ADAPTATION OF A PREDICTIVE MODEL OF TONGUE SHAPES

Chao Qin and Miguel Á. Carreira-Perpiñán

EECS, School of Engineering, University of California, Merced, USA

1 Abstract

It is possible to recover the full midsagittal contour of the tongue with sub-millimetric accuracy from the location of just 3–4 landmarks on it. This involves fitting a predictive mapping from the landmarks to the contour using a training set consisting of contours extracted from ultrasound recordings. However, extracting sufficient contours is a slow and costly process. Here, we consider adapting a predictive mapping obtained for one condition (such as a given recording session, recording modality, speaker or speaking style) to a new condition, given only a few new contours and no correspondences. We propose an extremely fast method based on estimating a 2D-wise linear alignment mapping, and show it recovers very accurate predictive models from about 10 new contours.

2 Motivation and idea

- Extracting sufficient good contours to estimate a predictive model is hard
- Need to obtain models for new conditions (e.g. a new recording session, recording modality, speaker or speaking style) $\Rightarrow$ adaptation
- The adaptation problem: Given the predictive mapping $\mathbf{f}$ (that maps landmarks to contours) and a small adaptation set of N contours $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, want to adapt $\mathbf{f}$ by estimating an invertible mapping $\mathbf{g}$ (with few parameters) that maps new data (from the new speaker) to old data (from the reference speaker)
- Advantages of our adaptation method:
  - No need for any correspondence
  - Little data needed for adaptation
  - Few parameters to estimate $\Rightarrow$ extremely fast
  - Applicable to both linear and nonlinear predictive mappings
  - Extendable to 3D shapes

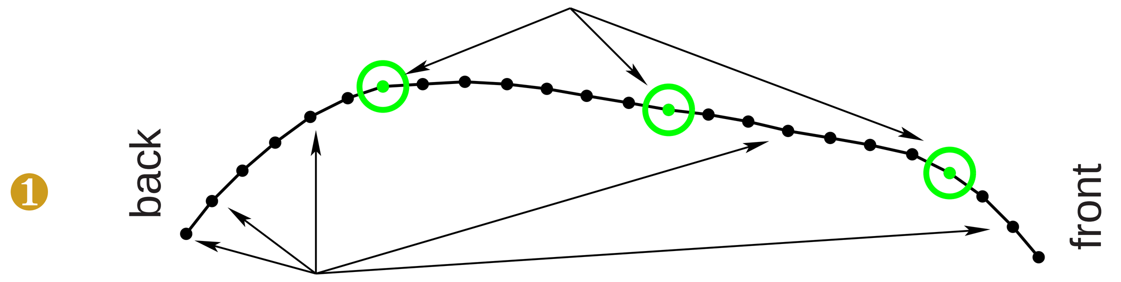
Idea of the method

❶ Learn a predictive mapping $\mathbf{f}$ from a dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ containing N contours

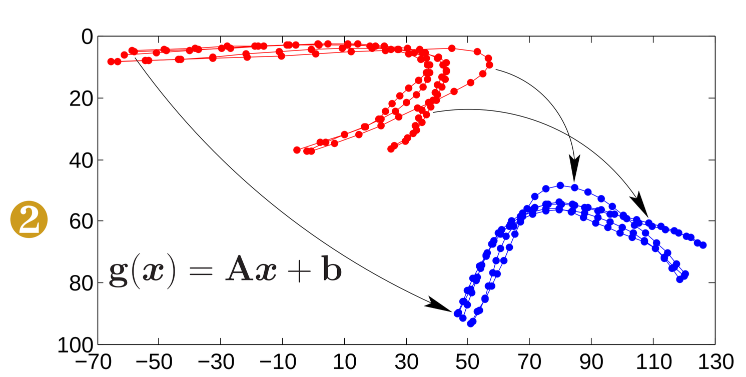
❷ Estimate 2D-wise transformation $\mathbf{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \mathbf{b}$

❸ Obtain the adapted mapping $\mathbf{f}' = \mathbf{g}_\mathbf{y}^{-1} \circ \mathbf{f} \circ \mathbf{g}_\mathbf{x}$

landmarks $\mathbf{x} = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_K^T)^T \in \mathbb{R}^{2K}$ $(K = 3)$

❶ 

full contour $\mathbf{y} = (\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_P^T)^T \in \mathbb{R}^{2P}$ $(P = 24)$

❷ 

❸

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

3 Predictive model of tongue shapes

The prediction problem: given the 2D locations of K landmarks located on the tongue midsagittal contour ($\mathbf{x}$), reconstruct the entire contour ($\mathbf{y}$), represented by P 2D points

- Linear mapping: $\mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{w}$
- Radial basis function (RBF) network: $\mathbf{f}(\mathbf{x}) = \mathbf{W}\Phi(\mathbf{x}) + \mathbf{w}$ with M Gaussian basis functions $\phi_m(\mathbf{x}) = \exp\left(-\frac{1}{2}\|(\mathbf{x} - \boldsymbol{\mu}_m)/\sigma\|^2\right)$ and width σ

4 Adaptation based on feature normalization

- Key aspect: apply the same transformation $\mathbf{g}$ to each 2D point of an $\mathbf{x}$– or $\mathbf{y}$–contour.

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

- The adapted predictive mapping $\mathbf{f}'$ is given by $\mathbf{g}_\mathbf{y}^{-1} \circ \mathbf{f} \circ \mathbf{g}_\mathbf{x}$
- Advantage of 2D-wise alignment mapping $\mathbf{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \mathbf{b}$:
  - Linear $\Rightarrow$ invertible and 6 parameters to estimate, $\mathbf{A}_{2\times 2}$ and $\mathbf{b}_{2\times 1}$.
  - Requires very little adaptation data (one contour is enough if $P > 2$ points)
- To estimate $\mathbf{g}$, we minimize the error function

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

which is easier to minimize than the square error error

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

- Solution:
  - Linear mapping: unique solution from a positive definite 6×6 linear system
  - RBF: we apply the BFGS algorithm and find no local optima. In practice, superlinear convergence in ~ 10 iterations

Computational complexity

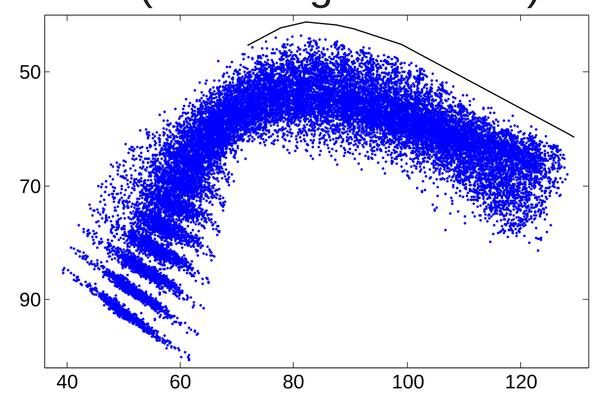
	Linear mapping	RBF
Theory	$\mathcal{O}(32NP)$	$\mathcal{O}(14NM(P+K))$ per BFGS iteration
$N = 10, M = 500, P = 24, K = 3$	4 ms	0.1 s

5 Experimental results

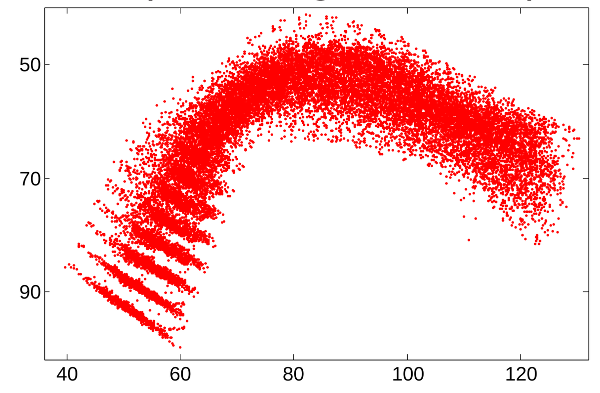
Datasets

- $8\,671$ ultrasound tongue contours ($P = 24$) from a set of 22 British TIMIT sentences for one Scottish speaker, `maaw0`
- Datasets **S1** ($3\,727$ contours from 10 utterances) from session 1 and **S2** ($4\,944$ contours from 12 utterances) from session 2

S1 (recording session 1)



S2 (recording session 2)



Predictive models

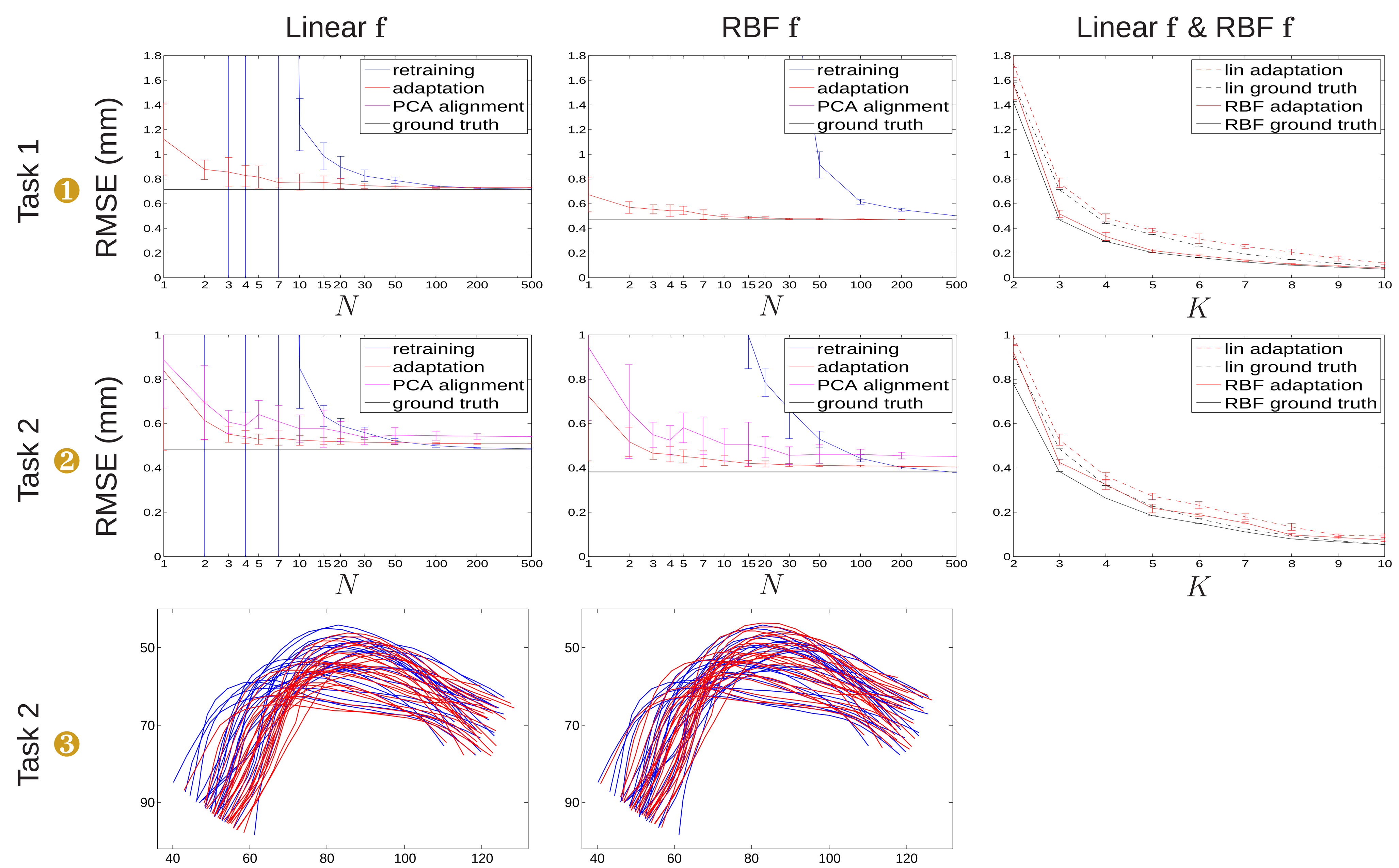
- K landmarks were chosen optimally from the P contour points
- Linear mapping: as a baseline and provide initial $\{\mathbf{A}, \mathbf{b}\}$ for the RBF
- RBF: $M = 500$, $\sigma = 55$, and $\lambda = 10^{-4}$, trained by cross-validation on the $2\,236$ contours of **S1**

Comparison methods

- Retraining: trains the predictive mapping from scratch on the adaptation data
- PCA alignment: finds $\{\mathbf{A}, \mathbf{b}\}$ by matching the mean and covariance (principal axes' angle and variance) of the original and the adaptation datasets
- Ground truth: retrains the predictive model with abundant data

Adaptation tasks

- Task 1 – *recover a known transformation* $\{\mathbf{A}_0, \mathbf{b}_0\}$: $\mathbf{A}_0 = \begin{pmatrix} 0.26 & 1.82 \\ -0.36 & 0.26 \end{pmatrix}$, $\mathbf{b}_0 = \begin{pmatrix} 52.8 \\ 67.4 \end{pmatrix}$
- Task 2 – *alignment between recording sessions*: to adapt the predictive model $\mathbf{f}$ of **S1** to data from **S2**



❶ – ❷ show pointwise RMSE E after adaptation/retraining w.r.t N contours (using $K = 3$ landmarks) and K landmarks (using $N = 10$). Errorbars are over 10 random choices of N. ❸ shows a joint, corrected dataset (**S1**,$\widehat{\textbf{S2}}$), obtained by using all the **S2** contours to adapt $\mathbf{f}$ from **S1** and aligning **S2** to **S1** with the resulting $\mathbf{g}$.

- Adaptation is much better than retraining and PCA alignment especially when the adaptation data is limited
- Adaptation is effective to minimize the misalignment among different recording sessions

6 Conclusion

We have proposed a fast (< 1 second) adaptation method for RBF predictive mappings of tongue contours and shown that a few contours suffice to achieve near-optimal accuracy. In synthetic transformations and in the problem of correcting for misalignments between different ultrasound recording sessions, just one contour reduces the error per contour point below 1 mm, and 10–20 contours bring it within 5–10% of the one obtained by training from scratch on a large training set. Future work will involve using more flexible adaptation models (e.g. using a different $\mathbf{A}$ and $\mathbf{b}$ per 2D point) and testing the model with data from different speakers, when the latter becomes available.

Work funded by NSF award IIS-0711186