Machine learning models of the tongue shape during speech

Miguel Á. Carreira-Perpiñán

[1]Electrical Engineering & Computer Science, University of California, Merced
mcarreira-perpinan@ucmerced.edu

Abstract. *We describe our ongoing work on data-driven models of the tongue shape. Recording techniques such as EMA and X-ray microbeam track the position of 3–4 pellets on the tongue. Our models allow a realistic reconstruction of the full shape of the tongue with submillimetric accuracy from the location of these pellets, and rapid adaptation of an existing model trained with lots of data from one speaker to a new speaker for which little data is available. These reconstruction models are useful in several applications, such as to display the tongue in a talking head animation, to visualise the vocal tract in speech production, therapy and learning, to track more robustly a signal (such as the speech or the ultrasound image), or in articulatory inversion and synthesis.*

1. Introduction

This paper describes our current work on data-driven models of the tongue shape (Qin et al., 2008; Qin and Carreira-Perpiñán, 2009, 2010; Qin et al., 2010) and some future directions. The overarching goal of our work is to obtain an accurate, realistic representation of the vocal tract of any speaker during normal, continuous speech as efficiently and cheaply as possible. Existing recording techniques provide only an incomplete answer to this problem. Electromagnetic articulography (EMA) and X–ray microbeam track at a high rate (> 100 Hz) the 2D or 3D location of around 7 landmarks (metallic pellets attached to articulator flesh points, such as the tongue tip or upper lip). Large-scale public databases such as MOCHA (Wrench, 2000) and the Wisconsin XRMB (Westbury, 1994) have proven very useful for speech research. However, the landmarks give a very sparse representation of the vocal tract (e.g. 3–4 pellets on the tongue, one or none on the velum, none in the pharyngeal cavity), and can make it hard to answer important questions such as where (and whether) the tongue touches the palate or whether the lips are open. Ultrasound gives a high-rate, detailed representation of the entire midsaggital tongue contour (see fig. 1), although segmenting and tracking it is difficult because of artifacts such as noise, invisibility of tongue parts, bone shadows or sound reflection (Li et al., 2005). Other techniques have their own limitations, such as radiation exposure (X–ray film) or large acoustic noise and low rate (MRI). All require expensive equipment, are uncomfortable for the subject, and require expert intervention to calibrate the equipment and postprocess the data to remove noise, mistracks or to segment the data. Finally, recovering the vocal tract shape directly from the acoustics (articulatory inversion) is a difficult problem and existing work has been based up to now on overly idealised models of the vocal tract, or on data-driven but landmark-based representations of it.

Models of the shape of the tongue or, in general, the vocal tract, have many applications. In speech science, they help our understanding of speech production and its

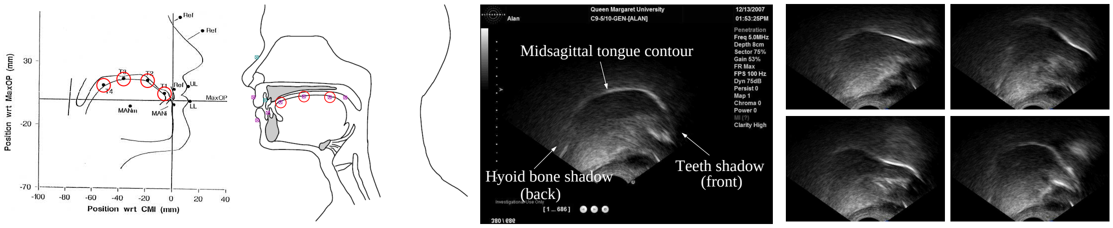


Figure 1. *Left two panels*: location of pellets in two articulatory databases: XRMB (left, 4 tongue pellets), MOCHA (right, 3 tongue pellets). *Rightmost panels*: typical tongue shapes during normal speech production in our ultrasound database for `maaw0` (lips to the right).

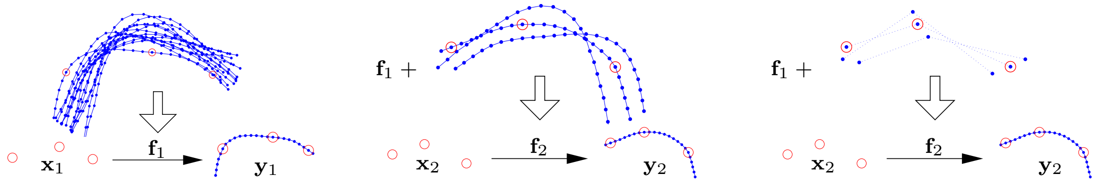
dynamics. In speech technology, they contribute to work on articulatory inversion and synthesis, as well as articulatory-based speech recognition, real-time vocal-tract visualisation and talking head animation, among others. Real-time tongue tracking using affordable imaging techniques such as ultrasound can be used as a clinical service for speech therapy and provide enhanced visualization for telemedicine devices, as well as for speech learning.

In our past work, we have focused on a specific goal: to reconstruct the full 2D midsaggital contour of the tongue given the location of 3–4 landmarks on it. This effectively results in a low-dimensional model of the tongue. A realistic tongue model must produce shapes that are both *physically feasible*, e.g. the tongue cannot penetrate the palate or the velum, and *typical*, i.e., shapes that are produced in actual, normal behaviour during speech, rather than shapes that are feasible but not normally produced. The complexity of the tongue motion make it difficult to achieve this with handcrafted models or with simple spline interpolation, so we advocate instead the use of machine learning algorithms trained on real data. However, as mentioned earlier, recording the vocal tract is expensive and cumbersome, so fast adaptation of a reference tongue model to new speakers or datasets with little data becomes crucial. We have so far defined the following three problems involving landmark-to-contour reconstruction and proposed machine learning techniques to solve them:

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

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

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



P1 is to learn a predictive model of the full tongue contour for a given speaker given many full contours from it. **P2** is to adapt the predictive model to a new speaker given a few full contours from the latter. **P3** is like **P2** but given partial contours containing only the 2D coordinates for the landmarks, and corresponds to reconstructing the tongue contours for the MOCHA and XRMB databases. The rest of this paper summarises our past work on these three problems (full details appear in the cited papers) and indicates directions of future work.

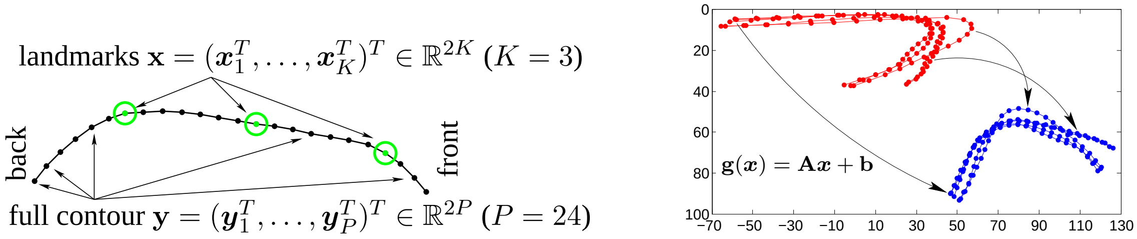

landmarks $\mathbf{x} = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_K^T)^T \in \mathbb{R}^{2K}$ $(K = 3)$
back
front
full contour $\mathbf{y} = (\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_P^T)^T \in \mathbb{R}^{2P}$ $(P = 24)$
0
20
40
60
80
100
−70
−50
−30
−10
10
30
50
70
90
110
130
$\mathbf{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \mathbf{b}$

Figure 2. *Left*: the prediction problem: 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. *Right*: our adaptation approach: transforming contours between speaker spaces with a 2D-wise mapping $\mathbf{g}$. *Red*: adaptation contours; *blue*: reference contours, recovered by $\mathbf{g}$.

2. P1: the prediction problem (with full contours)

We want to predict the full tongue contour $\mathbf{y} = (\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_P^T)^T \in \mathbb{R}^{2P}$ consisting of P points $\boldsymbol{y}_i \in \mathbb{R}^2$ given only the positions $\mathbf{x} = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_K^T)^T \in \mathbb{R}^{2K}$ of K landmarks $\boldsymbol{x}_i \in \mathbb{R}^2$ (fig. 2 left). The approach proposed in Kaburagi and Honda (1994) for *linear mappings* and in Qin et al. (2008) for *radial basis function (RBF) networks* fits a predictive mapping $\mathbf{f}$ by minimizing the predictive square error $E(\mathbf{f}) = \sum_{n=1}^{N'} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2$ (plus a regularization term for RBFs) given a sufficiently large training set, and $\mathbf{f}(\mathbf{x}) = \mathbf{W}\Phi(\mathbf{x}) + \mathbf{w}$ with M basis functions $\phi_m(\mathbf{x}) = \exp\left(-\frac{1}{2}\|(\mathbf{x} - \boldsymbol{\mu}_m)/\sigma\|^2\right)$. The RBF network is trained in an efficient but slightly suboptimal way (as commonly done) by fixing the centers $\boldsymbol{\mu}_m$ by k–means and cross-validating the number of basis functions M, width σ and the regularization parameter λ.

In order to be able to map landmarks $\mathbf{x}$ to a full tongue contour $\mathbf{y}$, we need ground-truth data for full tongue contours. We collected two datasets containing several thousand full contours (with $P = 24$ points each) from a male (`maaw0`) and a female speaker (`feal0`) with different Scottish accents using ultrasound at Queen Margaret University (Qin et al., 2008). Using the `maaw0` data, we learned an RBF network and showed it could achieve a test RMSE of 0.3–0.2 mm per point on the tongue using 3–4 pellets, respectively (note the ultrasound measurement error is about 0.4 mm per tongue point). Fig. 3 (left) shows the optimal location of the landmarks for $K = 2$ to 5. The landmarks are roughly equidistant along the tongue contour, but somewhat closer to each other near the tongue tip. The end landmarks are close to the contour ends (tip and back), but not right at the ends. The scale bar allows one to determine the positions in mm, and (after rescaling by the total tongue length) one can determine the approximately optimal placement for a different speaker. The approximate locations of the 3 pellets that were used in the MOCHA database are quite close to the optimal ones. The following recipe should yield near-optimal results: place two pellets 2 to 4 mm from the tongue ends (tip and root, i.e., as far forward and backward as possible), and place the remaining $K - 2$ pellets so all K pellets are regularly spaced. Fig. 3 (middle and right panels) shows the reconstruction error on the test set. The lowest error occurs at the landmarks themselves, and the highest error approximately in the midpoint between landmarks, or at the ends of the contour. Using only 2 landmarks yields an optimal error of 0.6 mm, while using 3 yields less than 0.3 mm and 4 yields 0.2 mm. Using more landmarks yields diminishing returns; it is also

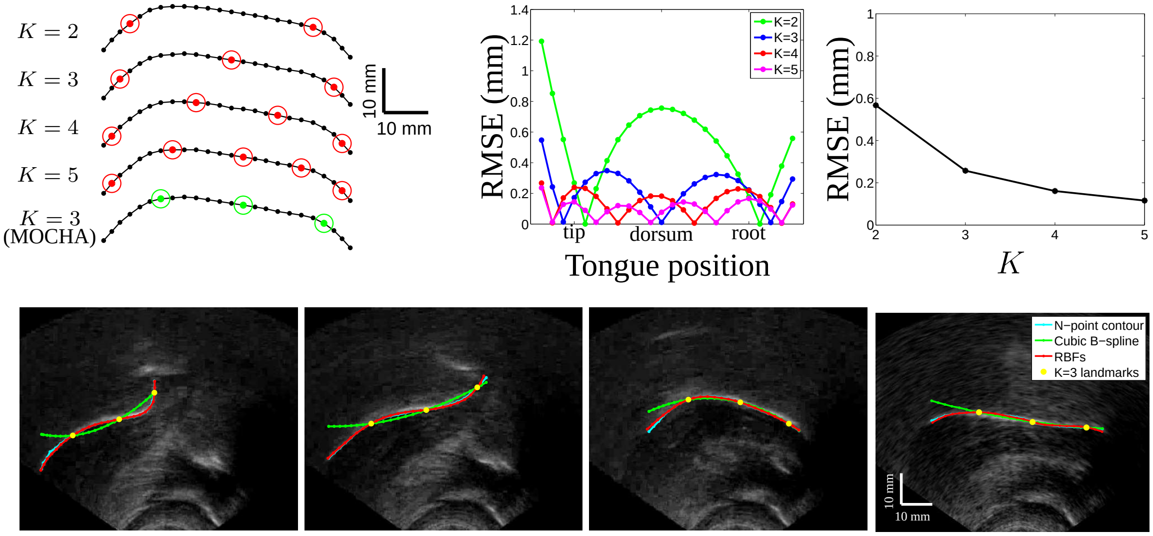


Figure 3. *Left*: optimal location of K landmarks (for $K = 2, 3, 4, 5$) depicted on a sample tongue contour (the tip is to the right and the root to the left). The bottom contour shows the approximate location of the 3 pellets used in the MOCHA database. *Middle*: error (RMSE, mm) incurred by the RBF prediction of the tongue contour wrt the ground-truth contour, for each contour point (averaged over all contours in the dataset) for different K. *Right*: average RMSE as a function of the number of landmarks K. *Bottom*: selected ultrasound frames comparing the true contour (cyan) and the contours estimated by spline interpolation (green) and our RBF prediction (red), for $K = 3$ landmarks (yellow dots).

practically harder to attach that many pellets to the tongue.

Fig. 3 (bottom) illustrates the complex shapes that the tongue can adopt, with significant changes in curvature, in particular when raising the tip. The contour predicted by the RBF network overlaps almost perfectly with the true contour, so the latter is barely visible. The contour obtained from a spline often deviates significantly from the true one. The advantage of the prediction based on a training set over a spline is largest when extrapolating beyond the end landmarks, near the root or the tip of the tongue.

3. P2: the adaptation problem with full contours

We are now given a small number N of full contours $\mathbf{y}_n$ from a new speaker. That is, each adaptation data item is a pair $(\mathbf{x}_n, \mathbf{y}_n)$ where $\mathbf{y}_n$ is the P–point contour and $\mathbf{x}_n$ the K–point input (a subset of $\mathbf{y}_n$). We follow a *feature normalization* approach (Woodland, 2001) (fig. 2 right) and adapt the existing predictive mapping $\mathbf{f}$ by estimating a *2D-wise alignment transformation* mapping $\mathbf{g}:\ \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that maps new data to old data, where $\mathbf{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \mathbf{b}$ is linear to ensure it is invertible and has few parameters (compared to mapping the entire $\mathbf{x}$– or $\mathbf{y}$–contour at once). Qin and Carreira-Perpiñán (2009) share a unique mapping $\mathbf{g}$ among all contour points (*global transformation approach*) while Qin et al. (2010) have a different mapping per point (*local transformation approach*).

Consequently, the inputs $\mathbf{x}$ and outputs $\mathbf{y}$ undergo *invertible linear transformations* $\mathbf{g}_{\mathbf{x}}$, $\mathbf{g}_{\mathbf{y}}$:

$$\tilde{\mathbf{x}} = \mathbf{g}_{\mathbf{x}}(\mathbf{x}) = \left(\begin{smallmatrix} \mathbf{A}\boldsymbol{x}_1+\mathbf{b} \\ \cdots \\ \mathbf{A}\boldsymbol{x}_K+\mathbf{b} \end{smallmatrix}\right) \text{ or } \left(\begin{smallmatrix} \mathbf{A}^{\mathbf{x}}_1\boldsymbol{x}_1+\mathbf{b}^{\mathbf{x}}_1 \\ \cdots \\ \mathbf{A}^{\mathbf{x}}_K\boldsymbol{x}_K+\mathbf{b}^{\mathbf{x}}_K \end{smallmatrix}\right), \quad \tilde{\mathbf{y}} = \mathbf{g}_{\mathbf{y}}(\mathbf{y}) = \left(\begin{smallmatrix} \mathbf{A}\boldsymbol{y}_1+\mathbf{b} \\ \cdots \\ \mathbf{A}\boldsymbol{y}_P+\mathbf{b} \end{smallmatrix}\right) \text{ or } \left(\begin{smallmatrix} \mathbf{A}^{\mathbf{y}}_1\boldsymbol{y}_1+\mathbf{b}^{\mathbf{y}}_1 \\ \cdots \\ \mathbf{A}^{\mathbf{y}}_P\boldsymbol{y}_P+\mathbf{b}^{\mathbf{y}}_P \end{smallmatrix}\right).$$

The adapted predictive mapping is given by $\mathbf{g}_{\mathbf{y}}^{-1} \circ \mathbf{f} \circ \mathbf{g}_{\mathbf{x}}$. For the local approach, adaptation requires estimating $6(K + P)$ parameters, which we do by minimising the predictive squared error between the full contours (given and predicted):

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

where we introduce new parameters $\mathbf{C}^{\mathbf{y}}_j = (\mathbf{A}^{\mathbf{y}}_j)^{-1}$, $\mathbf{d}^{\mathbf{y}}_j = -(\mathbf{A}^{\mathbf{y}}_j)^{-1}\mathbf{b}^{\mathbf{y}}_j$, so we work with $\mathbf{g}_{\mathbf{y}}^{-1}(\tilde{\mathbf{y}})$, simplifying the optimization (no matrix appears as an inverse). For the global approach, we estimate only the 6 parameters $\mathbf{A}_{2\times 2}$ and $\mathbf{b}_{2\times 1}$ by minimising instead the following error function

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

which also avoids inverses. Both error functions are efficiently optimized by the BFGS algorithm, initialized from the identity mapping. Note that we need no correspondences (pairs of inputs of the old and new speakers corresponding to the same sound).

In addition, we can regularise E by adding for each matrix $\mathbf{M}_{D\times D}$ a term of the form $C(\mathbf{M}) = \operatorname{tr}(\mathbf{M}^T\mathbf{M}) - D(\det(\mathbf{M}^T\mathbf{M}))^{1/D}$ (times a regularisation parameter $\lambda > 0$). This penalises large condition numbers in the matrices and, with very little adaptation data ($N \approx 10$), increases robustness to misspecification of landmarks and reduces overfitting.

We adapted a reference model trained on the `maaw0` speaker (male) to the `feal0` speaker (female). Figs. 4–5 show the results. The global approach achieves a test RMSE of 1.1–0.6 mm with as few as 10 adaptation contours, while the local one achieves 0.6–0.4 mm if 10–50 adaptation contours are available (for 3–4 pellets, resp.). Both methods are very effective at achieving submillimetric errors using little adaptation data and in a few seconds' CPU time. The global method is more robust with very few contours but stagnates with as few as 5 to 10 contours. The local method keeps reducing the error with more contours and stagnation happens only with many more contours, producing reconstruction results close to retraining the predictive model for the new speaker on abundant data (i.e., as in **P1**). With very few contours ($N < 10$), the local method needs regularization to reduce its variance, and performs worse than the global one. With more than 50 contours, retraining is the better option. Thus, the user has options to guide data collection and achieve the best result in each application.

4. P3: the adaptation problem with incomplete contours (for EMA/X–ray)

We are now given as adaptation data for a new speaker not the full contours with P points (as in **P2**) but only the much sparser K–landmark contours (N of them). Thus, we have no training data or ground truth for the remaining $P - K$ points at all. This is the problem with e.g. the MOCHA database, which contains the 2D locations of $K = 3$ pellets over time during speech for several unknown speakers, but not a single full contour—that is,

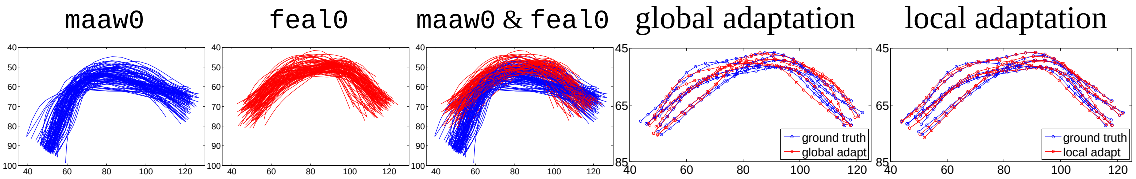


Figure 4. *Left 3 plots*: speaker datasets. *Right 2 plots*: maaw0 aligned to feal0 ($K = 3$). Only a subset of contours plotted to avoid clutter.

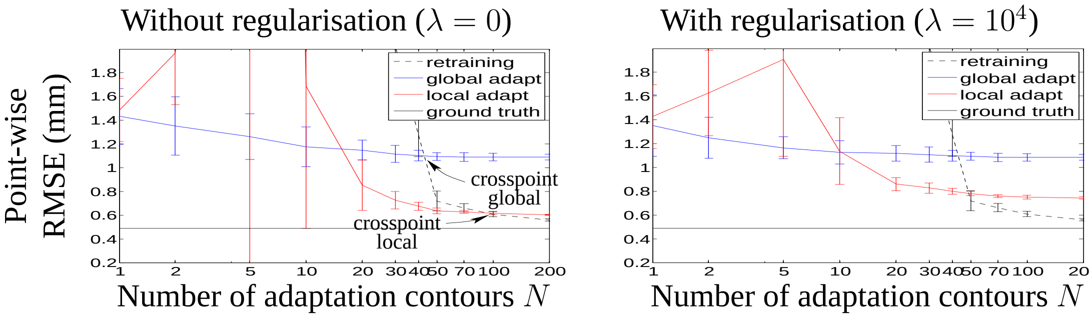


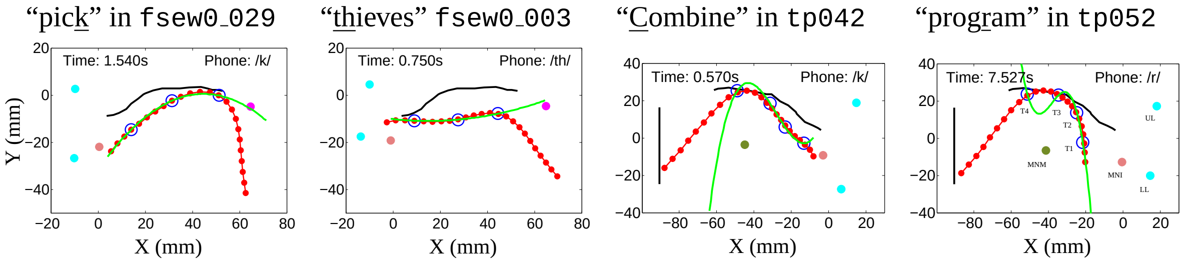
Figure 5. Predictive error E (RMSE per contour point in mm) after adaptation as a function of the number of adaptation contours N for $K = 3$ pellets at indices $[2, 9, 19]$. Errorbars over 10 random choices of the N adaptation contours. Note the crosspoints with the retraining curve.

most contour points are never observed. Qin and Carreira-Perpiñán (2010) extend the global adaptation method by considering as input $\mathbf{x}$ and also as output $\mathbf{y} = \mathbf{x}$ the pellet coordinates in these databases. We define the new problem (minimized with BFGS):

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

where $\mathbf{f}_{\mathbf{x}}$ is the components extracted from $\mathbf{f}$ corresponding to the K landmarks. This is equivalent to seeking $\{\mathbf{A}, \mathbf{b}\}$ such that the adapted model $\mathbf{g}_{\mathbf{x}}^{-1} \circ \mathbf{f}_{\mathbf{x}} \circ \mathbf{g}_{\mathbf{x}}$ best approximates the identity mapping and interpolates the landmarks. We then apply $\{\mathbf{A}, \mathbf{b}\}$ to reconstruct the entire contour as $\mathbf{g}_{\mathbf{y}}^{-1} \circ \mathbf{f} \circ \mathbf{g}_{\mathbf{x}}$. As before, we regularise F by penalising large condition numbers. The computational complexity of the adaptation algorithm per BFGS iteration is $\mathcal{O}(NMK)$ with N adaptation contours, M radial basis functions and K landmarks. Convergence occurs in around 10 iterations. Using $N = 1\,000$ takes just a few seconds.

Fig. 6 shows representative reconstructed tongue contours for speakers from the MOCHA (fsew0, female, 3 tongue pellets) and XRMB (jw11, male, 4 tongue pellets) databases. Although we do not have ground truth full contours to compare with, our reconstructions appear realistic by comparison with contours from the ultrasound database (fig. 1), correlate well with the phoneme articulation, interpolate well the K input pellets, and respect physical constraints (even though we did not impose this in any way when estimating the model): the tongue very rarely goes through the palate, velum or incisors, and moves smoothly over time. The key is in combining real measured data with flexible machine learning techniques so that they capture complex structure about the tongue motion during speech. Note how precisely reconstructed is the posterior tongue-palate contact in "pic$\underline{\text{k}}$" and the narrow alveolar constriction in "overal$\underline{\text{l}}$" and "$\underline{\text{th}}$ieves". This information, which is crucial for speech production and possibly for articulatory synthesis



“pick” in fsew0_029 “thieves” fsew0_003 “Combine” in tp042 “program” in tp052

Figure 6. Tongue reconstruction for MOCHA (left 2 panels, lips to the left) and XRMB (right two panels, lips to the right). Black curve: palate. Red curve: reconstructed tongue contour. Green curve: contour reconstructed by a cubic spline. The markers show the EMA pellets (tongue: open blue; lips: cyan; lower incisor: brown; velum: magenta).

and inversion, is not readily visible from the pellet locations alone (cf. Westbury et al., 1998). Simply using a spline to interpolate the K pellets gives a completely unrealistic reconstruction beyond the end pellets and can oscillate wildly between them.

5. Conclusion and future work

We have described a machine learning approach to learning and adapting models that can realistically reconstruct the entire 2D midsaggital tongue contour from the locations of a few landmarks with submillimetric accuracy, far outperforming using a spline interpolation. In effect, this yields a low-dimensional model of the tongue having $2K$ degrees of freedom if using K landmarks. Our models have immediate application to improving the recording setup (e.g. the choice of where to attach the pellets) and the inferences one can draw from EMA/X–ray microbeam data. For example, the reconstructed full tongue contour allows to pinpoint the time and location of a constriction (with the palate or teeth) that may not be obvious given just the recorded pellet locations. Beyond this, our models have application to speech science and technology in clinical and educational settings, as described in the Introduction.

Our approach is computationally efficient, requiring at most a few minutes of CPU time to train or adapt models with datasets containing thousands of contours. Our adaptation algorithms do not require correspondences, i.e., pairs of inputs of the old and new speakers corresponding to the same sound. The main limitation of our approach is that the models obtained are only as good as the data used to create them. In particular, adapting to shapes or sounds that are not present in the reference dataset will likely result in an imperfect reconstruction. However, the quality of the reconstructed contours for the MOCHA and XRMB datasets is remarkable even though it was achieved with a small-scale reference model (based on just a few tens of sentences recorded with ultrasound). We think that the same algorithms will be able to achieve noticeably better results with larger-scale training sets.

Our prediction and adaptation algorithms carry over with obvious modifications to 3D tongue shapes, and in general to the full vocal tract shape (e.g. as recorded with MRI) or to other shapes (e.g. for model-based segmentation of ultrasound images of the heart or other organs). Thus, our work can be seen more generally as providing an automatic, data-driven mechanism to construct and adapt realistic, low-dimensional shape models based on landmarks.

Our ongoing work includes: (1) handling missing data in the training and test sets in both prediction and adaptation, as happens in ultrasound images, where the tongue contour often disappears partially (particularly in the back and tip); (2) automatically determining the optimal level and type of regularisation needed for adaptation, depending on the amount of data available; (3) more general low-dimensional models of the tongue and vocal tract, not necessarily based on a landmark representation.

Software and data will be made available at `http://eecs.ucmerced.edu`.

Acknowledgements

Work funded by NSF awards CAREER #0754089 and #0711186. This paper is based on work done with my graduate students Chao Qin and Mohsen Farhadloo at UC Merced. We thank Alan Wrench (Queen Margaret University) for providing the ultrasound data and for useful discussions, and Korin Richmond and Steve Renals (University of Edinburgh) and Ricardo Gutiérrez-Osuna (Texas A&M University) for useful discussions.

References

Kaburagi, T. and Honda, M. Determination of sagittal tongue shape from the positions of points on the tongue surface. *J. Acoustic Soc. Amer.*, 96(3):1356–1366, 1994.

Li, M., Kambhamettu, C., and Stone, M. Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics*, 19(6–7):545–554, September 2005.

Qin, C. and Carreira-Perpiñán, M. Á. Adaptation of a predictive model of tongue shapes. In *Proc. Interspeech*, pages 772–775, 2009.

Qin, C. and Carreira-Perpiñán, M. Á. Reconstructing the full tongue contour from EMA/X-Ray microbeam. In *Proc. ICASSP*, pages 4190–4193, 2010.

Qin, C., Carreira-Perpiñán, M. Á., and Farhadloo, M. Adaptation of a tongue shape model by local feature transformations. In *Proc. Interspeech*, pages 1596–1599, 2010.

Qin, C., Carreira-Perpiñán, M. Á., Richmond, K., Wrench, A., and Renals, S. Predicting tongue shapes from a few landmark locations. In *Proc. Interspeech*, pages 2306–2309, 2008.

Westbury, J. R. *X-Ray Microbeam Speech Production Database User's Handbook Version 1.0*. University of Wisconsin, Madison, June 1994.

Westbury, J. R., Hashi, M., and Lindstrom, M. J. Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication*, 26:203–226, 1998.

Woodland, P. C. Speaker adaptation for continuous density HMMs: A review. In *Adaptation Methods for Speech Recognition, ISCA Tutorial and Research Workshop (ITRW)*, pages 11–19, Sophia Antipolis, France, August 29–30 2001.

Wrench, A. A. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. In *Phonus*, volume 5, Saarbrücken, 2000.