ADAPTATION OF A TONGUE SHAPE MODEL BY LOCAL FEATURE TRANSFORMATIONS

Chao Qin, **Miguel Á. Carreira-Perpiñán** and **Mohsen Farhadloo**

EECS, School of Engineering, University of California, Merced

1 Abstract

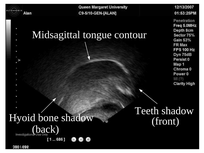
Reconstructing the full contour of the tongue from the position of 3 to 4 landmarks on it is useful in articulatory speech work. This can be done with submillimetric accuracy using nonlinear predictive mappings trained on hundreds or thousands of contours extracted from ultrasound images. Collecting and segmenting this amount of data from a speaker is difficult, so a more practical solution is to adapt a well-trained model from a reference speaker to a new speaker using a small amount of data from the latter. Previous work proposed an adaptation model with only 6 parameters and demonstrated fast, accurate results using data from one speaker only. However, the estimates of this model are biased, and we show that, when adapting to a different speaker, its performance stagnates quickly with the amount of adaptation data. We then propose an unbiased adaptation approach, based on local transformations at each contour point, that achieves a significantly lower reconstruction error with a moderate amount of adaptation data.

2 Introduction

Accurately reconstructing the full contour of the tongue from only 3–4 landmarks is possible by training a nonlinear predictive mapping on a large dataset of contours extracted from ultrasound images. But:

- Recording many ultrasound images is a tiresome and lengthy process.
- Extracting accurate tongue contours from ultrasound images is cumbersome and requires an expert user (since the automatic methods are unreliable).
- We want to obtain models for new conditions (e.g. a new recording session, recording modality, speaker or speaking style) ⇒ adaptation necessary.

Adapting a well-trained predictive model of a reference speaker given a small number of segmented contours from a new speaker is a practical solution.



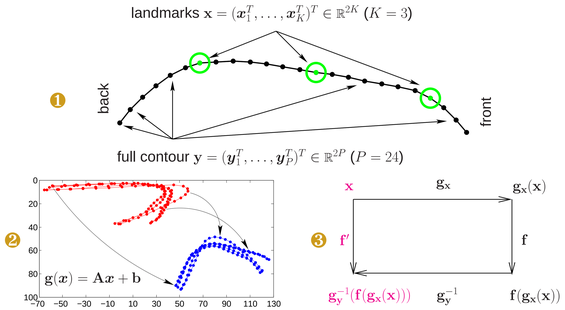
Idea of the approach

❶ Learn a predictive mapping $\mathbf{f}$ that maps the K landmarks to P contour points, from a dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}$ containing many contours.

❷ Given an adaptation dataset, estimate 2D-wise linear transformations $\mathbf{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}+\mathbf{b}$ (where $\mathbf{g}\colon \mathbb{R}^2 \to \mathbb{R}^2$) and obtain the adapted mapping $\mathbf{f}' = \mathbf{g}_\mathbf{y}^{-1} \circ \mathbf{f} \circ \mathbf{g}_\mathbf{x}$.

Advantages of our adaptation method:

- No need for any correspondence.
- Little data needed for adaptation.
- Few parameters to estimate ⇒ fast and robust.
- Applicable to both linear and nonlinear predictive mappings.
- Extendable to 3D shapes.



3 Predictive model of tongue shapes (Qin et al 2008)

The prediction problem: given the 2D locations of K landmarks ($\mathbf{x}$) on the tongue midsagittal contour, 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: $\mathbf{f}(\mathbf{x}) = \mathbf{W}\boldsymbol{\Phi}(\mathbf{x}) + \mathbf{w}$ 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.

Both mappings are estimated from contours obtained from ultrasound images. This beats spline interpolation of the landmarks.

4 Adaptation with global feature transformations (Qin & Carreira-Perpiñán 2009)

Idea: apply the same invertible transformation to each $2D$ point.
Given the predictive mapping $\mathbf{f}$ and a small number N of adaptation contours $\{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$, we estimate an invertible linear mapping $\mathbf{g}$ that transforms new data (from the new speaker) to old data (from the old speaker):

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

Our optimal solution $\{\mathbf{A}, \mathbf{b}\}$ is the minimizer of the proxy error function

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

Only 6 parameters $\mathbf{A}_{2\times 2}$ and $\mathbf{b}_{2\times 1}$ to estimate ⇒ little data needed (small N) to achieve reasonable accuracy. But:

- The accuracy stagnates as N grows (error 0.7 mm more than the ground truth).
- It doesn't allow for different transformations in different points of the tongue.
- Minimizing $F(\mathbf{A}, \mathbf{b})$ introduces a small bias in the transformation.

5 Adaptation with local feature transformations

Idea: apply different invertible transformations to each $2D$ point.
We estimate invertible linear mappings $\mathbf{g}_\mathbf{x}(\mathbf{x})$ and $\mathbf{g}_\mathbf{y}(\mathbf{y})$:

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

It is easier to work with the parameters $\mathbf{C}_i^\mathbf{y} = (\mathbf{A}_i^\mathbf{y})^{-1}$, $\mathbf{d}_i^\mathbf{y} = -(\mathbf{A}_i^\mathbf{y})^{-1}\mathbf{b}_i^\mathbf{y}$ of $\mathbf{g}_\mathbf{y}^{-1}$. The optimal solution $\{\mathbf{A}_1^\mathbf{x}, \mathbf{b}_1^\mathbf{x}, \dots, \mathbf{A}_K^\mathbf{x}, \mathbf{b}_K^\mathbf{x}, \mathbf{C}_1^\mathbf{y}, \mathbf{d}_1^\mathbf{y}, \dots, \mathbf{C}_P^\mathbf{y}, \mathbf{d}_P^\mathbf{y}\}$ is the minimizer of the the predictive square error function:

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

There are $6(K+P)$ parameters ⇒ more accurate adaptation possible (error only 0.1 to 0.3 mm more than the ground truth). Besides:

- There is no estimation bias: E is the ideal error to minimize, not F.
- Minimizing E is simpler now because the parameters are decoupled.

Optimizing $E(\mathbf{A}^\mathbf{x}, \mathbf{b}^\mathbf{x}, \mathbf{C}^\mathbf{y}, \mathbf{d}^\mathbf{y})$

We have tried two methods:

1. BFGS: we initialize from the solution $\{\mathbf{A}, \mathbf{b}\}$ obtained with global adaptation. This optimizes in the $6(K+P)$ dimensional space of $\{\mathbf{A}^\mathbf{x}, \mathbf{b}^\mathbf{x}, \mathbf{C}^\mathbf{y}, \mathbf{d}^\mathbf{y}\}$.
2. Alternating optimization:
   - Fix $(\mathbf{A}^\mathbf{x}, \mathbf{b}^\mathbf{x})$, minimize E over each $(\mathbf{C}_i^\mathbf{y}, \mathbf{d}_i^\mathbf{y})$ (P linear systems of 6×6).
   - Fix $(\mathbf{C}^\mathbf{y}, \mathbf{d}^\mathbf{y})$, minimize E over $(\mathbf{A}^\mathbf{x}, \mathbf{b}^\mathbf{x})$ by BFGS in the $6K$ dim. space.

   Alternating optimization converges slowly, but applies if P is large.

Regularizing E

To increase robustness and prevent overfitting with very little adaptation data, we penalize transformations with large condition numbers by adding the term $\lambda\mathcal{C}(\mathbf{A}^\mathbf{x}, \mathbf{C}^\mathbf{y}) = \lambda\left(\sum_{i=1}^K C(\mathbf{A}_i^\mathbf{x}) + \sum_{i=1}^P C(\mathbf{C}_i^\mathbf{y})\right)$ to the objective function, where $\lambda > 0$ and $C(\mathbf{A}) = \operatorname{tr}(\mathbf{A}^T\mathbf{A}) - D(\det(\mathbf{A}^T\mathbf{A}))^{1/D}$. This satisfies $C(\mathbf{A}) \geq 0$ and $C(\mathbf{A}) = 0 \Leftrightarrow \operatorname{cond}(\mathbf{A}) = 1$.

7 Conclusion

We have introduced a new method for fast adaptation of a tongue model based on local transformations that align each contour point separately. This is more flexible than the global method and eliminates its estimation bias. The local method asymptotes close to retraining with abundant data, and distinctly outperforms retraining and the global method when the number of adaptation contours is not very small (10 to 50). Thus the user should use the global, local, or retraining methods with less than 10, 10 to 50, and more than 50 contours, respectively.

Code at `http://eecs.ucmerced.edu`.

Work funded by NSF award IIS-0711186.

6 Experimental results

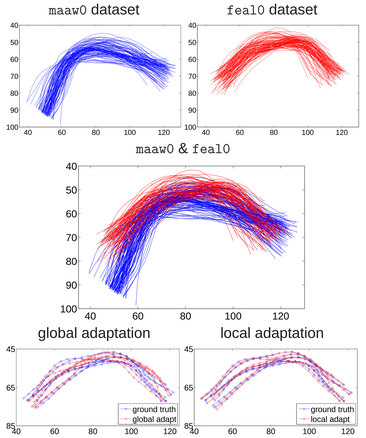
Datasets $3\,727$ ultrasound tongue contours (with $P = 24$ points) of a male speaker (`maaw0`) and $2\,909$ ultrasound tongue contours of a female speaker (`feal0`) from a set of British TIMIT sentences.

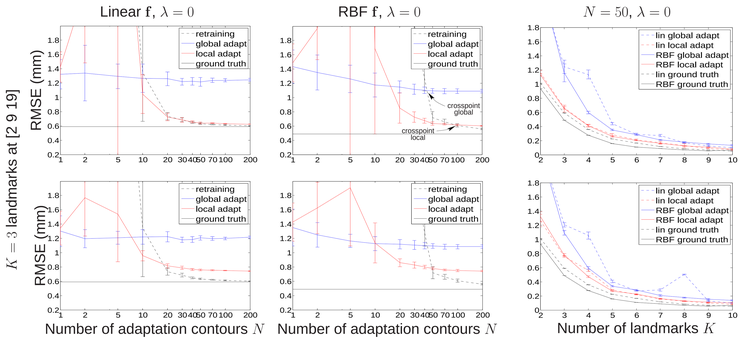
Predictive models Using the ultrasound tongue contours of `maaw0`, we learn linear and RBF predictive mappings $\mathbf{f}$.

- K landmarks were chosen optimally.
- Linear mapping: we use it as a comparison baseline and to provide initial $\{\mathbf{A}, \mathbf{b}\}$ for the RBF training.
- RBF: $M = 500$ gaussians of width $\sigma = 55$ mm trained by cross-validation on the $2\,236$ contours of `maaw0`.

Adaptation The predictive models of the male speaker `maaw0` are adapted to the female speaker `feal0` using up to 500 contours from `feal0` for adaptation/retraining and the remaining $2\,409$ contours for testing.

Comparisons The local adaptation method is compared with (1) global adaptation; (2) training from scratch on the adaptation data (retraining); (3) the best model achievable, obtained by retraining the predictive model with all 500 contours (ground truth); (4) using the predictive model directly on `feal0` (no adaptation) incurs an error > 2 mm.





Predictive error E after adaptation. Errorbars over 10 random choices of N adaptation contours. Note the crosspoints with the retraining curve, which indicate the region where each method is best:

- With $N < 10$ contours available, global adaptation performs better than local adaptation and retraining.
- With $10 < N < 50$ contours available, local adaptation performs better.
- With $N > 50$ contours available, it is better to do retraining.