Tongue Tracking in Ultrasound Images with Active Appearance Models

Anastasios Roussos, Athanassios Katsamanis and Petros Maragos

School of ECE, National Technical University of Athens 15773, Greece

{troussos,nkatsam,maragos}@cs.ntua.gr http://cvsp.cs.ntua.gr

1. Outline

- Ultrasound (US) imaging of speakers' tongues
  - Widely used for human speech production analysis and modeling
  - Captures the shape & dynamics of tongue during speech
  - Simple to use, no radiation, high frame rates
- Automatic tongue tracking
  - Extremely helpful for large datasets of acquired US videos
  - Difficulties: high speckle noise, non-visible tongue parts

Contributions

- We propose a novel tracking method
  - Built on a variant of Active Appearance Models
  - Incorporates prior about tongue shape variation
  - Bayesian formulation of the tracking
- Properties of the method
  - Robust even in cases of bad tongue visibility
  - Also extrapolates the contour in the non-visible parts
  - Improved performance compared to other, previously proposed techniques

2. Preliminaries

- Acquired speech articulation data of the same speaker:
  - Ultrasound imaging @ 66 Hz
  - Electro-Magnetic sensors on US probe & head @ 40 Hz
  - Magnetic Resonance Imaging (static)
  - X-ray videos @ 25 Hz
- Exploitation of the X-rays to model the tongue shape
  - The entire tongue contour is visible, in contrast to US images
  - Usage of a Vocal Tract (VT) grid to represent the tongue shape
    - fixed pose w.r.t. the speaker's palate
    - it bypasses the point correspondence problem
- Estimation of the VT grid's pose at every US frame, using EM sensors data & the head's MRI

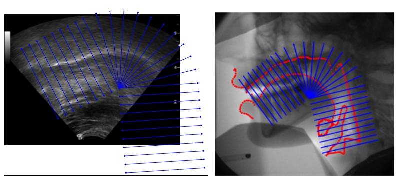


Fig. 1. Ultrasound and X-ray images of the speaker, with the registered Vocal Tract grid superimposed.

- Preprocessing of the US frames using our method of [2]

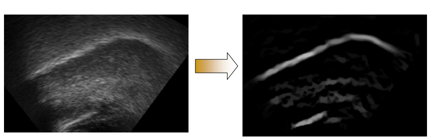
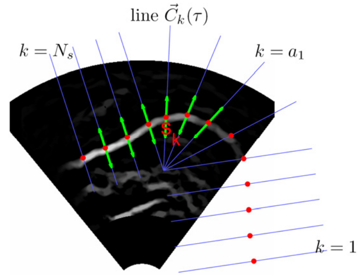


Fig. 2. Original and filtered US frame, using the method of [2]

3. Tongue Appearance Representation



- Tongue appearance
  - Shape

$$\boldsymbol{s} = [s_1, .., s_{N_s}]^T$$

  - s_k : scalar that determines the intersection of the tongue contour with the grid line k
  - Texture

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

  - Only the texture-active grid lines G_{act} are used for texture, since some parts of the tongue contour are never or rarely visible
  - $W = \{-d, -d+1, .., d\} \cdot \delta\ell$: sampling window
  - $u_k(\tau) = u(\vec{C}_k(\tau))$: restriction of the image to grid line k
- Differences from classic AAMs
  - Various modifications to exploit application-specific properties
  - Reduced complexity of the appearance representation & model
  - Lighter optimization problem for the model fitting

4. Modeling Appearance Variation

- Shape model

$$\boldsymbol{s} \approx \boldsymbol{s}_0 + \mathrm{Q_s}\boldsymbol{b}$$

- $\boldsymbol{b}$: normalized shape parameters vector with $\mathrm{p}(\boldsymbol{b}) = \mathcal{N}(\boldsymbol{b}|\boldsymbol{0}, \mathrm{I}_{N_b})$
- Principal Component Analysis (PCA) to learn $\boldsymbol{s}_0$ and $\mathrm{Q_s}$
  - Training vectors from manually annotated tongue contours on 700 X-ray frames
- Texture model

$$\boldsymbol{g} = \boldsymbol{g}_0 + \mathrm{Q_g}\boldsymbol{\lambda} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{\lambda}$: texture parameters with $\mathrm{p}(\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{0}, \mathrm{I}_{N_\lambda})$
- $\boldsymbol{\varepsilon}$: texture reconstruction error with:

$$\mathrm{p}(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \Sigma_{\boldsymbol{\varepsilon}}), \quad \Sigma_{\boldsymbol{\varepsilon}} = \widetilde{\mathrm{Q}}_{\mathrm{g}} \mathrm{diag}(\rho_1, .., \rho_{N_g}) \widetilde{\mathrm{Q}}_{\mathrm{g}}^T$$

- Training of the model
  - Manual annotations at 400 US frames. This training set is divided into 2 subsets T_1 and T_2
  - Subset T_1 is used to learn $\boldsymbol{g}_0$ and $\mathrm{Q_g}$ using PCA
  - Subset T_2 is used to learn the optimum parameters $\rho_1, .., \rho_{N_g}$

5. Tongue Tracking

- Tracking via fitting of the appearance model in every US frame
- MAP estimation of shape & texture parameters $\boldsymbol{b}$ and $\boldsymbol{\lambda}$ by maximizing:

$$\mathrm{p}(\boldsymbol{b}, \boldsymbol{\lambda}|\underbrace{u(x,y)}_{\textit{filtered US frame}}) \propto \mathrm{p}(u|\boldsymbol{b}, \boldsymbol{\lambda})\,\mathrm{p}(\boldsymbol{b}, \boldsymbol{\lambda}) = \mathrm{p}(\boldsymbol{\varepsilon})\,\mathrm{p}(\boldsymbol{b})\,\mathrm{p}(\boldsymbol{\lambda})$$

$$\boldsymbol{\varepsilon} = \boldsymbol{g}(\boldsymbol{s}(\boldsymbol{b})) - \boldsymbol{g}_0 - \mathrm{Q_g}\boldsymbol{\lambda}$$

- Equivalently: minimization of the energy:

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

- Gradients of the energy:

$$\nabla_{\boldsymbol{b}} E = \boldsymbol{b} + \mathrm{Q_s}^T (\partial \boldsymbol{g}/\partial \boldsymbol{s})^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{\varepsilon}$$

$$\nabla_{\boldsymbol{\lambda}} E = \boldsymbol{\lambda} - \mathrm{Q_g}^T \Sigma_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{\varepsilon}$$

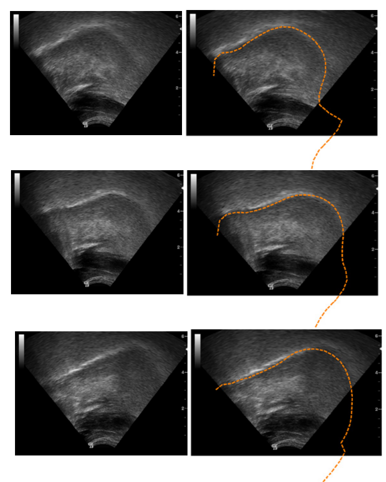
where:

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

- Optimization algorithm
  - Gradient descent
  - Initialization of parameters:
    - $\boldsymbol{b}_0$: from previous frame result
    - $\boldsymbol{\lambda}_0$: maximization of the posterior $\mathrm{p}(\boldsymbol{\lambda}|\boldsymbol{g}(\boldsymbol{s}(\boldsymbol{b}_0)))$

6. Experimental Results

I. Results of the proposed method



Original US frames *Same frames + extracted contour*

Fig. 3. Tongue tracking & extrapolation in a US image sequence, using the proposed method.

- Parameters of the method:

	Dimensionality of original vector	Number of model parameters	Variance explained (% of the total)
Shape	30	6	96%
Texture	1215	35	93%

II. Comparisons with other tracking methods

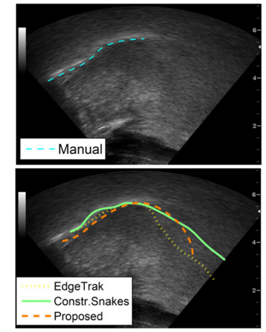
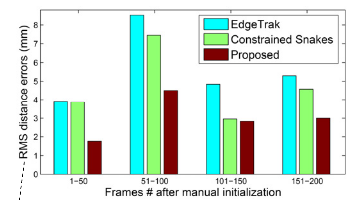


Fig. 4. Frame from a sequence where the tongue tracking methods have been applied. *Top*: manually annotated contour. *Bottom*: comparison of the methods' results.

- Quantitative evaluation



$$e_d = \sqrt{(d_{om}^2 + d_{mo}^2)/2}$$

where d_{om} (d_{mo}) is the RMS distance of the points of the output (manual) contour from the manual (output) contour.

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.

Acknowledgments : 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.