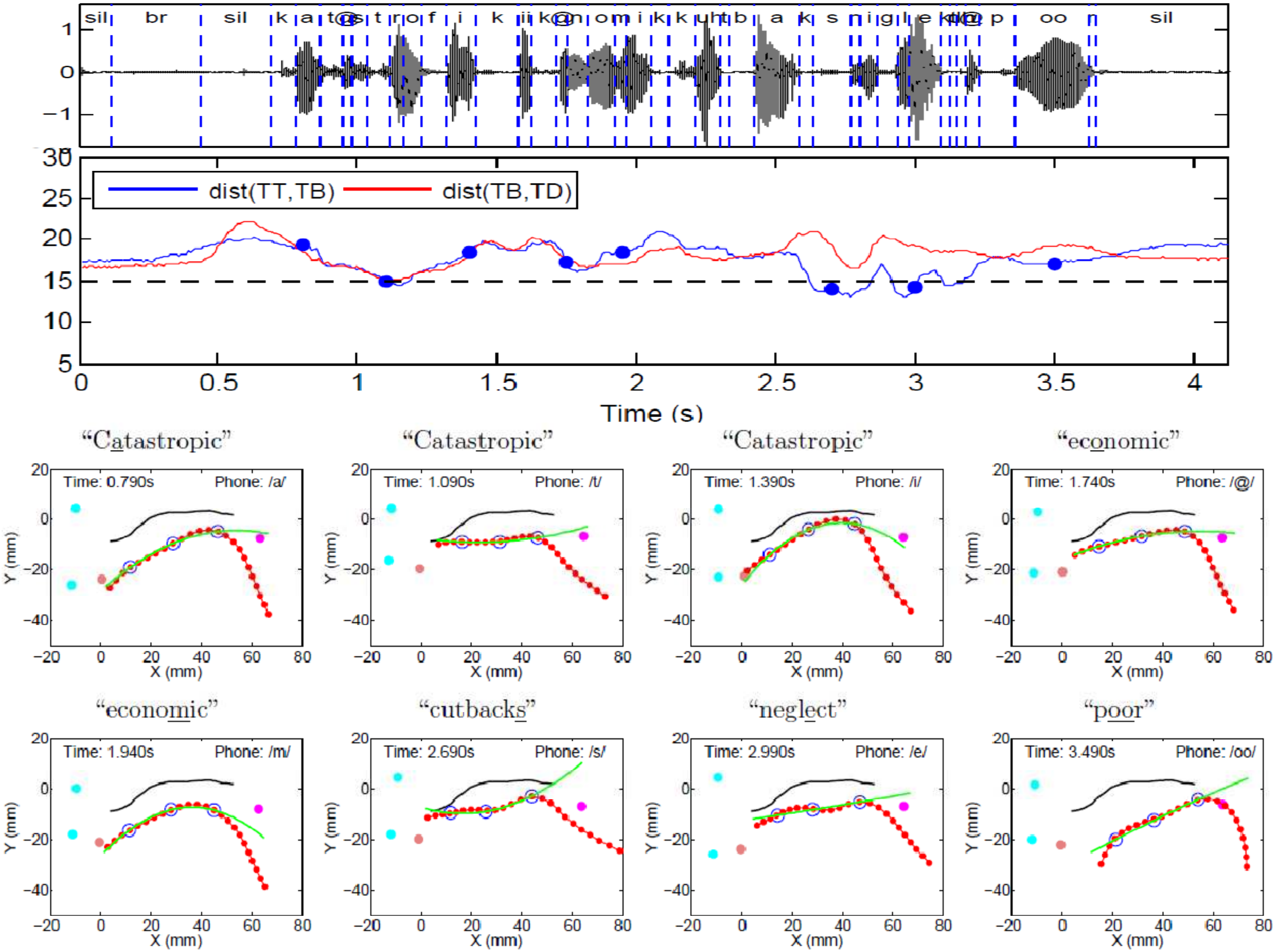


frame 300	frame 177	frame 382	frame 300
maaw0_054	maaw0_177	maaw0_177	maaw0_069

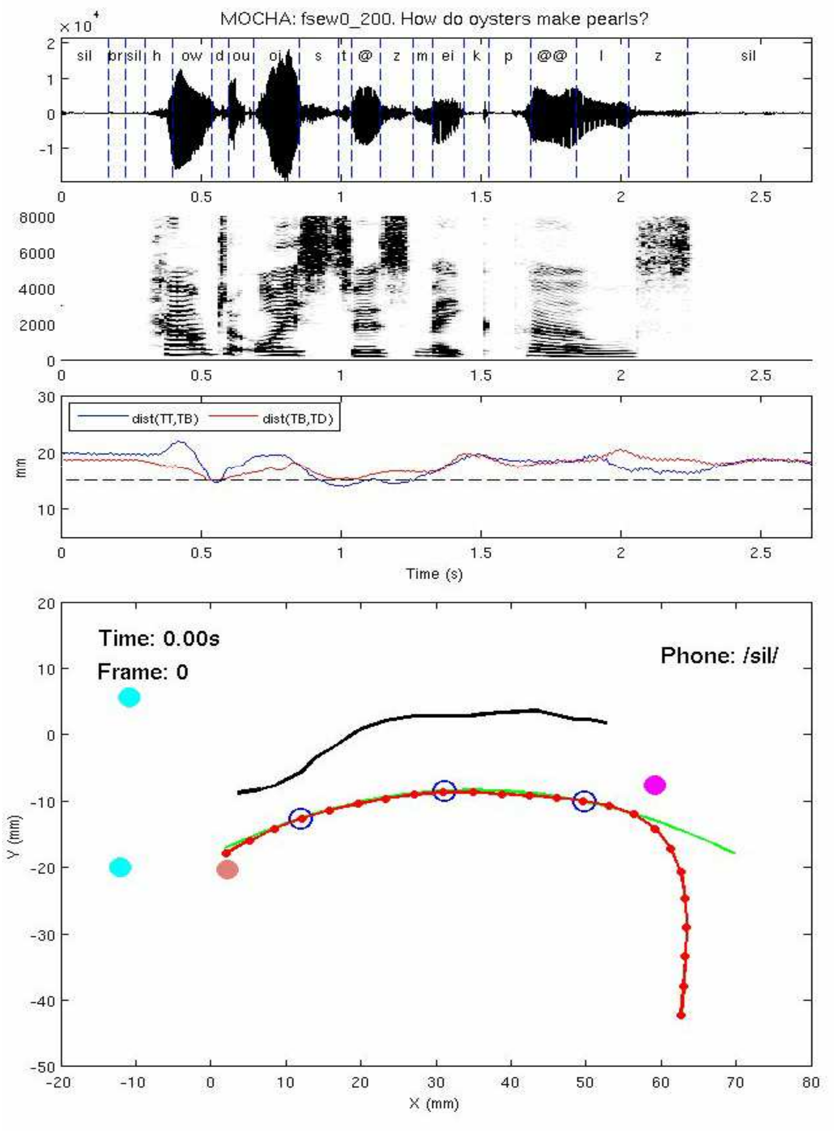


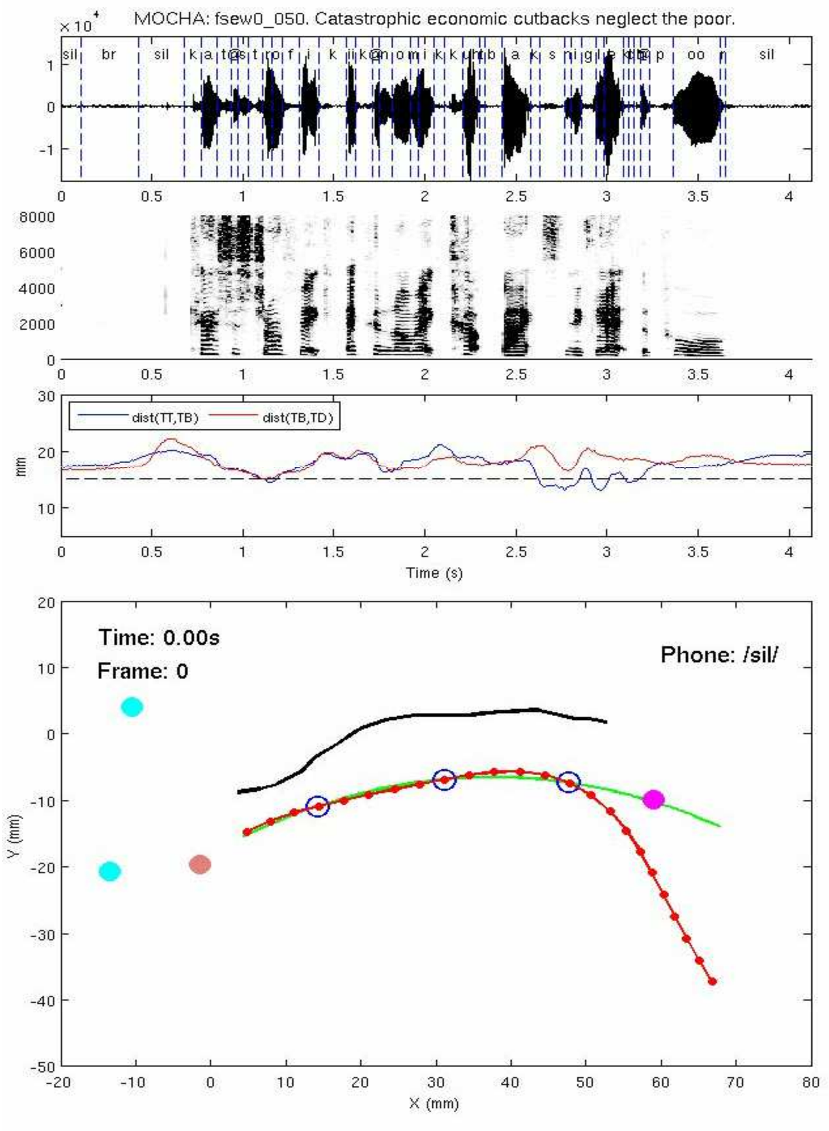
P3: reconstruction w.r.t pellets misspecification

MOCHA: fsew0_050. Catastrophic economic cutbacks neglect the poor

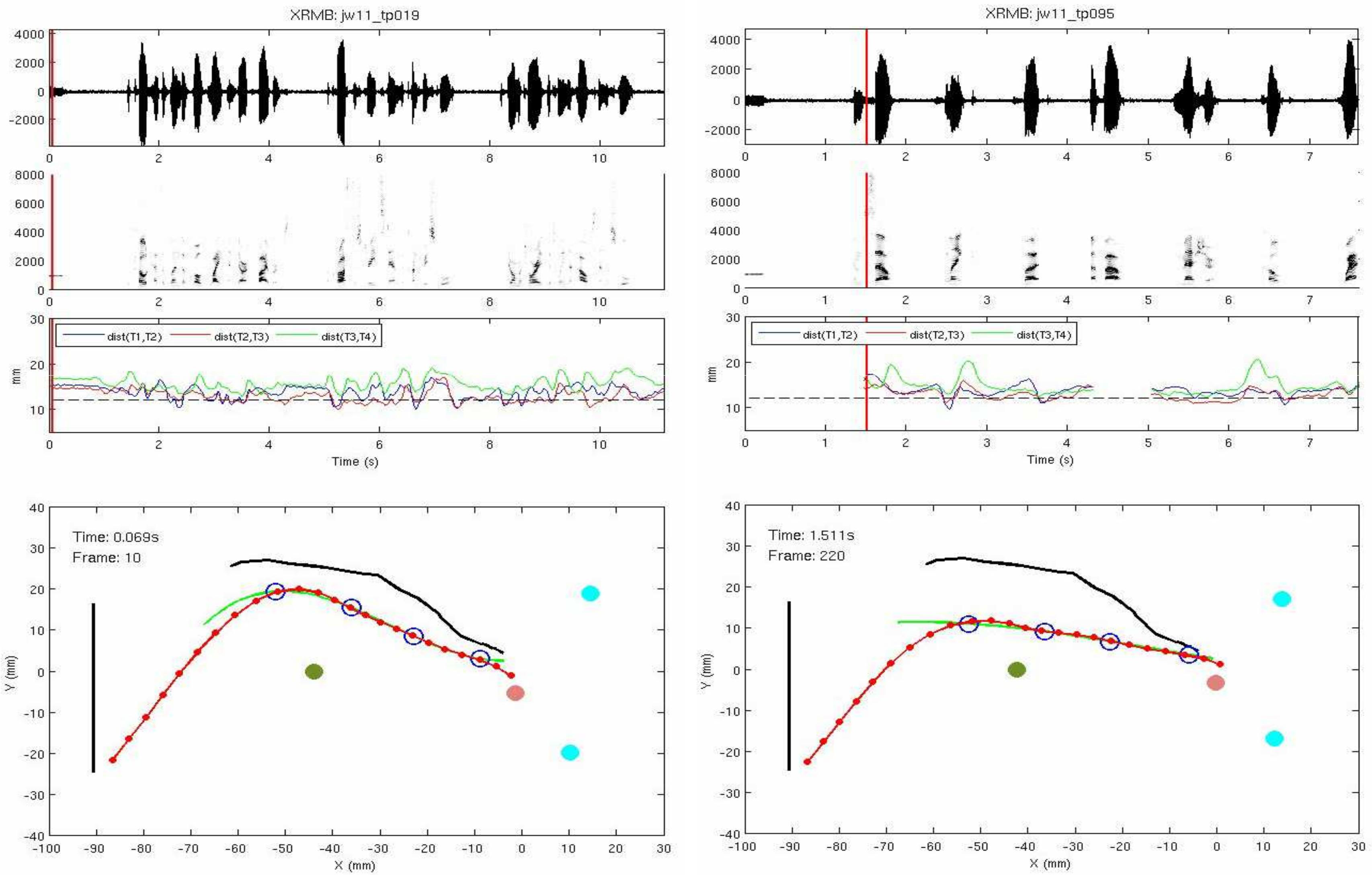


P3: reconstruction of tongue shapes for MOCHA





P3: reconstruction of tongue shapes for XRMB



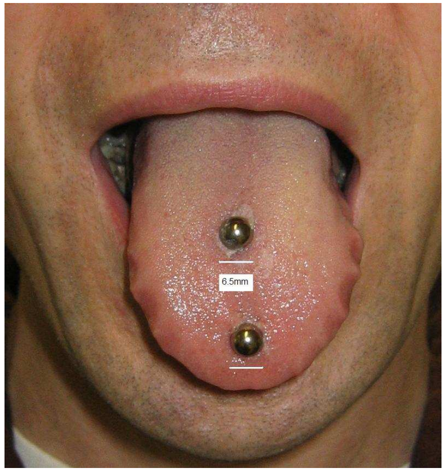
Conclusions

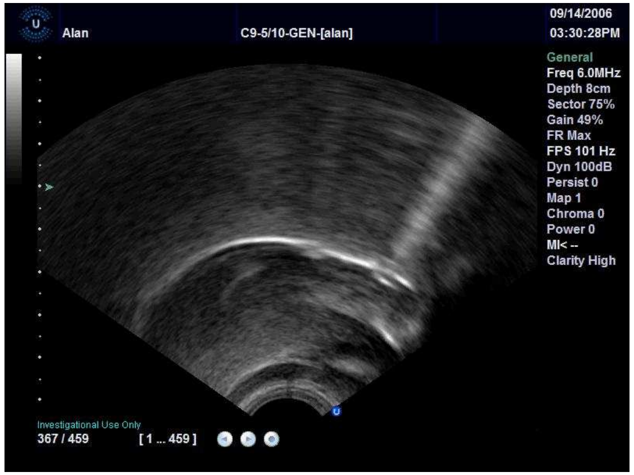
- An algorithm that can recover realistic tongue shapes given partial contours (containing just 3-4 points) for a never seen speaker.
- We applied it to two public datasets, MOCHA and XRMB.
- The reconstructed tongue satisfies physical constraints without having to specify the latter in the model.
- It provides information not easily inferred from the MOCHA/XRMB data, e.g. the location of tongue-palate constrictions.
- Matlab software available from the authors.

Acknowledgement

- Alan Wrench from Queens Margaret University
- Korin Richmond and Steve Renals from CSTR, Edinburgh
- Work funded by NSF awards IIS-0754089 (CAREER) and IIS-0711186
- XRMB funded (in part) by NIDCD grant R01 DC 00820

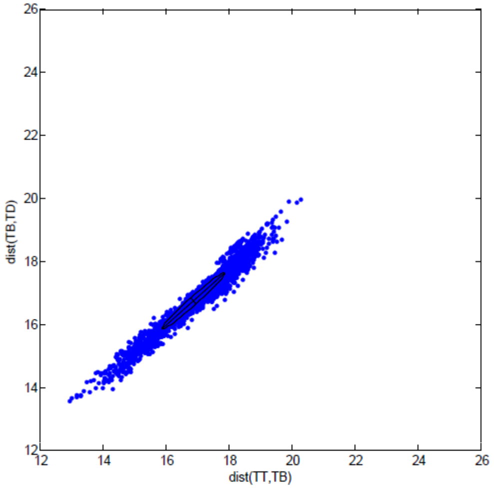
P3: tongue stretching problem



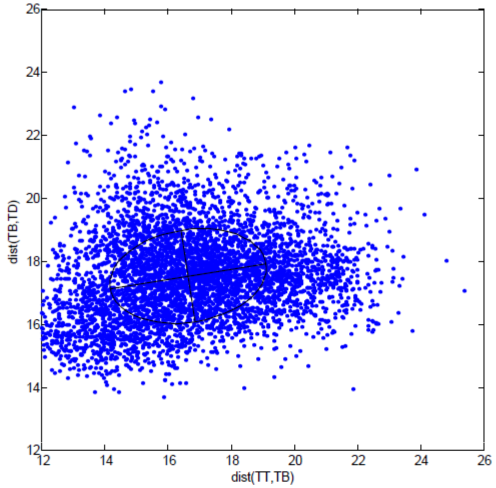


P3: scatterplot of inter-pellet distances

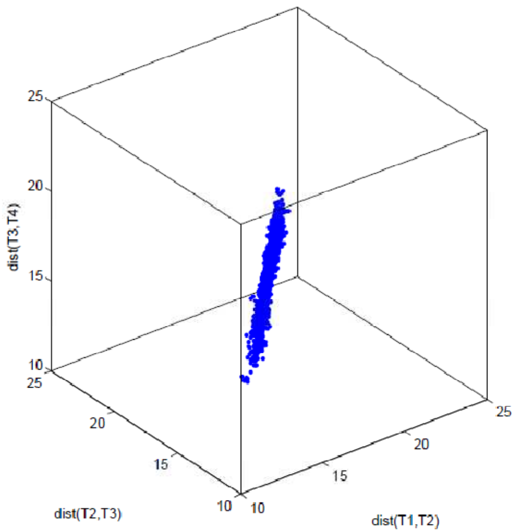
Ultrasound



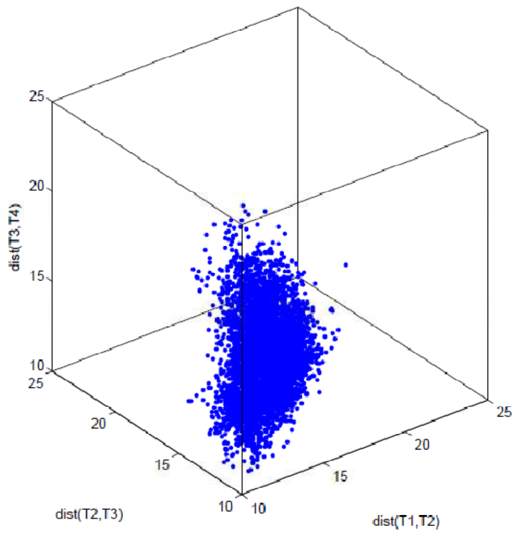
MOCHA

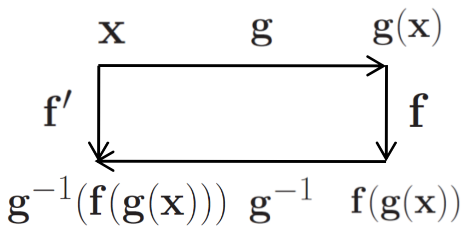


Ultrasound



XRMB




x
g
g(x)
f′
f
g⁻¹(f(g(x)))
g⁻¹
f(g(x))