Learning and adaptation of a tongue shape model with missing data

Mohsen Farhadloo and **Miguel Á. Carreira-Perpiñán**

Electrical Engineering and Computer Science

University of California, Merced

`http://eecs.ucmerced.edu`

Introduction

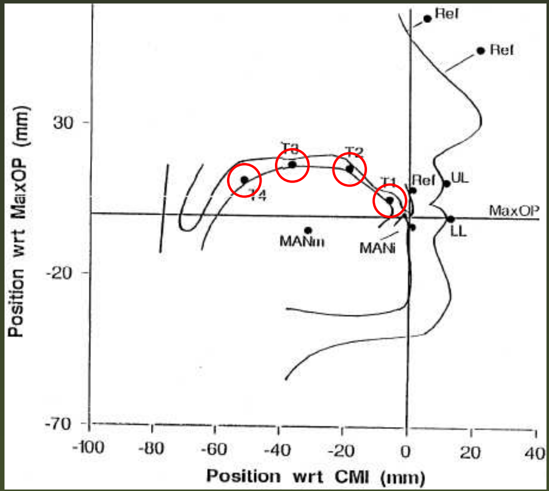
- We consider a realistic, data-driven model of the tongue shape, in particular its midsagittal contour.
- Applications: talking heads, articulatory synthesis and inversion, tracking in ultrasound and MRI, and reconstructing the tongue contour in articulatory databases such as MOCHA.
- Landmark-based models use as control parameters the location on the tongue contour of a fixed number of fleshpoints (landmarks), given which the entire tongue shape is reconstructed.

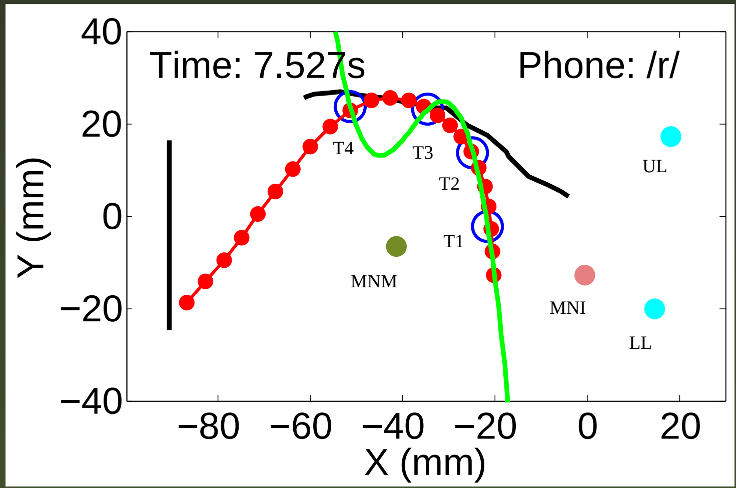
  This is a low-dimensional model of the tongue contour.
- We consider two fundamental problems:
  - Training the model for a speaker, given a large dataset of contours.
  - Adapting the model to a new speaker, given a few contours.
- In this paper, we solve both problems when there is missing data.

Reconstructing EMA/X-ray microbeam (Qin et al., ’10)

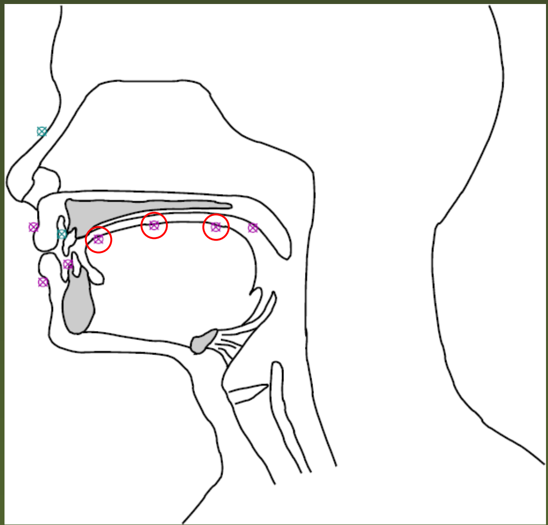
Our original motivation: reconstruct the tongue contour in EMA/X-ray microbeam articulatory databases by adapting a tongue model constructed for a reference speaker.

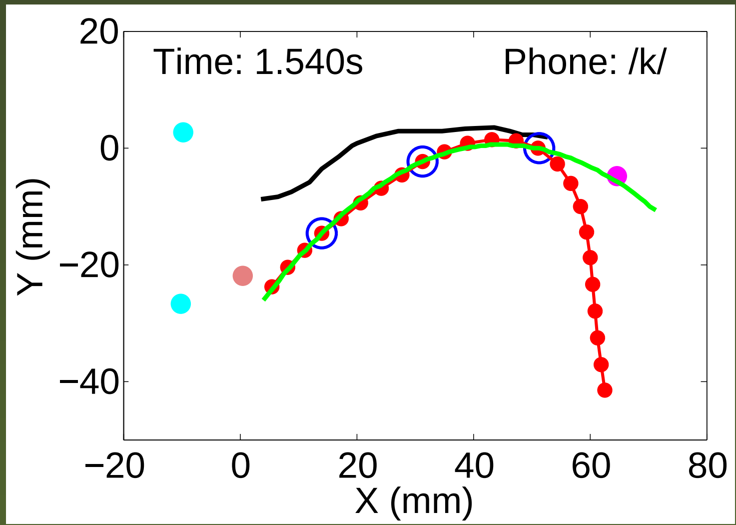
XRMB





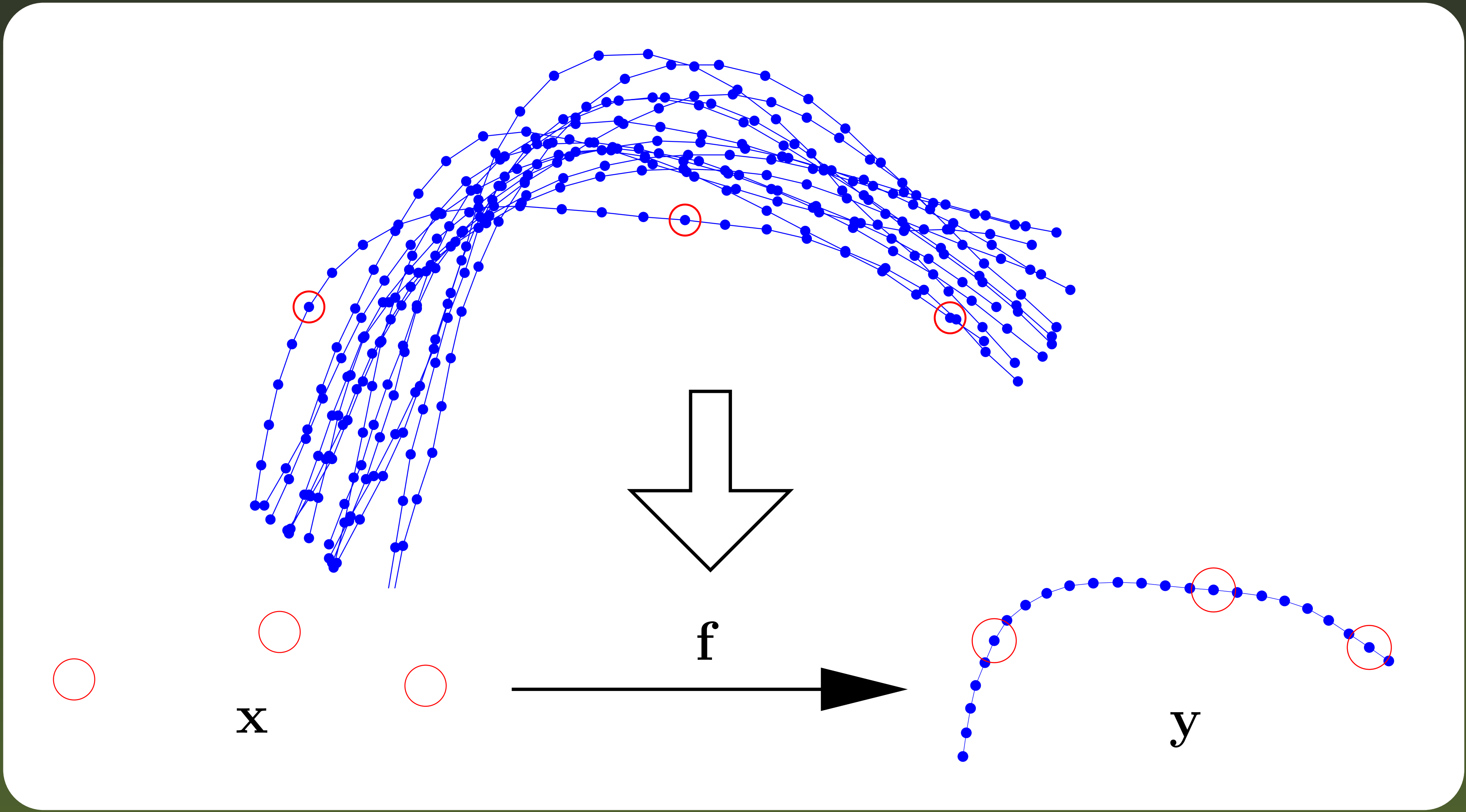
MOCHA





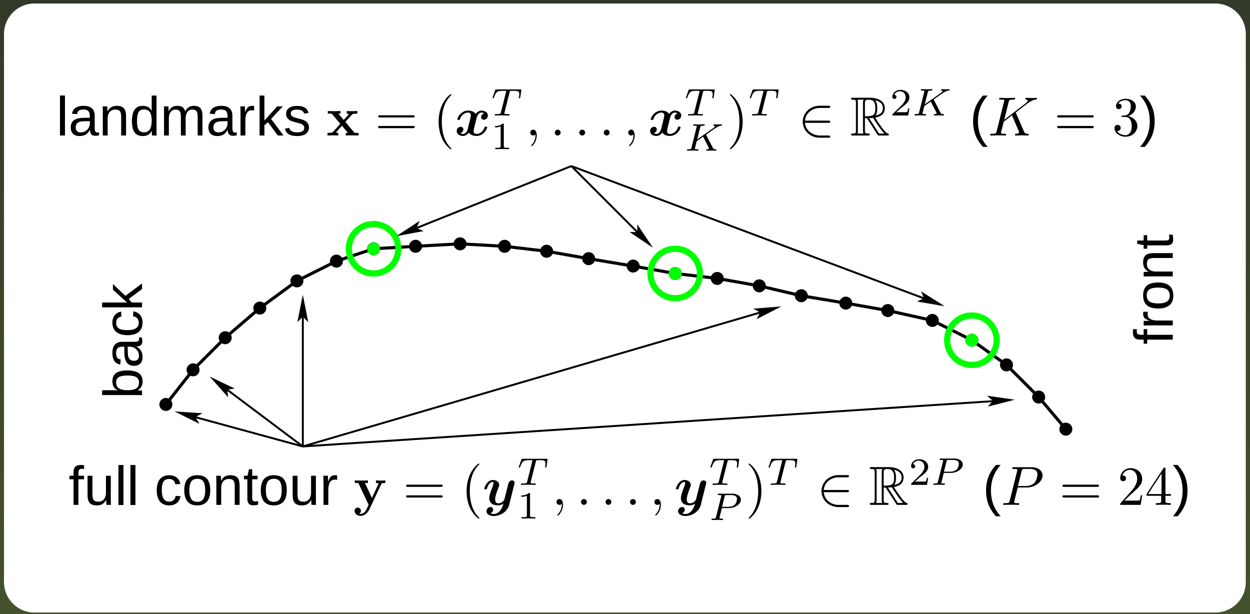
Predictive model of the tongue contour (Kaburagi et al., '94; Qin et al., '08)

The training problem: learn a predictive model f of the full tongue contour for a given speaker given many full contours from it.



Predictive model of the tongue contour (cont.)

Given the 2D locations of K landmarks located on the tongue contour ($\mathbf{x}$), reconstruct the entire contour ($\mathbf{y}$), represented by P 2D points.



We can obtain full contours from ultrasound recordings (semiautomatic segmentation process).

Predictive model of the tongue contour (cont.)

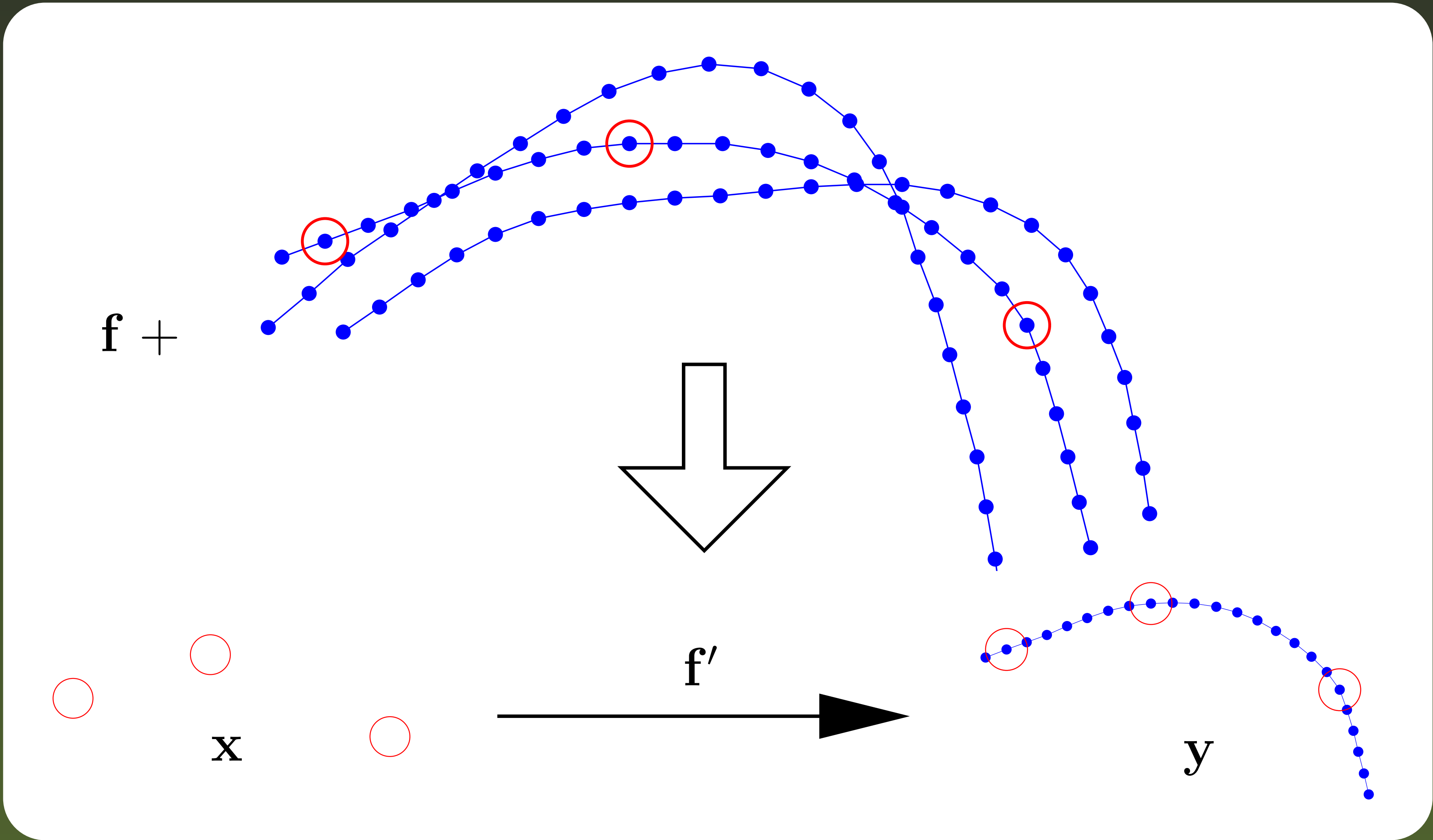
- Learn a predictive mapping $\mathbf{f}$ in order to reconstruct the full tongue contour given only a few landmarks, from a dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ containing many contours.
  Radial basis function (RBF) network: $\mathbf{f}(\mathbf{x}) = \mathbf{W}\Phi(\mathbf{x}) + \mathbf{w}$
  M Gaussian basis functions $\phi_m(\mathbf{x}) = \exp\left(-\frac{1}{2}\left\|(\mathbf{x} - \boldsymbol{\mu}_m)/\sigma\right\|^2\right)$, width σ.
- Minimize the following objective function given N contours:

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

- This achieves submillimetric error per contour point (below the ultrasound measurement accuracy).
- It beats spline interpolation of the landmarks.

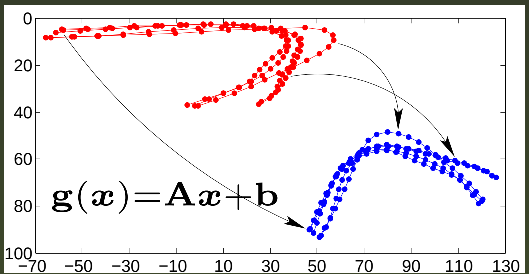
Adapting a predictive model to a new speaker (Qin et al. ’09, ’10

The adaptation problem: adapt the predictive model $\mathbf{f}$ to a new, target speaker given a few full contours from the latter.



Adapting a predictive model to a new speaker (cont.)

- We transform contours between speaker spaces with invertible linear mappings $\mathbf{g}_{\mathbf{x}}(\mathbf{x})$ and $\mathbf{g}_{\mathbf{y}}(\mathbf{y})$ constructed with 2D-wise mappings $\mathbf{g}$:



$$
\begin{tikzcd}
\mathbf{x} \arrow[rr, "\mathbf{g}_{\mathbf{x}}"] \arrow[d, "\mathbf{f}'"'] & & \mathbf{g}_{\mathbf{x}}(\mathbf{x}) \arrow[d, "\mathbf{f}"] \\
\mathbf{g}_{\mathbf{y}}^{-1}(\mathbf{f}(\mathbf{g}_{\mathbf{x}}(\mathbf{x}))) & & \mathbf{f}(\mathbf{g}_{\mathbf{x}}(\mathbf{x})) \arrow[ll, "\mathbf{g}_{\mathbf{y}}^{-1}"]
\end{tikzcd}
$$

$$\mathbf{g}_{\mathbf{x}}(\mathbf{x}) = \begin{pmatrix} \mathbf{A}_1^{\mathbf{x}}\boldsymbol{x}_1 + \mathbf{b}_1^{\mathbf{x}} \\ \dots \\ \mathbf{A}_K^{\mathbf{x}}\boldsymbol{x}_K + \mathbf{b}_K^{\mathbf{x}} \end{pmatrix}$$

$$\boldsymbol{x}_1, \dots, \boldsymbol{x}_K \in \mathbb{R}^2$$

and same for $\mathbf{g}_{\mathbf{y}}^{-1}(\mathbf{y})$.

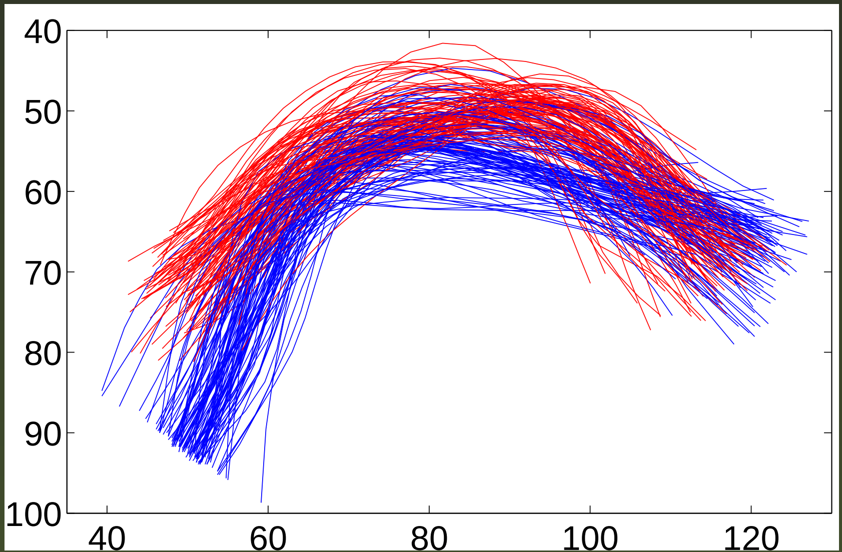
- Total $6(K+P)$ parameters ($\ll$ # parameters of $\mathbf{f}$).
- We minimize the following objective function given N adaptation contours using BFGS (details in paper):

$$E(\mathbf{A}^{\mathbf{x}}, \mathbf{b}^{\mathbf{x}}, \mathbf{C}^{\mathbf{y}}, \mathbf{d}^{\mathbf{y}}) = \sum_{n=1}^{N} \left\| \mathbf{y}_n - \mathbf{g}_{\mathbf{y}}^{-1}(\mathbf{f}(\mathbf{g}_{\mathbf{x}}(\mathbf{x}_n))) \right\|^2 .$$

- This requires no correspondences (i.e., the adaptation contours need not match any sound of the reference).

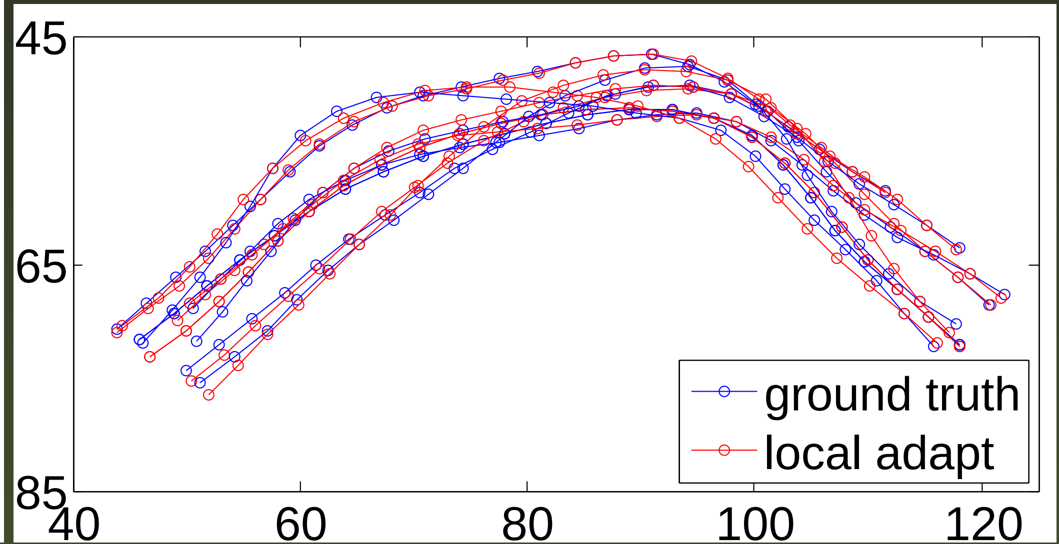
Adapting a predictive model to a new speaker (cont.)

Contours before adaptation



Contours after adaptation

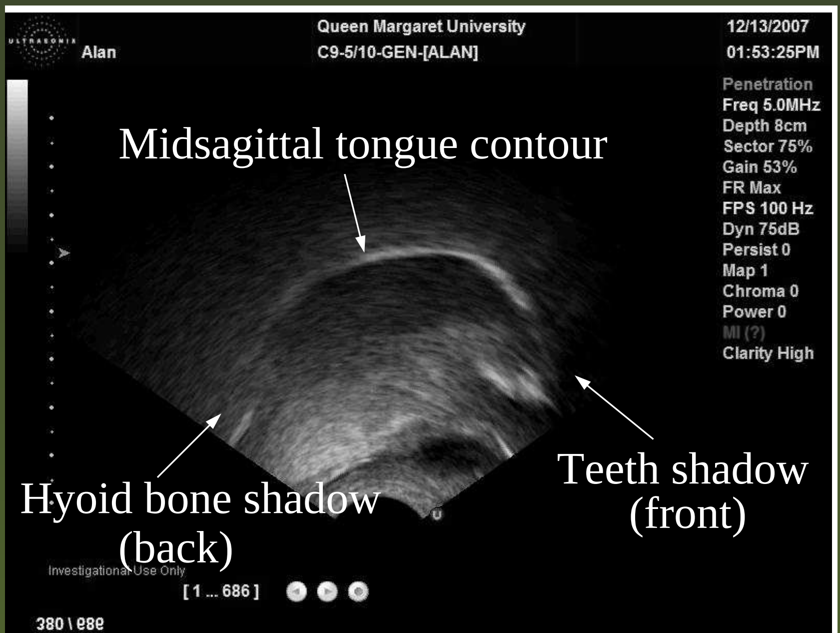
(not all contours shown, to avoid clutter)

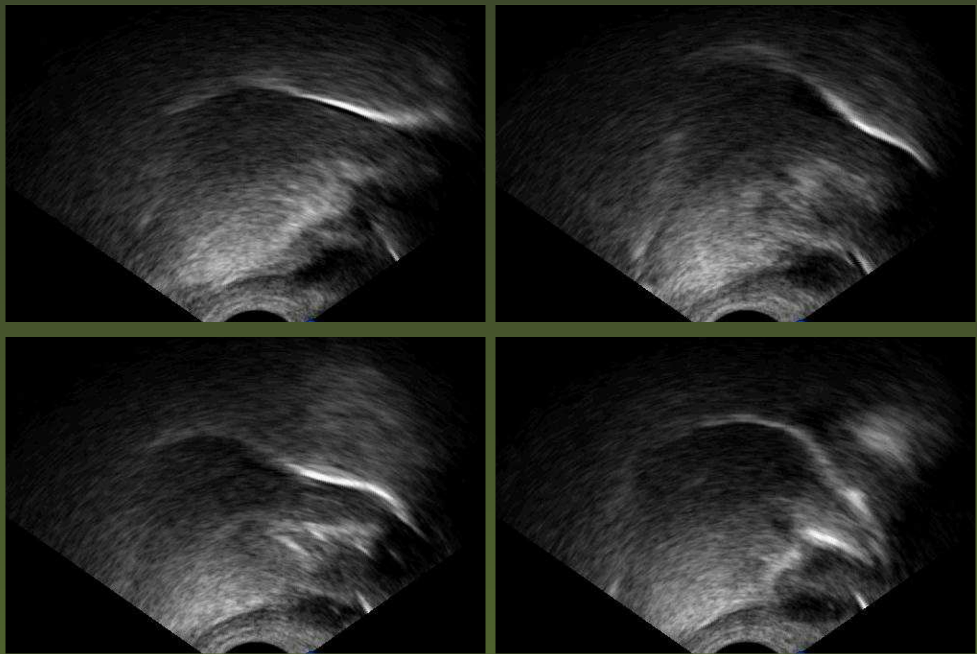


Training and adaptation with missing data

Missing data in ultrasound (incomplete contours) caused by:

- Noise and shadows occlude portions of the contour.
- Back/tip of tongue may exit window of visibility of the probe.
- Tongue surfaces disappear if parallel to the probe.
- Errors in (manual or automatic) segmentation of the tongue contour.





Training and adaptation with missing data (cont.)

We cannot afford to discard incomplete contours:

- Wasteful (recording and segmentation are costly and cumbersome).
- Can severely reduce the number of complete contours available, particularly in the adaptation setting.

The tongue contours have implicit temporal and spatial redundancy.

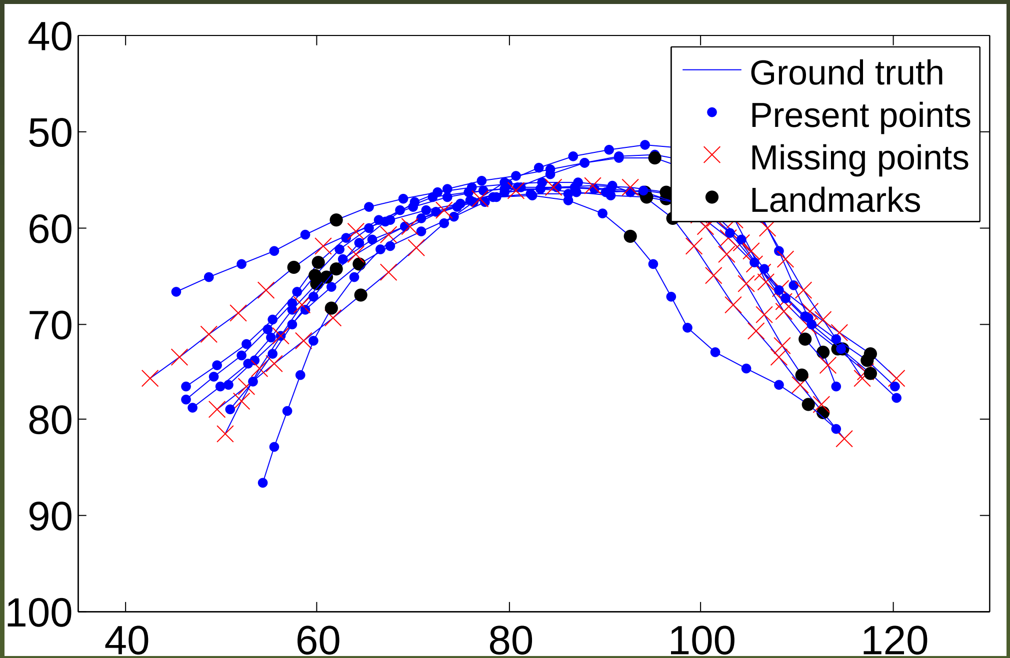
Assume now we are given a dataset of contours $\{(\mathbf{x}_n, \mathbf{y}_n)\}$, each of which may contain missing points.

Training and adaptation with missing data (cont.)

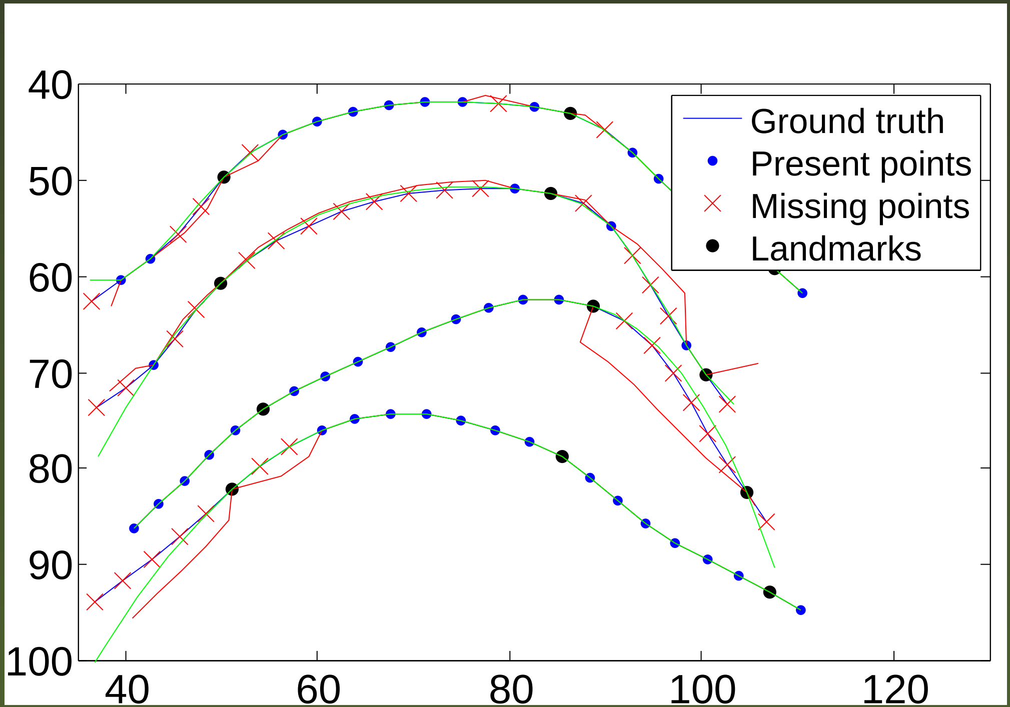
Approach 1: reconstruct the missing data, the train/adapt as usual.

- Mean imputation.
- Spline imputation.

Sample contours with missing runs



Mean and spline imputation with missing data at random/in runs



Training and adaptation with missing data (cont.)

Approach 2: directly train/adapt without reconstructing any missing data ("missing data deleted" technique: drop missing terms from objective):

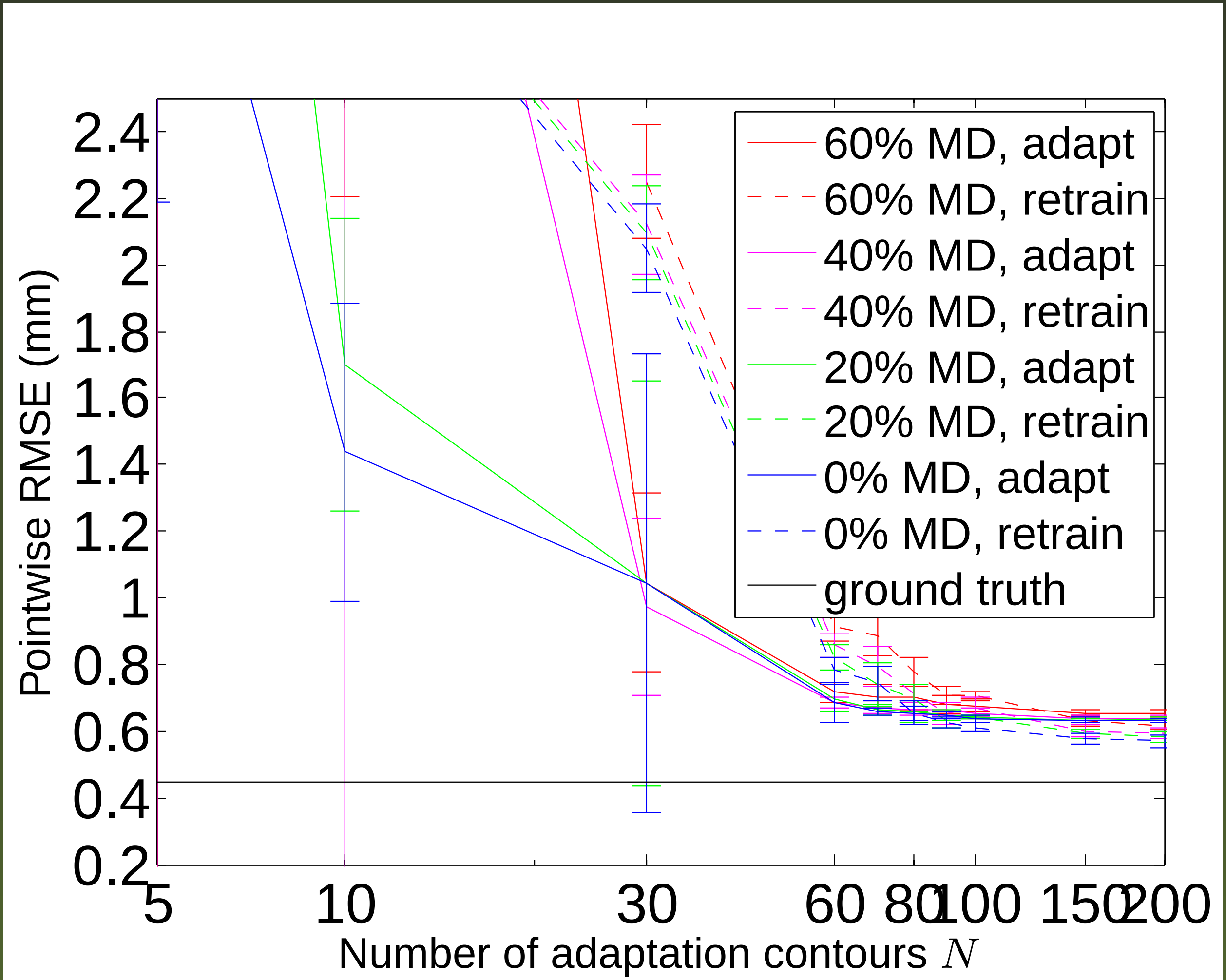
- Training: $E(\mathbf{f}) =$
  - With complete data: $\sum_{n=1}^{N} \sum_{j=1}^{2P} (y_{jn} - (\mathbf{f}(\mathbf{x}_n))_j)^2$.
  - With missing data: $\sum_{\text{present}\, n,j} (y_{jn} - (\mathbf{f}(\mathbf{x}_n))_j)^2$.
- Adaptation: $E(\mathbf{A}^{\mathbf{x}}, \mathbf{b}^{\mathbf{x}}, \mathbf{C}^{\mathbf{y}}, \mathbf{d}^{\mathbf{y}}) =$
  - With complete data: $\sum_{\text{present}\, n,j} (y_{jn} - (\mathbf{g}_{\mathbf{y}}^{-1}(\mathbf{f}(\mathbf{g}_{\mathbf{x}}(\mathbf{x}))))_j)^2$.
  - With missing data: $\sum_{\text{present}\, n,j} (y_{jn} - \mathbf{g}_{\mathbf{y}}^{-1}(\mathbf{f}(\mathbf{g}_{\mathbf{x}}(\mathbf{x})))_j)^2$.
- Computational cost:
  - Both missing-data objective functions have $(1-\rho)PN$ terms where $\rho \in [0,1]$ is the proportion of missing data $\Rightarrow$ faster.
  - Imputation methods first reconstruct all contours (so $\rho = 0$) and then minimize $\Rightarrow$ slower, and also larger error.

Experimental results: setup

- Ultrasound database: two speakers (one male, one female) with different Scottish accents ($\approx 10\,000$ contours). We used the male speaker to obtain a reference model, which we adapted to data from the female speaker having missing values. We take $K = 3$ landmarks and $P = 24$ contour points.
- Missing patterns:
  - Missing at random (from 0% to 60% MD).
    representative of random ultrasound noise
  - Missing run (8 consec. points) at front/mid/back (= 33% MD).
    representative of shadowing and other effects
- Comparison methods:
  - Optimal baseline: many contours, no missing data.
  - Retraining a new model from scratch (disregarding the reference model $\mathbf{f}$).
  - Mean imputation and spline imputation.
  - Direct training/adaptation without reconstructing missing data.

Experimental results (cont.)

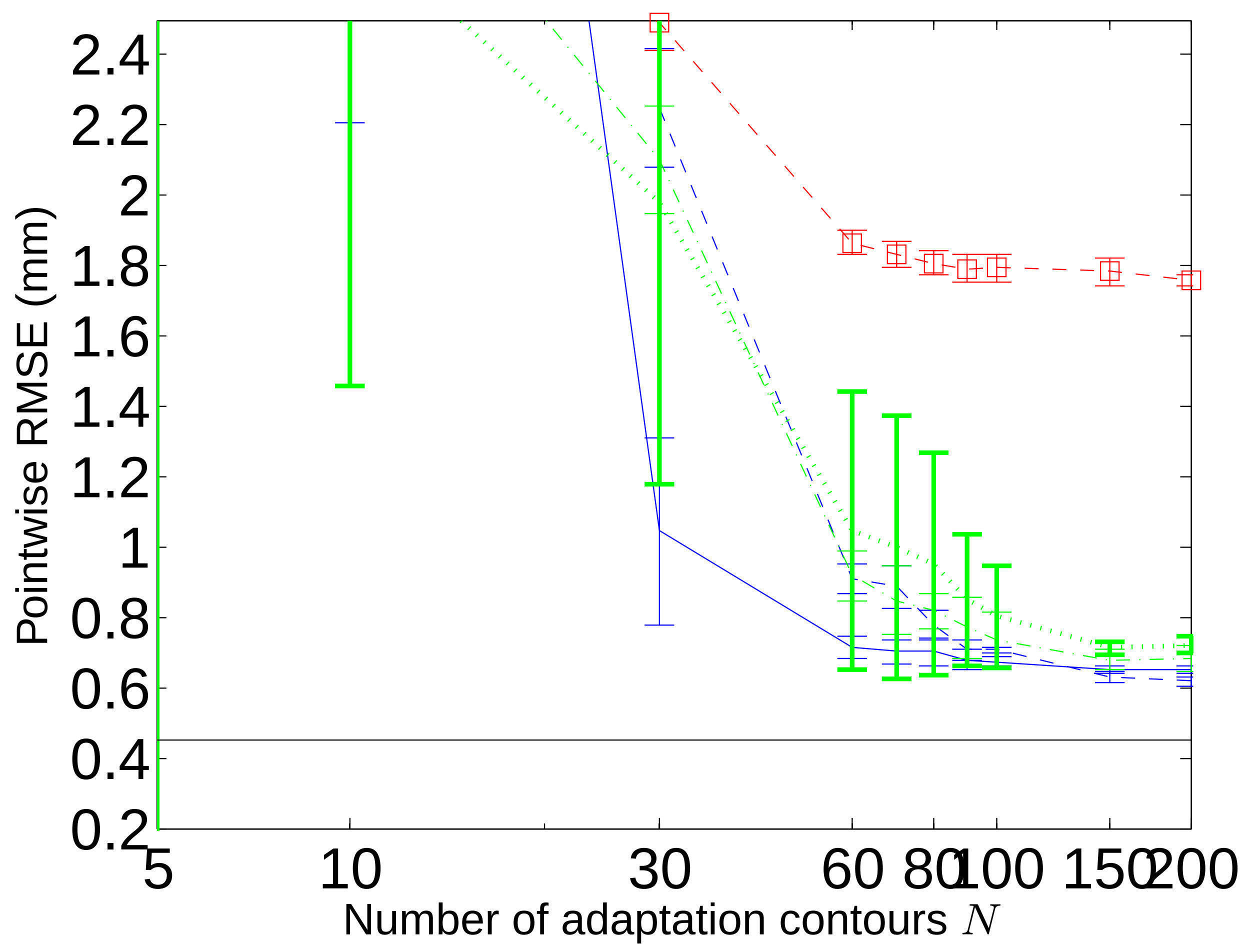
Missing at random pattern, predictive error E for adaptation and retraining with different amounts of missing data, as a function of the number of adaptation contours N.



- Adaptation beats retraining for $N < 100$ contours.
- With as few as $N = 30$ contours and up to 60% missing data, we achieve an error within 0.5 mm from the optimal baseline.
- With very few contours ($N < 30$), up to 20% missing data is still tolerated.

Experimental results (cont.)

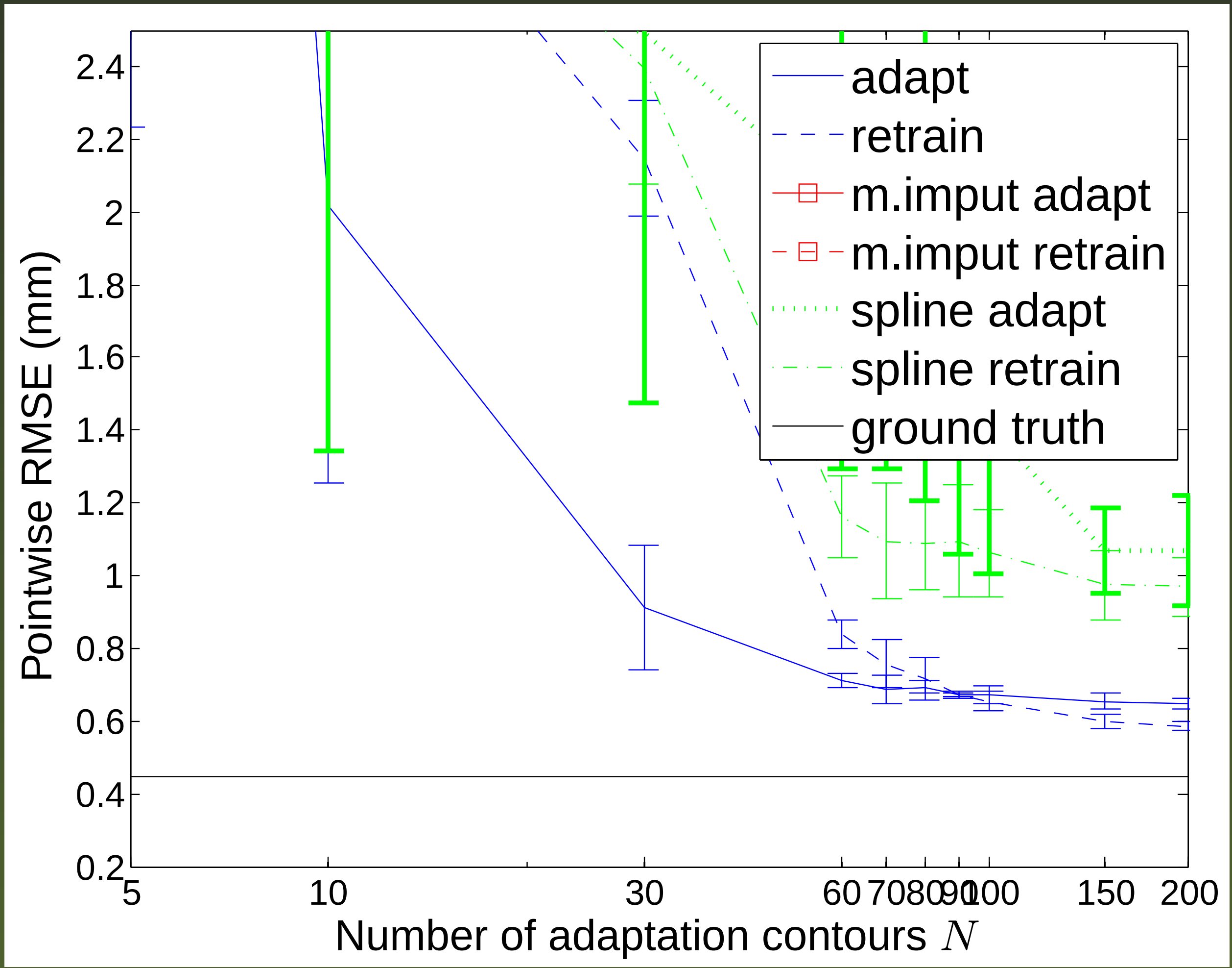
Missing at random pattern (60% missing data), comparing direct retraining/adaptation (blue lines) with retraining/adaptation after mean imputation (red) and spline imputation (green).



- Direct retraining/adaptation does best.
- Spline imputation does somewhat worse.
- Mean imputation does poorly.

Experimental results (cont.)

Missing runs pattern (33% missing data), comparing direct retraining/adaptation (blue lines) with retraining/adaptation after mean imputation (red) and spline imputation (green).



- Direct retraining/ adaptation does best, with an error similar to the missing-at-random case.
- Spline imputation does quite worse.
- Mean imputation is off the chart.

Conclusions

- We have extended a landmark-based training and adaptation model of the tongue shape to deal with missing data.
- With significant amounts of missing data, we achieve an accuracy comparable to that using complete data, and with less computation time.
- No need to reconstruct the missing data.
- Limitation: the landmarks themselves cannot be missing.
- Could use to increase the temporal resolution of ultrasound by skipping scan lines (reducing the spatial resolution), thus trading off missing data in the temporal and spatial domains.

Work funded by NSF award IIS–0711186.