RECONSTRUCTING THE FULL TONGUE CONTOUR FROM EMA/X-RAY MICROBEAM

Chao Qin *Miguel Á. Carreira-Perpiñán*

EECS, School of Engineering, University of California, Merced. Merced, CA, USA
Email: {cqin,mcarreira-perpinan}@ucmerced.edu

ABSTRACT

Existing large-scale articulatory databases describe the tongue shape through the 2D positions of 3–4 fixed landmarks on the tongue surface. The ability to reconstruct the full tongue contour from these landmarks would increase the utility of these databases in speech research. We give an algorithm to adapt a predictive model of the tongue contour, that has been learned using ultrasound data for a given speaker, to a new speaker for which only landmark coordinates are given. We show realistic reconstructions of the full tongue contour in the MOCHA and XRMB databases.

Index Terms— Tongue reconstruction, model adaptation, articulatory databases.

1. INTRODUCTION

Existing large-scale articulatory databases, such as MOCHA (using electromagnetic articulography, EMA) [1] and Wisconsin XRMB (using X-ray microbeam) [2], provide the vocal tract shape during continuous speech, and have been very useful for research in articulatory inversion and synthesis, vocal tract visualization, and speech production and therapy (e.g. [3, 4]). However, their representation of the vocal tract is limited. For example, the tongue is represented by the 2D locations of 3–4 pellets attached to its tip, body and dorsum. Reconstructing the tongue shape with a spline (linear [3] or cubic [4, 5]) produces shapes that are often unrealistic and penetrate the palate, velum or teeth. Previous work [6, 7] has shown that (1) the tongue contour can be reconstructed very accurately from a few landmarks for a given speaker by training a nonlinear predictive model using tongue shapes recorded (e.g. with ultrasound) from that same speaker; and that (2) such a model can be quickly and accurately adapted to a new speaker given a few full contours for the latter. Here, we go one step beyond and propose an algorithm to reconstruct realistically the tongue shapes from articulatory databases that provide 2D coordinates for only 3–4 landmarks *but no full contours*. The method applies also to 3D shapes or full vocal tract shapes.

Reconstructing the tongue contour from landmarks using data for a single speaker has been done using linear [8, 9] and nonlinear models [6]. Nonlinear reconstruction yields lower errors and is more amenable to our proposed adaptation algorithm (see sec. 2.3). Except for [7], there seems to be little work on automatic, data-driven adaptation of such models to new speakers, although some papers have used manual adaptation of articulatory models [5]. Our work is also related to recent efforts in fusing information from different articulatory modalities such as EMA, ultrasound and MRI [10].

2. THE RECONSTRUCTION ALGORITHM

We consider three problems (fig. 1). **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.

2.1. Data-driven predictive model of the tongue contour (P1)

Given a sufficiently large training set containing N' full tongue contours $\mathbf{y} = (\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_P^T)^T \in \mathbb{R}^{2P}$ of P points $\boldsymbol{y}_i \in \mathbb{R}^2$ and the positions $\mathbf{x} = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_K^T)^T \in \mathbb{R}^{2K}$ of $K < P$ landmarks $\boldsymbol{x}_i \in \mathbb{R}^2$ (a subset of the P points), we fit a predictive mapping $\mathbf{f}$ by minimizing the predictive square error $E(\mathbf{f}) = \sum_{n=1}^{N'} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2$. The mapping $\mathbf{f}$ can be linear ($\mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{x} + \mathbf{w}$) or a radial basis function (RBF) network ($\mathbf{f}(\mathbf{x}) = \mathbf{W}\Phi(\mathbf{x}) + \mathbf{w}$ with M Gaussian basis functions $\phi_m(\mathbf{x}) = \exp\left(-\frac{1}{2}\|(\mathbf{x} - \boldsymbol{\mu}_m)/\sigma\|^2\right)$). In [6], we obtained errors below 0.3 mm using contours $\mathbf{y}_n$ extracted from ultrasound images. Note that what this essentially achieves is a realistic, data-driven model of the tongue midsaggital contour with $2K$ dof.

2.2. Adaptation of the model given full contours (P2)

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 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. The key aspect of this *feature normalization* approach is to apply the same transformation to each 2D point of an $\mathbf{x}$– or $\mathbf{y}$–contour. Consequently, the inputs $\mathbf{x}$ and outputs $\mathbf{y}$ undergo *invertible linear transformations* $\mathbf{g}_\mathbf{x}$, $\mathbf{g}_\mathbf{y}$:

$$\tilde{\mathbf{x}} = \mathbf{g}_\mathbf{x}(\mathbf{x}) = \begin{pmatrix} \mathbf{A}\boldsymbol{x}_1 + \mathbf{b} \\ \cdots \\ \mathbf{A}\boldsymbol{x}_K + \mathbf{b} \end{pmatrix} \qquad \tilde{\mathbf{y}} = \mathbf{g}_\mathbf{y}(\mathbf{y}) = \begin{pmatrix} \mathbf{A}\boldsymbol{y}_1 + \mathbf{b} \\ \cdots \\ \mathbf{A}\boldsymbol{y}_P + \mathbf{b} \end{pmatrix}. \quad (1)$$

The adapted predictive mapping is given by $\mathbf{g}_\mathbf{y}^{-1} \circ \mathbf{f} \circ \mathbf{g}_\mathbf{x}$. Then, adaptation requires estimating only the 6 parameters $\mathbf{A}_{2\times 2}$ and $\mathbf{b}_{2\times 1}$. To estimate $\{\mathbf{A}, \mathbf{b}\}$ we define the error function between the full contours (given and predicted):

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

which, in the RBF case, is efficiently optimized by the BFGS algorithm, initialized from the identity mapping. (See [7] for details.) Using only $N = 10$–20 adaptation contours and in less than 1 second CPU time, in [7] we achieved errors only slightly larger than if training with abundant data from the new speaker (i.e., as in **P1**); see also fig. 2**a**.

2.3. Adaptation without full contours (P3)

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

P1: Training a predictive model $\mathbf{f}_1$ for speaker 1 with 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 landmark positions

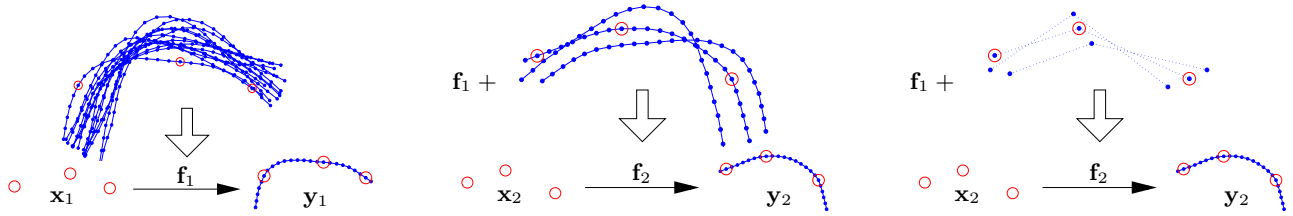

Fig. 1. Three problems involving landmark-to-contour reconstruction (here $K = 3$ landmarks and $P = 24$ points). This paper focuses on **P3**, while papers [6] and [7] focused on **P1** and **P2**, respectively.

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. Given this information alone, how can we reconstruct the full tongue contour from the K points? We propose an extension of our 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):

$$\min_{\mathbf{A},\mathbf{b}} F_{\mathbf{x}}(\mathbf{A},\mathbf{b}) = \sum_{n=1}^{N} \|\mathbf{g}_{\mathbf{x}}(\mathbf{x}_n) - \mathbf{f}_{\mathbf{x}}(\mathbf{g}_{\mathbf{x}}(\mathbf{x}_n))\|^2 \quad (3)$$

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 $\mathbf{g}_{\mathbf{y}}^{-1} \circ \mathbf{f} \circ \mathbf{g}_{\mathbf{x}}$.

Note this approach does not work if $\mathbf{f}$ is linear, because then $\mathbf{f}_{\mathbf{x}}$ became the identity when minimizing $E(\mathbf{f}) = \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{f}(\mathbf{x}_n)\|^2$ during training, and $F_{\mathbf{x}}(\mathbf{A},\mathbf{b}) = 0$ in (3) for any $\{\mathbf{A},\mathbf{b}\}$. In constrast, a Gaussian RBF with a finite number of basis functions approximates the identity to high but not perfect accuracy, and only within a finite domain of the input $\mathbf{x}$, thus (3) has a well-defined minimum that implicitly aligns the new speaker's input with the domain of the old speaker's one. Also note that *we do not need correspondences*, i.e., pairs of inputs of the old and new speakers corresponding to the same sound; achieving such correspondences is not only time-consuming but also ill defined, as it is not clear what sounds from both speakers should be considered the "same".

We have found an additional problem when applying the method proposed to our ultrasound data. Essentially, our models (trained and adapted) are as good as the data used to train the predictive mapping $\mathbf{f}$. When tracking the tongue contour in ultrasound images, it is very difficult to detect compression or stretching of the tongue because the air-tongue interface is featureless (and the tip or back of the tongue can partially disappear)—a situation similar to the aperture problem in computer vision. Thus, our training contours show mostly equidistant contour points, and we observe that a (small) proportion of the MOCHA frames show distances between pellets differing by up to 30%. If the adaptation data contains such frames, the adapted model can be far from the best one (fig. 3**a**). Note this problem is caused not by our adaptation algorithm but by our contour data, and the ultimate solution would be to collect tongue contours that show compression and stretching as naturally occurring during speech (perhaps attaching metal pellets to the subject's tongue with ultrasound imaging). It is possible to use only MOCHA frames with roughly equidistant pellets (fig. 3**b**), but this discards useful adaptation data and is unreliable. However, we have found one way of achieving very good overall adaptation with our existing data: to regularize problem (3) to encourage $\mathbf{A}$ to have a low condition number. This works because much of the misalignment between speakers can be explained by a scaling and rigid motion (which has $\mathrm{cond}\,(\mathbf{A}) = 1$), and we do observe that poorly adapted models obtained when using all sorts of MOCHA frames indeed yield a poorly conditioned $\mathbf{A}$. We then solve

$$\min_{\mathbf{A},\mathbf{b}} F_{\mathbf{x}}(\mathbf{A},\mathbf{b}) + \lambda C(\mathbf{A}), \quad \lambda \geq 0. \quad (4)$$

Directly minimizing $C(\mathbf{A}) = \mathrm{cond}\,(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$ is difficult, so we use instead the much simpler

$$C(\mathbf{A}) = \mathrm{tr}\,(\mathbf{A}^T\mathbf{A}) - D \det(\mathbf{A}^T\mathbf{A})^{1/D} \quad \text{for } \mathbf{A}_{D\times D}, \quad (5)$$

which satisfies $C(\mathbf{A}) \geq 0$ and $C(\mathbf{A}) = 0$ iff $\mathrm{cond}\,(\mathbf{A}) = 1$ (so it is minimal when $\mathrm{cond}\,(\mathbf{A})$ is minimal), and is piecewise quadratic for $D = 2$. In our experiments we find that, for a wide range of λ, this method reliably obtains the best results of all options and realistically reconstructs the full tongue contour (within and beyond the MOCHA pellets) for most frames. Frames with significantly non-equidistant pellets do show distortions, but this is unavoidable with our data.

Before training the predictive model and running the adaptation algorithm, we need to determine which of the P contour points are the K landmarks so that this matches as closely as possible the landmark locations in the new dataset. MOCHA and XRMB give approximate information as to how the pellets were attached to the tongue (e.g. "2 mm from the tongue tip") that can be used for this purpose. The location of the reconstructed tongue relative to the velum/teeth/palate can also be used to refine this estimate.

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 around 2 seconds in a PC.

3. EXPERIMENTAL RESULTS

We used the database of [7] of contours extracted from ultrasound from a Scottish speaker (`maaw0`): 2 236 full contours ($P = 24$ points and $K = 3$ landmarks) of dataset **S1** were used to train an RBF predictive model $\mathbf{f}$ with $M = 500$ basis functions, width $\sigma = 55$ mm and regularization parameter 10^{-4}, all obtained by cross-validation.

3.1. Reconstruction error with known ground truth

Here, we use additional contours (different from the training ones) from the same speaker of [7] to adapt and test the model. We transformed them using $\mathbf{A} = \left(\begin{smallmatrix} 1.1 & -0.05 \\ -0.1 & 1.2 \end{smallmatrix}\right)$ and $\mathbf{b} = \left(\begin{smallmatrix} 10 \\ -10 \end{smallmatrix}\right)$ and then split them into 991 contours for testing, and the rest (up to 500) for use in adaptation. Fig. 2**a** shows the reconstruction error (at each contour point) of our baseline models: applying the known $\{\mathbf{A},\mathbf{b}\}$

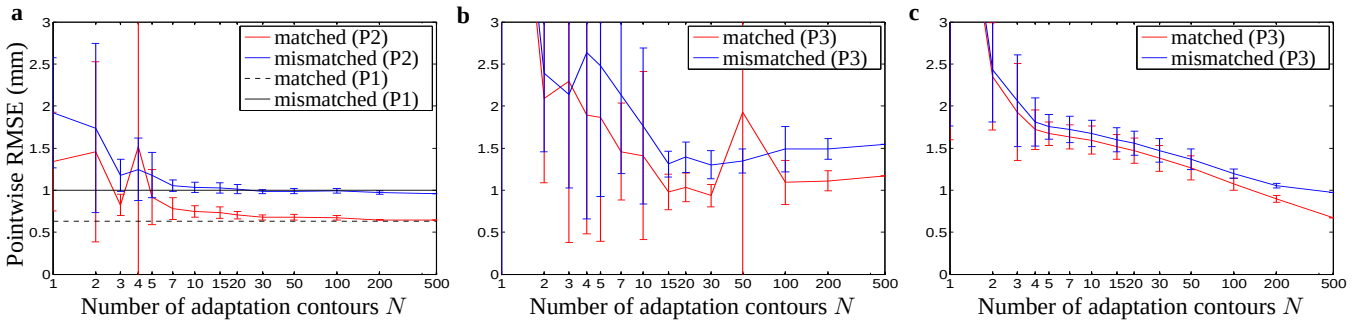

Fig. 2. Reconstruction error (RMSE at each contour point in mm) as a function of the number of adaptation contours N for two landmark placements (matched $[4\ 9\ 14]$, mismatched $[4.2\ 9.2\ 14.2]$). **a**: adaptation using full contours (problem **P2**) and training using many full contours (problem **P1**), given as baselines. **b**: adaptation using partial contours (problem **P3**) and no regularization. **c**: adaptation using partial contours (problem **P3**) and regularization ($\lambda = 2$). Errorbars over 10 random choices of the N adaptation contours.

to **f** (black line), and using different numbers N of full contours to adapt **f** (red line). These correspond to problems **P1** and **P2**, resp. Fig. 2**b** and **c** use different numbers N of partial contours to adapt **f** (red line), with and without regularization, respectively. We also consider two choices of landmarks' placement: the one used when training ("matched", points $[4\ 9\ 14]$ of the 24–point contour) and one with landmarks displaced by 20% ("mismatched", $[4.2\ 9.2\ 14.2]$).

The advantage of regularizing the condition number of $\mathbf{A}$ (both in reducing the error and its variance) is obvious, particularly with mismatched landmarks, which gives some robustness to landmark misspecification. With large enough N, the results with the correct landmark choice are almost as good as when adapting using the full contours; $N = 100$ contours suffice to reduce the error to 1 mm, while larger N can reduce it to 0.7 mm (note the measurement error in ultrasound and MOCHA/XRMB is around 0.5 mm).

3.2. Reconstruction of MOCHA/XRMB tongue contours

Figures 3–4 show results for MOCHA (speaker `fsew0`), which has $K = 3$ tongue pellets. We estimated their locations in our predictive model as $[4\ 9\ 14]$, as this gave visually the best results.

Effect of regularization and data selection. Fig. 3 shows results using $N = 3\,600$ partial contours for adaptation. In fig. 3**a** the N contours were randomly selected from the MOCHA database and no regularization was used ($\lambda = 0$). The reconstructed contour oscillates wildly, its ends are too long and it can even appear upside-down; note the diagonal of $\mathbf{A}$ has very different values of opposite sign. In fig. 3**b** we first eliminated all MOCHA contours having interpellet distance below a certain threshold (see section 2.3) and randomly selected N partial contours from these for adaptation without regularization ($\lambda = 0$). The reconstructed contours are now better, but the result is sensitive to the threshold used and we lose useful adaptation data. In fig. 3**c** we used contours without selecting them (as in 3**a**), and regularization with $\lambda = 10^4$ (for this dataset, $\lambda \in [10^2, 10^4]$ gave similar results). The reconstructed contours are the best. The resulting $\{\mathbf{A}, \mathbf{b}\}$ are essentially a translation and uniform scaling, as one might expect. These conclusions hold over different choices of the N contours and the value of N, and demonstrate the need for regularization. The experiments below use a predictive model adapted with $N = 10^4$ contours and regularization.

Realistic tongue contour reconstruction. Fig. 4 shows representative reconstructed tongue contours for MOCHA. Although, by the very nature of our goal, we do not have ground truth full contours from the MOCHA speaker to compare with, we do have strong indirect evidence that our reconstructions are quite realistic: (1) The reconstructed contours interpolate well the 3 input pellets, and they

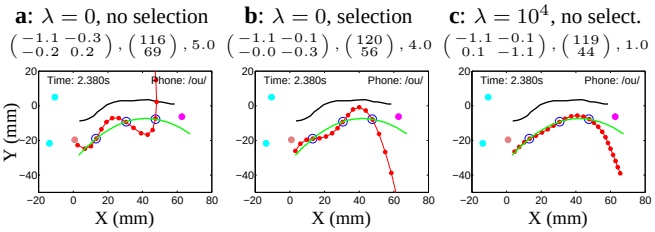

Fig. 3. Effect of regularization and data selection in adaptation in MOCHA. **a**: no regularization, randomly selected adaptation set. **b**: no regularization, carefully selected adaptation set. **c**: adaptation with regularization $\lambda = 10^4$, randomly selected adaptation set. Color scheme as in fig. 4; $\mathbf{A}$, $\mathbf{b}$ and $\mathrm{cond}\,(\mathbf{A})$ over each plot.

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 lower incisor (6% of all 10 000 frames of 460 utterances we tested, and then by less than 1 mm); see also fig. 6. (2) Comparing visually our contours with those from the ultrasound database (fig. 5) shows similar shapes, in particular in the back of the tongue (beyond the innermost pellet): "pick" (fig. 4) and frame 177/`maaw0_177` (fig. 5); "overall" (fig. 4) and frame 300/`maaw0_054` (fig. 5). (3) The contours correlate well with the phoneme articulation. Note how precisely reconstructed is the posterior tongue-palate contact in "pick" and the narrow alveolar constriction in "overall" and "thieves"; see also fig. 6. This information, which is crucial for speech production and possibly for articulatory synthesis and inversion, is not readily visible from the pellet locations alone (cf. [3]).

We do obtain less realistic reconstructions for a small proportion of frames, usually those having a small interpellet distance, or that are not well represented in our ultrasound dataset. We think this could be improved by collecting a more comprehensive contour dataset, without the need for changes in the algorithm.

Comparison with an interpolating spline. Fig. 3–4 (MOCHA) and 6 (XRMB) also show the reconstruction using a cubic interpolating spline (green curve). (For XRMB speaker `jw11`, which uses $K = 4$ tongue pellets, we chose landmarks $[3\ 7\ 11\ 15]$; other parameters as for MOCHA.) Although the spline can often give a reasonable contour between the pellets, beyond them it generally looks completely unrealistic (e.g. see "overall" or "caused" in fig. 4). The spline can also oscillate wildly between the pellets, as in fig 6. Constraining the spline a priori not to go through the palate, velum or teeth seems very difficult, while such constraints are implicitly learned in the data-driven approach we propose.

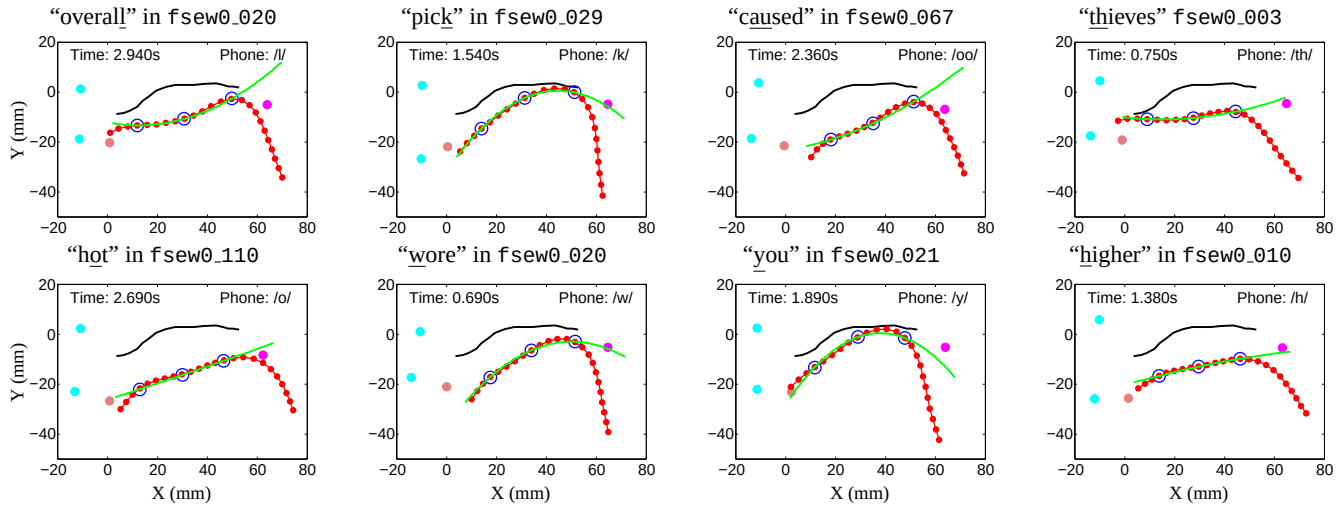

Fig. 4. Tongue reconstruction for MOCHA; $\lambda = 10^4$, $\mathbf{A} = \left(\begin{smallmatrix} -1.13 & -0.07 \\ 0.06 & -1.05 \end{smallmatrix}\right)$, $\mathbf{b} = \left(\begin{smallmatrix} 121 \\ 48 \end{smallmatrix}\right)$, $\mathrm{cond}\,(\mathbf{A}) = 1.07$. Black curve: estimate of the palate computed as the convex hull of all the tongue pellets in the entire MOCHA data for speaker `fsew0`. 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). Lips to the left. See utterance animations, and Matlab packages MOCHAtools/XRMBtools that implement the tongue reconstruction algorithm for the MOCHA/XRMB databases, at `http://faculty.ucmerced.edu/mcarreira-perpinan`.

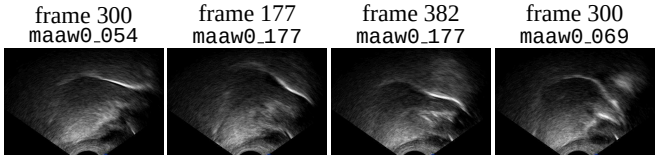

Fig. 5. Typical tongue shapes during normal speech production in the ultrasound database (lips to the right).

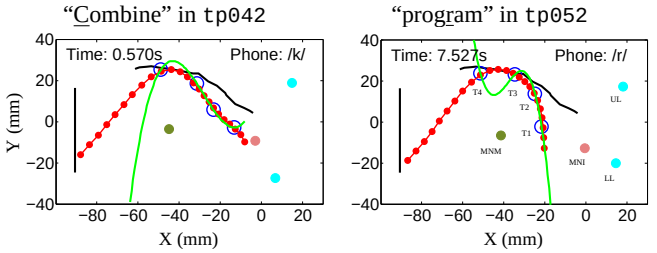

Fig. 6. XRMB results; $\lambda = 10^4$, $\mathbf{A} = \left(\begin{smallmatrix} 1.07 & -0.46 \\ -0.15 & -0.67 \end{smallmatrix}\right)$, $\mathbf{b} = \left(\begin{smallmatrix} 130 \\ 62 \end{smallmatrix}\right)$, $\mathrm{cond}\,(\mathbf{A}) = 1.75$. Color scheme as in fig. 4, but the palate was actually traced from a mold from the speaker. Lips to the right.

4. CONCLUSION

We propose an efficient algorithm that recovers realistic tongue contours for articulatory databases, based only on the 2D coordinates for the tongue pellets provided in the latter (without the need for correspondences, full contours or any other information). The reconstructed tongue satisfies physical constraints (e.g. not going through the palate, teeth or velum) without having to apply the latter explicitly, and provides detailed information not readily available in the database such as the precise location of tongue-palate constrictions. This could be very useful for research in speech production and articulatory synthesis and inversion. The algorithm is applicable to any 2D or 3D shapes and thus opens the door for reconstructing the entire vocal tract shape of an unknown speaker from a few landmarks on it, provided one can train a predictive model for a reference speaker using data for the full vocal tract of the latter (recorded with e.g. MRI or X-ray). Training, adaptation and visualization Matlab software is available from the authors. **Acknowledgments.** Work funded by NSF awards IIS–0754089 (CAREER) and IIS–0711186. XRDB funded (in part) by NIDCD grant R01 DC 00820.

5. REFERENCES

[1] A. A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," in *Phonus*, vol. 5. Institute of Phonetics, Saarbrücken, 2000.

[2] J. R. Westbury, *X-Ray Microbeam Speech Production Database User's Handbook Version 1.0*, Jun. 1994.

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

[4] C. Qin and M. Á. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Proc. Interspeech*, 2007, pp. 74–77.

[5] A. Toutios, S. Ouni, and Y. Laprie, "Protocol for a model-based evaluation of a dynamic acoustic-to-articulatory inversion method using electromagnetic articulography," in *Proc. 8th Int. Seminar on Speech Production*, 2008, pp. 317–320.

[6] C. Qin, M. Á. Carreira-Perpiñán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," in *Proc. Interspeech*, 2008, pp. 2306–2309.

[7] C. Qin and M. Á. Carreira-Perpiñán, "Adaptation of a predictive model of tongue shapes," in *Proc. Interspeech*, 2009.

[8] T. Kaburagi and M. Honda, "Determination of sagittal tongue shape from the positions of points on the tongue surface," *J. Acoustic Soc. Amer.*, vol. 96, no. 3, pp. 1356–1366, Sep. 1994.

[9] P. Badin, E. Baricchin, and A. Vilain, "Determining tongue articulation: From discrete fleshpoints to continuous shadow," in *Proc. Eurospeech*, 1997, pp. 47–50.

[10] M. Aron, M.-O. Berger, and E. Kerrien, "Multimodal fusion of electromagnetic, ultrasound and MRI data for building an articulatory model," in *Proc. 8th Int. Seminar on Speech Production*, 2008, pp. 349–352.