ADAPTATION OF A TONGUE SHAPE MODEL BY LOCAL FEATURE TRANSFORMATIONS

Chao Qin, Miguel Á. Carreira-Perpiñán and Mohsen Farhadi
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 submetric 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 adopt 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 adaptation 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 with training a nonlinear predictive mapping on a large dataset of contours extracted from ultrasound images. But recording many ultrasound tongue movements is a tiresome and lengthy process.

- Extracting accurate tongue contours from ultrasound images is cumbersome and requires great care.
- 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.

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

The prediction problem: given the 2D locations of N landmarks (x,y) on the tongue midsagittal contour, reconstruct the entire contour (y), represented by 2D points.

- Linear mapping: we use it as a comparison baseline and linear methods are unreliable.
- We have tried two methods:
  - (i) Global adaptation: adapt the solution (A, b) obtained with global adaptation. This is the optimal in L_2 (dimensional) space of (A, b).
  - (ii) Alternating optimization:
    - Fix a parameter y, minimize x over all N adaptation contours (obtained from ultrasound images). This steers adaptation of landmarks.
    - Fix all parameters x, maximize y over all adaptation contours.

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

Idea: apply the same invertible transformation to each point.

Given the predictive mapping f and a small number N of adaptation contours (x_i, y_i), we estimate an invertible linear mapping g that transforms new data from (the new speaker) to data from (the old speaker).

Our optimal solution (A, b) is the minimizer of the proxy error function

\[ P(A, b) = \sum_{i=1}^{N} \frac{1}{2} ||g(x_i) - y_i||^2 \]

Only 6 parameters A_{d1}, b_{d1}, . . . , A_{dN}, b_{dN} to estimate in total (small N) to achieve reasonable accuracy. But:

- The accuracy stagnates as x_i grows (error 0.7 mm more than the ground truth).
- It doesn’t allow for different transformations in different points of the tongue.

Fixing P(A, b) introduces a small bias in the transformation.

5 Adaptation with local feature transformations
Idea: apply different invertible transformations to each point.

We estimate linear mappings g_i(x) and e_i(x) (i=1, . . . , N) with the ground truth y_i:

\[ g_i(x) = A_i x + b_i \]

It is easier to work with the parameters C_i, d_i, \phi_i = (A_i', b_i', C_i', d_i') of the optimal solution (A_i', b_i', C_i', d_i') since it is the minimizer of the proxy error function:

\[ P(C_i, d_i, \phi_i) = \sum_{i=1}^{N} \frac{1}{2} ||g_i(x_i) - y_i||^2 \]

There are 4 (K, \lambda) parameters \lambda so more accurate adaptation possible (error only 0.1 mm more than the ground truth). Besides:

- There is no estimation bias; f is the ideal error to minimize, not P.
- Minimizing P is simpler now because the parameters are decoupled. Optimizing (A_i', b_i', C_i', d_i') is simpler.

6 Experimental results

Table: 3D ultrasound tongue contours (with P = 46 features) in a dataset of 100 male and 100 female speakers collected with ultrasound tongue movements.

Predictive models (using the ultrasound tongue contours of a female): we learn linear and RBF predictive mappings f, g and evaluate their performance on a test set of 100 ultrasound tongue contours from an unseen female speaker.

Comparisons The local adaptation method is compared with the global adaptation (2) training from scratch on the adaptation data (retraining). The best model achievable is the ideal model with all 500 contours (ground truth) (0) using the predictive model directly on ultrasound (no adaptation) errors are 0.2 mm.

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 approximates 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 local, or retraining methods with less than 10, 10 to 50, and more than 50 contours, respectively.


Work funded by NSF award IIS-0711186.

8 Results

Number of adaptation contours

Number of ultrasound contours

Error in adaptation, over 10 random choices of N adaptation contours. Note the axes in the retraining curve, which indicate the region where each method is best:

- With 0 to 10 contours available, global adaptation performs better than local adaptation and retraining.
- With 10 to 50 contours available, local adaptation performs better.
- With > 50 contours available, it is better to do retraining.