TONGUE TRACKING IN ULTRASOUND IMAGES WITH ACTIVE APPEARANCE MODELS

Anastasios Roussos, Athanassios Katsamanis and Petros Maragos

School of ECE, National Technical University of Athens, Greece

E-mails: {troussos,nkatsam,maragos}@cs.ntua.gr

ABSTRACT

Tongue Ultrasound imaging is widely used for human speech production analysis and modeling. In this paper, we propose a novel method to automatically detect and track the tongue contour in Ultrasound (US) videos. Our method is built on a variant of Active Appearance Modeling. It incorporates shape prior information and can estimate the entire tongue contour robustly and accurately in a sequence of US frames. Experimental evaluation demonstrates the effectiveness of our approach and its improved performance compared to previously proposed tongue tracking techniques.

1. INTRODUCTION

The shape and dynamics of the human tongue during speech are crucial in the analysis and modeling of the speech production system. Currently, ultrasound (US) imaging is one of the most convenient ways to acquire such information. It is relatively simple to use, does not expose the subject to radiation and achieves frame rates that can capture the fast tongue movement. Since even a few minutes of recorded speech correspond to tens of thousands of US frames, automatic extraction of the tongue contour at every time instant can be significantly helpful. This is a quite difficult problem, given that the US images contain high amounts of speckle noise, some parts of the tongue contour are not visible and the remaining parts are only weakly visible. In this paper, we introduce a novel approach to tackle the automatic tongue tracking problem building on Active Appearance Models.

Few methods addressing this problem are reported in the literature. Li et al. [1] developed the *EdgeTrak*, a publicly available semi-automatic system for tongue tracking in US videos. It is based on a *Snake* model that is designed for this application and incorporates information on edge gradient, intensity and contour orientation. It works quite well in frame subsequences where the same part of the tongue is visible, but when a part disappears, the corresponding tracked contour is erroneous and the method cannot afterwards recover from such errors. Therefore, this system very often needs manual refinements. More recently, Aron et al. [2] introduced various improvements. Their method is also based on *Snakes*, but preprocesses the US frames to enhance the tongue visibility and poses boundary constraints on the snake to prevent it from shrinking. In addition, the contour is initialized in every frame using the information from the optical flow between two consecutive frames and two electromagnetic (EM) sensors that are glued on the tongue. This method, which we refer to as *Constrained Snakes*, has been reported to be more accurate than the Edgetrak system. On the other hand, it also needs manual refinements, though less often than Edgetrak.

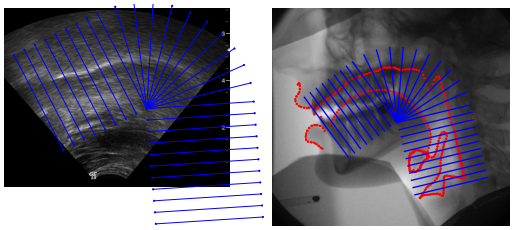
In this paper, we propose a novel tracking method that incorporates prior information about the shape variation of the contour of the tongue. This method is robust even in cases of bad tongue visibility. Further, it not only extracts the visible tongue contour parts in every frame, but also extrapolates the contour in the nonvisible parts, thanks to the model of shape variation. The methodology of Active Appearance Models (AAMs) [3, 4] is used, properly adapted to the specific application. The shape variation model of the tongue is trained on annotated X-ray videos of the speaker's head during speech. The texture model, i.e. the model of the US image intensities around the tongue contour [1], is trained on manually annotated US frames. The tracking problem is formulated in a bayesian framework, involving estimation of the model parameters in every frame. The experimental results demonstrate the effectiveness of the proposed method and its improved performance, as compared to the methods of [1] and [2].

2. PRELIMINARIES

Acquisition setup The acquisition setup that is used in this work is the one described in [2, 5]. Among other modalities, it includes US imaging of a speaker's tongue at 66 Hz. Figure 1(a) shows an example of an acquired US frame. Only a part of the intersection of the inner vocal tract wall with the midsagittal plane is visible. We refer to this intersection as *tongue contour*, since its major part corresponds to the tongue.

Using X-rays to model the tongue shape variation To model the shape variation of the tongue contour we use X-ray videos of the same speaker during speech (Fig.1(b)). In contrast to US images, the entire tongue contour is visible in the X-ray images and this makes them more appropriate for shape modeling. As will be revealed in section 3, the applied shape modeling methodology requires that the shape is

[1] As usually in AAMs, the term "*texture*" is used as in computer graphics.



(a) (b)

Fig. 1. Ultrasound and X-ray images of the speaker, with the registered Vocal Tract grid superimposed.

represented by a constant number of points. But the main difficulty in the shape modeling of the tongue is that, contrary to other applications (e.g. faces or hands), the point correspondence between frames is not obvious; it is not possible to manually annotate the same set of tongue points in different tongue images. Therefore, similarly to [6], we use a vocal tract grid (which we refer to as *VT grid*) that has a constant shape and whose *pose* (i.e. position, orientation and scale) is fixed with respect to the palate (Fig.1(b)). The intersections of the tongue contour with the lines of this grid form the points that represent the tongue shape.

Note that the tongue shape model could be based on other modalities that are easier to acquire than X-rays, such as MRI. The same model could be used for a different speaker as well after reliable adaptation.

Estimation of the VT grid's pose To build the tongue shape model and afterwards use it for tracking, it is necessary that the VT grid's pose is estimated at every X-ray and US frame (Fig.1). This can be achieved if we know the position of the palate in every frame. For the X-rays, this is straightforward, since the palate is visible. However, this is not the case for the US frames. Therefore, the MRI of the speaker's head, where the palate is also visible, is registered with every US frame. This registration is performed by the method described in [5], using position data of EM sensors on the US probe and the head.

Preprocessing of the US frames To improve the visibility of the tongue contour we first filter the US images using the method proposed in [2]. This method eliminates the US speckle patterns (see image in Fig.2) and has been found to enhance the robustness of tongue tracking. These filtered US frames, together with the VT grid's pose for each frame, are the inputs of the proposed tongue tracking system.

3. AAM-BASED TONGUE TRACKING

3.1. Representation of Tongue Appearance in US frames

Using a framework similar to AAMs [3, 4], we represent the *appearance* of the tongue in the filtered US frames. This consists of the *shape* of the tongue contour and the *texture*, namely the intensities around the visible parts of this contour.

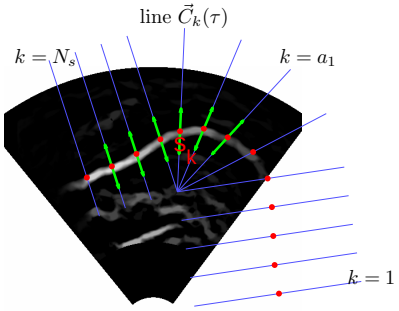


Fig. 2. Filtered US frame and representation of tongue appearance in US images. The points (dots) on the VT grid represent the tongue contour. The vectors are on the texture-active grid lines and show the windows where the image is sampled to form the texture vector.

Let the registered VT grid consist of N_s lines. The tongue shape is represented by the shape vector $\boldsymbol{s} = [s_1, .., s_{N_s}]^T$, which contains the scalars s_k that determine the intersection points of the N_s lines of the VT grid with the tongue contour. Each s_k is the distance of the intersection point from the starting point of the corresponding grid line (see Fig.2).

The texture vector $\boldsymbol{g}(\boldsymbol{s})$ represents the texture around some parts of the tongue contour, as these are defined from the grid lines subset G_{act}=$\{a_1, .., a_{N_a}\}$⊂$\{1, .., N_s\}$, which we call *texture-active grid lines* (see Sec.3.3 for how choosing this subset). The reason for using only a subset of the grid lines is that some parts of the tongue contour are never or rarely visible in the US images. The N_g-dimensional vector $\boldsymbol{g}(\boldsymbol{s})$ is formed by sampling the preprocessed US frame $u(x, y)$ in 1D windows on the grid lines $k \in G_{act}$ around the corresponding tongue points s_k (see Fig.2):

$$\boldsymbol{g}(\boldsymbol{s}) = \Big[\underbrace{[u_{a_1}(s_{a_1}+t)]^T_{t\in W}}_{1\times N_W} \cdots \underbrace{[u_{a_{N_a}}(s_{a_{N_a}}+t)]^T_{t\in W}}_{1\times N_W}\Big]^T$$

where $W = \{-d, -d+1, .., d\} \cdot \delta\ell$ is the sampling window with step $\delta\ell$ and N_W=2d+1 samples, and $u_k(\tau) = u(\vec{C}_k(\tau))$ is the restriction of $u(x, y)$ to the k-th grid line $\vec{C}_k(\tau)$. [2]

3.1.1. *Differences from classic AAMs*

Compared to conventional AAMs [3], the used representation contains various modifications, in order to exploit application-specific properties. The a priori knowledge of the pose of the VT grid in every US frame allows us to reduce the complexity of the appearance representation and model. First, the shape points are represented by the scalars s_k instead of 2D coordinates (N_s vs $2N_s$ measurements respectively). Also, there is no need to independently account

[2] Not any global linear mapping of image intensities is considered in the formation of the texture vector, since all US frames are acquired under similar conditions.

for any alignment of the shapes via similarity transforms; by using the VT grid, the shape variation caused by the movement of the speaker's palate is implicitly excluded from the model. In addition, the texture vector is formed by scanning on the grid lines around the shape points, instead of scanning entire 2D regions and warping on a mean shape, therefore it typically has much smaller dimensionality.

Due to the above simplifications, the AAM fitting involves simpler analytic expressions and can be done by solving a lighter and more reliable optimization problem than in the classic approach (Sec. 3.4). Note that the use of a grid that is designed for a specific deformable shape could be adopted for other applications as well and not only for tongue tracking. If the pose parameters of such a shape-specific grid are not known a priori as herein, they could be inferred by inserting them in the optimization process of the AAM fitting, incorporating also constraints about their dynamics.

3.2. Modeling Shape Variation

For the prior tongue shape model we exploit the manually extracted tongue contours from N_x X-ray frames of the same speaker during speech. Using the information about the VT grid position in every frame, N_x training shape vectors are extracted. As in [3], a linear model is considered for the shape vector:

$$\boldsymbol{s} \approx \boldsymbol{s_0} + \mathrm{Q_s}\boldsymbol{b}\,, \tag{1}$$

where $\boldsymbol{s_0}$ is a base shape, $\boldsymbol{b}$ is the vector that contains the N_b shape parameters, with $N_b{<}N_s$, and $\mathrm{Q_s}$ is an $N_s{\times}N_b$ matrix whose columns define the modes of shape variation. $\mathrm{Q_s}$ and $\boldsymbol{s_0}$ are statistically learned from the N_x training vectors, using Principal Component Analysis (PCA): $\boldsymbol{s_0}$ is the mean vector and the columns of $\mathrm{Q_s}$ contain eigenvectors that correspond to the N_b biggest eigenvalues of the training set's covariance matrix. The scale of each one of these eigenvectors is chosen so that $\boldsymbol{b}$ has unit covariance. Consequently, assuming that $\boldsymbol{b}$ follows a Gaussian distribution, the optimum approximation of its probability density function (pdf) is $\mathrm{p}\,(\boldsymbol{b}){=}\mathcal{N}(\boldsymbol{b}|\mathbf{0},\mathrm{I}_{N_b})$, where I_M denotes the $M{\times}M$ identity matrix and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)$ stands for the multivariate Gaussian pdf on $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

3.3. Modeling Texture Variation

We use here manual annotations from N_{us} US frames, where only the parts of the tongue contour that are visible have been marked. As texture-active are classified the VT grid lines that intersect the annotated visible parts of the tongue in more than $\pi_{act}N_{us}$ training frames. Since $\pi_{act}{<}1$, in some training US frames, the shape coordinates of some texture-active grid lines may be missing, because the tongue visibility in this lines could be absent or very low for the human annotator. In these cases, the shape vector is extrapolated by MAP estimation of the shape parameters, using the derived model of shape variation. Afterwards, the N_{us} training texture vectors can be formed. This training set is separated into 2 subsets, T_1 and T_2. In analogy to the shape modeling, a linear model is used for the texture vector:

$$\boldsymbol{g} = \boldsymbol{g_0} + \mathrm{Q_g}\boldsymbol{\lambda} + \boldsymbol{\varepsilon}\,, \tag{2}$$

where $\boldsymbol{\lambda}$ is the texture parameters vector of dimension $N_\lambda{<}N_g$ and $\boldsymbol{\varepsilon}$ is the error of the reconstruction of the texture vector $\boldsymbol{g}$ using the texture model[3]. The training set T_1 is used to learn $\boldsymbol{g_0}$ and $\mathrm{Q_g}$, using the same procedure described in Sec.3.2. Assuming a Gaussian distribution for $\boldsymbol{\lambda}$, we have $\mathrm{p}\,(\boldsymbol{\lambda}){=}\mathcal{N}(\boldsymbol{\lambda}|\mathbf{0},\mathrm{I}_{N_\lambda})$. It is also assumed that $\boldsymbol{\varepsilon}$ follows a zero mean Gaussian distribution with covariance matrix of the form $\Sigma_{\boldsymbol{\varepsilon}}{=}\widetilde{\mathrm{Q}}_\mathrm{g}\mathrm{diag}(\rho_1,..,\rho_{N_g})\widetilde{\mathrm{Q}}_\mathrm{g}^T$, where the columns of $\widetilde{\mathrm{Q}}_\mathrm{g}$ contain all the orthonormal eigenvectors of the covariance matrix of $\boldsymbol{g}$, as derived from the training set T_1. Therefore, for $\boldsymbol{\varepsilon}$ we have $\mathrm{p}\,(\boldsymbol{\varepsilon}){=}\mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0},\Sigma_{\boldsymbol{\varepsilon}})$. The optimum parameters $\rho_1,..,\rho_{N_g}$ are learned using the reconstruction errors $\{\boldsymbol{\varepsilon}_i\}$ of the training set T_2.

Note that the tongue shape and texture are considered statistically independent (*independent AAM*). The reason is that not the entire tongue contour is visible in the US frames and thus the shape model had to be learned from a different modality (X-ray) than the texture model.

3.4. Tracking via Model Fitting

In the described framework, the extraction of the tongue contour from the US frames can be achieved via fitting of the appearance model in every frame. For that, a MAP estimation of $\boldsymbol{b}$ and $\boldsymbol{\lambda}$ is used. Our goal is to maximize the posterior:

$$\mathrm{p}\,(\boldsymbol{b},\boldsymbol{\lambda}|u(x,y)) \propto \mathrm{p}\,(u|\boldsymbol{b},\boldsymbol{\lambda})\,\mathrm{p}\,(\boldsymbol{b},\boldsymbol{\lambda}) = \mathrm{p}\,(\boldsymbol{\varepsilon})\,\mathrm{p}\,(\boldsymbol{b})\,\mathrm{p}\,(\boldsymbol{\lambda})$$

where $u(x,y)$ is the filtered US frame and $\boldsymbol{\varepsilon}{=}\boldsymbol{g}(\boldsymbol{s}(\boldsymbol{b}))\text{–}\boldsymbol{g_0}\text{–}\mathrm{Q_g}\boldsymbol{\lambda}$ is the texture reconstruction error. The above maximization is equivalent to the minimization of the following energy:

$$E(\boldsymbol{b},\boldsymbol{\lambda}) = -\ln \mathrm{p}\,(\boldsymbol{b},\boldsymbol{\lambda}|u) = C{+}\tfrac{1}{2}\left\{\|\boldsymbol{b}\|^2{+}\|\boldsymbol{\lambda}\|^2{+}\boldsymbol{\varepsilon}^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\varepsilon}\right\}$$

where C is a constant. The gradients of $E(\boldsymbol{b},\boldsymbol{\lambda})$ can be easily derived, using the chain rule:

$$\begin{aligned}\nabla_{\boldsymbol{b}}E &= \boldsymbol{b} + {\mathrm{Q_s}}^T\left(\partial\boldsymbol{g}/\partial\boldsymbol{s}\right)^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\varepsilon}\\ \nabla_{\boldsymbol{\lambda}}E &= \boldsymbol{\lambda} - {\mathrm{Q_g}}^T\Sigma_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\varepsilon}\end{aligned} \tag{3}$$

where the k-th column, $1{\le}k{\le}N_s$, of the Jacobian $\partial\boldsymbol{g}/\partial\boldsymbol{s}$ is:

$$\frac{\partial\boldsymbol{g}}{\partial s_k} = \begin{cases}\left[0\cdots\cdots0\right]^T, & \text{if } k\notin G_{act}\\ \left[\underbrace{0\cdots0}_{(k-1)N_W}\,[u'_k(s_k{+}t)]_{t\in W}^T\,\underbrace{0\cdots0}_{(N_s-k)N_W}\right]^T, & \text{if } k\in G_{act}\end{cases}$$

[3]In the shape model, we do not include such an error because the tongue contour is not directly observable in the US frames.

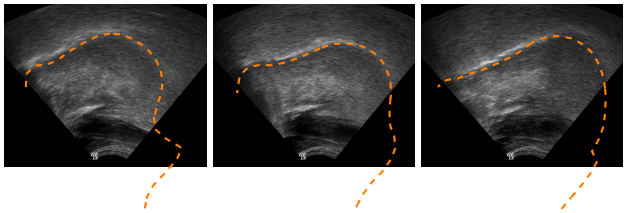


Fig. 3. Tongue tracking and extrapolation of the whole inner vocal tract wall in a US image sequence, using the proposed method: results from frames of the sequence.

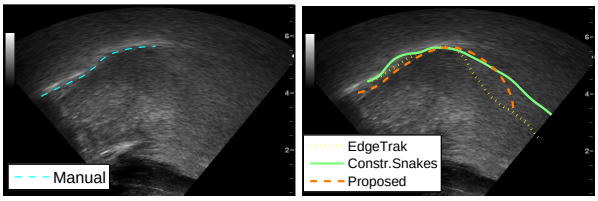


Fig. 4. Frame from a sequence where the tongue tracking methods have been applied. *Left*: manually annotated contour. *Right*: comparison of the methods' results.

We use a simple gradient descent for the minimization of E, based on (3). At every frame, we initialize $\boldsymbol{b}$ from the previous frame result and $\boldsymbol{\lambda}$ from the result of the maximization of the posterior $\mathrm{p}\left(\boldsymbol{\lambda}|\boldsymbol{g}(\boldsymbol{s}(\boldsymbol{b_0}))\right)$. Note however that more efficient minimization methods can be applied, since the 2nd-order partial derivatives of E have simple analytic forms, too.

4. EXPERIMENTAL RESULTS

In our experiments, the dimension of the shape vector was N_s=30 and the shape model was learned using N_x=700 training X-ray frames and N_b=6 parameters, so that the model could explain 96% of the shape vectors' variance. Also, the dimension of the texture vector was N_g=1215 and in the texture model training, we used N_{us}=400 US frames and we set π_{act}=50% and N_λ=35, which explained 93% of the texture vectors' variance.

Figure 3 demonstrates results of the proposed tracking method. We observe that the visible part of the tongue contour has been accurately detected and extrapolated in a sensible way. For evaluation, we have applied the proposed method, as well as the method of Edgetrak [1] (without any manual refinements) and Constrained Snakes [2], to a sequence of 200 US frames for which manually annotated tongue contours are available. In Fig. 4, an example from these results is shown (for the proposed method, only the tracked points on the texture-active grid lines are considered here). Edgetrak appears to erroneously extend the tongue contour while the Constrained Snakes demonstrate improved performance. The proposed method, which is the only one that incorporates prior shape information, yields the most plausible result. In Fig. 5, quantitative measures are shown, using the symmetric RMS distance error e_d=$\sqrt{(d_{om}^2 + d_{mo}^2)/2}$, where d_{om} (d_{mo}) is the RMS distance of the points of the output (manual) contour from the manual (output) contour. It seems that the proposed method outperforms the previous approaches. The above results demonstrate the potential of this novel approach for tongue tracking.

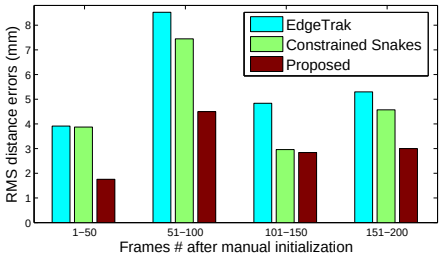


Fig. 5. Evaluation of the proposed method in comparison with state-of-the-art approaches. Manually determined tongue contours were used as ground truth. The results have been binned for better visualization.

Acknowledgements This work was supported by European Community FP6 FET *ASPI* (contract no. 021324). We would like to thank M. Aron and the rest of the *ASPI* participants for invaluable help on the articulatory data and for fruitful discussions. Also, we are grateful to G. Papandreou for fruitful discussions on AAMs.

5. REFERENCES

[1] M. Li, X. Khambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," *Clinical Linguistics and Phonetics*, vol. 6, no. 19, pp. 545–554, 2005.

[2] M. Aron, A. Roussos, M.O. Berger, E. Kerrien, and P. Maragos, "Multimodality Acquisition of Articulatory Data and Processing," in *Proc. EUSIPCO*, 2008.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. PAMI*, vol. 23, no. 6, pp. 681–685, 2001.

[4] G. Papandreou and P. Maragos, "Adaptive and constrained algorithms for inverse compositional active appearance model fitting," in *Proc. CVPR*, 2008.

[5] M. Aron, A. Toutios, M.-O. Berger, E. Kerrien, B. Wrobel-Dautcourt, and Y. Laprie, "Registration of multimodal data for estimating the parameters of an articulatory model," in *Proc. ICASSP*, 2009.

[6] S. Maeda, *Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model*, chapter in Speech Production and Speech Modeling, pp. 131–149, Kluwer, 1990.