Tongue Tracking in Ultrasound Images using EigenTongue Decomposition and Artificial Neural Networks

Diandra Fabre[1,2], Thomas Hueber[1,2], Florent Bocquelet[1,2,3], Pierre Badin[1,2]

[1] Univ. Grenoble Alpes, GIPSA-Lab, F38000 Grenoble, France
[2] CNRS, GIPSA-Lab, F38000 Grenoble, France
[3] INSERM / CEA, LETI, Clinatec, Grenoble, France

{diandra.fabre, thomas.hueber, florent.bocquelet, pierre.badin}@gipsa-lab.grenoble-inp.fr

Abstract

This paper describes a machine learning approach for extracting automatically the tongue contour in ultrasound images. This method is developed in the context of visual articulatory biofeedback for speech therapy. The goal is to provide a speaker with an intuitive visualization of his/her tongue movement, in real-time, and with minimum human intervention. Contrary to most widely used techniques based on active contours, the proposed method aims at exploiting the information of all image pixels to infer the tongue contour. For that purpose, a compact representation of each image is extracted using a PCA-based decomposition technique (named EigenTongue). Artificial neural networks are then used to convert the extracted visual features into control parameters of a PCA-based tongue contour model. The proposed method is evaluated on 9 speakers, using data recorded with the ultrasound probe hold manually (as in the targeted application). Speaker-dependent experiments demonstrated the effectiveness of the proposed method (with an average error of ~1.3 mm when training from 80 manually annotated images), even when the tongue contour is poorly imaged. The performance was significantly lower in speaker-independent experiments (*i.e.* when estimating contours on an unknown speaker), likely due to anatomical differences across speakers.

Index Terms: ultrasound imaging, biofeedback, tongue, speech therapy, ANN, segmentation, speech production.

1. Introduction

Rehabilitating an articulation or swallowing disorder may require correcting the placement of the tongue. In that context, a visual biofeedback system can help a patient to better understand the origin of the trouble by displaying his/her 'own' tongue movements. Medical ultrasonography (*i.e.* 2D ultrasound imaging) is a clinically-safe and a non-invasive way ([1]) of imaging the tongue either in the midsagittal or in the coronal plane (the probe is placed beneath the speaker's chin), with both good spatial (~0.5 mm) and temporal resolution (~80 fps). Several studies showed the benefit of ultrasound biofeedback for treating different speech disorders [2] [3] [4]. However, one of the main issues of ultrasound biofeedback is the difficulty for the patient to 'read' (and thus to interpret) the images. This can be explained by several reasons [5]: (1) ultrasound image of the tongue does not show the limits of the oral cavities (such as the palate or the pharyngeal wall), and (2) some parts of the upper surface of the tongue can be very poorly imaged when they are not oriented orthogonally to the ultrasound beam. In a preliminary version of one of our biofeedback systems [6], we addressed the first issue by superimposing a generic vocal tract template to the live image stream (this template was rescaled manually to fit approximately the patient's morphology). In this study, we addressed the other issue. Our goal is to augment the ultrasound image by highlighting the tongue contour. Therefore, an automatic segmentation of the tongue contour procedure is required. In order to be used in a practical speech therapy context, this procedure should ideally: be robust to low-quality images (*i.e.* with badly imaged tongues) and probe displacements (in case it is hold manually), run in real-time, and involve minimal human intervention for adapting the system to a new patient.

Several approaches for extracting the tongue contour in ultrasound images have been proposed in the literature. One of the most widely used technique in phonetic research is an adaptation of active contours (snakes) to the problem of tongue segmentation [5] (by introducing specific shape constrains in the internal energy term). Manual intervention is required for initializing the active contours on the first frame of a sequence; the rest of the tracking remains automatic. This approach provides good results, as long as the tongue contour is clearly visible. However, the performance decreases drastically when part of the contour disappears (as mentioned in [7]). Other studies proposed to use an external model of the tongue to regularize the segmentation process. In [7], Roussos et al. proposed an approach based on an Active Appearance Model (AAM) for which the shape model was pre-trained on a set of 700 X-ray images, annotated manually (texture parameters of the AAM were derived from a filtered version of the ultrasound image). One advantage of this technique is its ability to extrapolate the tongue contour in the front and back regions of the vocal tract. Those regions are usually hidden in ultrasound by the acoustic shadows of hyoid bone and jaw. In [8], Loosvelt et al. proposed to use a biomechanical model of the tongue to constrain the displacement of contour points over time. Both approaches ([9] and [8]) seem to outperform the state-of-the-art (snake-based) technique [5]. However, the first one does not seem to be well adapted to a speech therapy context since it requires X-ray data of the patient. For the second one, it is not mentioned in [8] if evaluation was carried out on several speakers.

Statistical machine learning can also be used to address the problem of tongue segmentation in ultrasound images. In [10], Fasel & Berry proposed to model the relationships between the intensity of *all image pixels* and the position of the tongue contour, using a translational Deep Belief Network (tDBN).

This work seems to rely on the following hypothesis: the deformation of other visible structures (muscle, fat, as well as the speckle's patterns) can help recovering missing information about the tongue contour. This technique was evaluated on a multi-speaker database (7 speakers, corpus of words with the letter /l/ in frame sentences) and outperformed clearly snake-based technique (with an average error of 0.75 mm). However, as common in deep learning, a large amount of data was needed to train the model efficiently. In [10], more than 8000 annotated images were necessary to train a DBN with 646 inputs (which corresponds to all the pixels of the original image resized to 19x34), and 5514 neurons dispatched on 3 hidden layers.

The tongue segmentation method proposed in the present study is in line with Fasel & Berry's work. As detailed in Section 2, it models the statistical relationships between pixels' intensity over the entire ultrasound image and the tongue contour, using an artificial neural network (ANN). However, in order to use this system for practical biofeedback therapy, we aimed at reducing the amount of data needed to train the network (which requires human manual intervention). For that purpose, in order to alleviate the complexity of the network, we investigate the use of a more compact representation than the raw pixels, using the PCA-based decomposition technique *EigenTongue*, originally described in [11] (in the context of silent speech interface). As described in Section 3, the proposed method was evaluated on 9 speakers. The performance of the proposed method was evaluated using both a speaker-dependent and speaker-independent approaches.

2. General methodology

The proposed technique is based on: (1) a PCA-based decomposition technique named *EigenTongue* used to parameterize the pixels intensity, (2) a PCA-based model of the tongue contour referred here to as *EigenContour*, and (3) an artificial neural network for modeling the relationships between these two representations. A block diagram of the method is given in Figure 1. The different steps involved in this method are detailed in the next subsections.

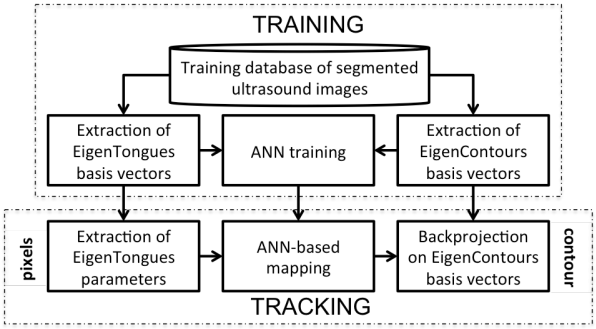

Figure 1: *Block diagram of the proposed segmentation technique*

2.1. EigenTongue decomposition

The EigenTongue decomposition technique is a straightforward adaptation of the Eigenfaces method proposed by [12]. Previous studies on articulatory recognition [13], articulatory-acoustic mapping [14], and cross-speaker articulatory mapping [15] showed that this technique provides a compact and articulatory-consistent representation of tongue position in ultrasound images. Besides, it encodes also the other visible structure in the image (muscle, fat, speckle patterns, etc.), which can be exploited to recover missing information about the tongue contours. The EigenTongue decomposition technique can be summarized as follows. In the training stage, a subset of ultrasound frames is selected from the recorded dataset and resized to limit some of the redundancy between pixels. A decomposition basis that best explains the variation of pixel intensity in the training frames was then extracted using PCA (basis vectors for ultrasound are called *EigenTongues*). In the feature extraction stage, each new ultrasound frame is projected onto the set of EigenTongues. The features used for the mapping experiments are defined as the first components in that space (such features are referred here to as *EigenTongues parameters*). The number of components is typically determined by keeping the eigenvectors that carry 80% of the variance of the training set.

2.2. EigenContour decomposition

In the training stage, the tongue contour is annotated manually with respect to a polar grid, as illustrated in Figure 2. The geometry of this grid is adapted to the morphology of the speaker. It is centered on the probe position (given by the ultrasound system). The two extreme left and right grid lines (back and front of the mouth) are aligned on the acoustic shadows created by the hyoid bone and the jaw, respectively. This grid positioning aims at dealing with morphological differences across speakers. The manual annotation is facilitated by the use of Bézier splines: the user manipulates control points of the splines instead of pointing each grid-contour intersection. The tongue contour is represented by the x/y coordinates of its intersections with each grid lines. If the tongue contour does not cross a specific grid, the corresponding point is labeled as 'missing' (this could happen for the extreme right lines of the grid for back tongue articulation).

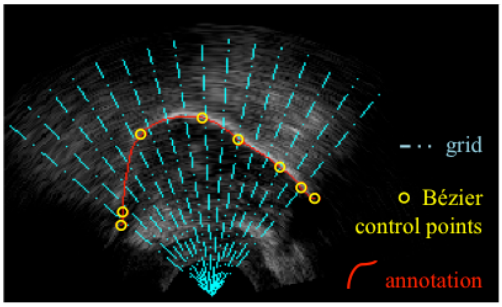

Figure 2: *Example of annotation grid*

Similarly to the ultrasound images, PCA is used to obtain a compact representation of the annotated tongue contours. However, a probabilistic PCA (p-PCA) was used to deal with missing valued in annotated tongue contours [16]. Similarly to the EigenTongue decomposition techniques (section 2.1), the basis vectors obtained by p-PCA are called EigenContours. For the ANN-based mapping experiments, each contour of the training set is represented by the first components in the EigenContours space (the resulting features are referred here to as *EigenContours parameters*).

In the training stage, the relationships between EigenTongues and EigenContours parameters are modeled using a single-layer perceptron. In the segmentation stage (AAN-based

mapping), the estimated EigenContours parameters are back-projected onto the matrix of basis vectors in order to get the x/y coordinates of the tongue contour. Segmentation is performed on a frame-by-frame basis, is computationally light and can thus run in real-time.

3. Experimental protocol and results

3.1. Data acquisition

For this particular study, we recorded a database of 9 speakers (4 females and 5 males). In order to be as close as possible to a practical speech therapy context, each speaker was asked to manually hold the probe him/herself. Tongue movements were captured in the midsagittal plane using the Terason T3000 system, a 128 elements micro convex transducer, and the *Ultraspeech* software [17]. Ultrasound frequency range was set to 3-5 MHz, scanning angle to 140°, and penetration depth to 7 cm. Ultrasound images were recorded at 60 Hz with a resolution of 320×240 pixels. At this size, the resolution of the ultrasound images was 0.5 mm / pixel. Each speaker was asked to utter two repetitions of the first list of the Combescure corpus of phonetically balanced sentences [18], for a total of 20 sentences. This results in approximately 5000 images per speaker.

3.2. Training data selection

A preliminary experiment on a single speaker, not included in our database, demonstrated that decreasing the number of training images from 1200 to 100 would increase the error by about 0.5 mm only. Thus, instead of randomly choosing a set of images for each speaker, we divided the data in 20 clusters by using the k-means algorithm on the EigenTongues parameters. The number of clusters was determined based on the phonetic knowledge of the different tongue articulations existing in French. We picked 5 images by cluster. This procedure led to a final set of 100 images for each speaker. The resulting 900 images were then manually annotated, using the procedure described in Section 2.2.

3.3. ANN training procedure

ANN were trained similarly to [19], with the following specificities. Weights were first randomly initialized using a Gaussian distribution with a 0.0001 standard deviation. The error criterion was defined as the Mean Squared Error (MSE) between predicted and expected values. The minimization of the error was performed with the conjugate gradient method using a 3 lines search, on successive batches: at each epoch, the training data samples were randomly shuffled. Non-linear units used the logistic sigmoid as activation function. Input and output data were z-scored before being fed to the ANN. Regularization of the ANN during training was done both using early-stopping and adding an L2 penalty to the network weights (the L2-norm of the weights were included in the cost function, with a cost-factor of 0.01).

3.4. Metrics and evaluation

Similarly to [5], the accuracy of the estimated tongue contour on the k^{th} image was assessed using the MSD measurement (Mean Sum of Distance), defined as:

$$MSD_k = \frac{1}{2N}\sum_{i=1}^{N}\left(\min_j\left(v_i \to u_j\right)+\min_j\left(u_i \to v_j\right)\right) \quad (1)$$

where $u_i \to v_j$ denotes the Euclidean distance between the i^{th} point u_i of the estimated contour and the j^{th} point v_j of the (manually) annotated contour (N *is* the number of grid lines). The average error in millimeters over the K images of the test set was defined as: $MSD_{RMS} = R \cdot \sqrt{(1/K)\sum_{k=1}^{K} MSD_k^2}$ where R is the resolution of the ultrasound system (R=0.5mm/pixel in our case). For each of the 9 recorded speakers, all experiments were systematically assessed using a complete 5-fold cross-validation procedure. For each of the 5 folds, 80 images were used to train the ANN (i.e. estimating the model parameters and triggering early stopping), the remaining 20 images were used for test.

3.5. Speaker-dependent approach

The performance of the proposed method was evaluated for the 9 recorded speakers, first using a *speaker-dependent* approach. In this scenario, both EigenTongues and EigenContours models were estimated from single speaker data as well as the ANN parameters.

Ultrasound images were resized to 64×64 pixels. A set of 40 EigenTongues parameters was retained for parameterization (this number of components explained at least 80% of the observed variance, for all the 9 speakers).

To quantify the amount of manual intervention needed by the proposed method, we explored the evolution of the performance as a function of the size of the training corpus (and thus the number of manually annotated images). For each fold of the cross-validation procedure, we selected successively the 20, 40, 60, and 80 most distinct images from the available 80 training images using a k-means algorithm (for these experiments, the set of EigenTongues remained calculated from the 80 available frames).

A set of 8 EigenContours parameters was used to describe the tongue (explaining 90% of the variance). The error obtained by reconstructing the original tongue contours from only the

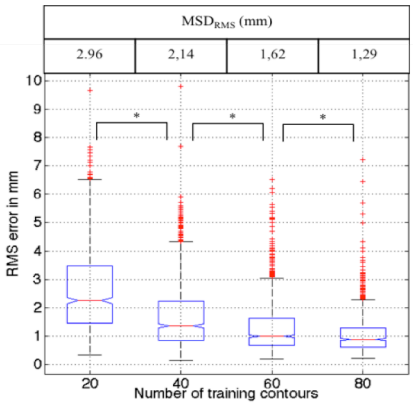

Figure 3: *Performance obtained for the speaker-dependent experiments as a function of the amount of training images. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. (*) denotes a statistically significant difference between conditions, evaluated using an ANOVA test.*

corresponding 8 EigenContours parameters was found to be 0.28 mm. This value is therefore the *best possible performance* with this parameterization.

The number of neurons on the ANN hidden layer was determined using cross-validation. However, no improvements were observed with more than 35 neurons, which was the ANN configuration used for all the speaker-dependent experiments. The performances obtained for each size of the training set are presented in Figure 3.

As expected, the best performance was obtained when using the largest training set (1.29 mm when using 80 training images, compared to 2.96 mm when using only 20 training images). A good tradeoff between manual intervention and accuracy of the segmentation process can be obtained with 60 images (1.62 mm). Besides, as shown in Figure 4, the proposed method seems to deal correctly with low-quality images for which parts of the tongue contour are totally missing. This supports our approach based on the use of information of the whole image when segmenting the tongue contour.

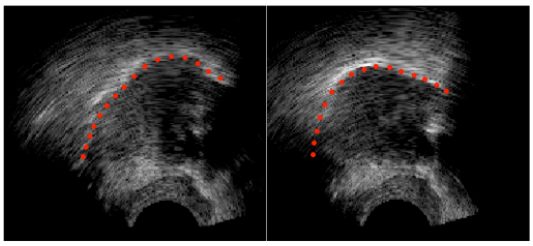

Figure 3: *Examples of images where parts of the tongue contour are missing. In red, the segmentation performed by the neural network.*

3.6. Speaker-independent approach

The performance of the proposed method was then evaluated using a *speaker-independent* approach. In this scenario, the system was trained on a multi-speaker database (8 speakers, 100 images each), and was evaluated on a so-called 'target speaker' (which was not seen during training). These experiments aim at evaluating the capability of the ANN to generalize to a new speaker's morphology and articulatory idiosyncrasies (to our knowledge, this problem has so far not been addressed in the literature). Two scenarios were investigated. In the first one, no observations of the target speaker were included in the training dataset. This corresponds to the 'ideal case', for which the proposed segmentation method becomes fully automatic. In the second one, we added some observations of the target speaker in the training set. In that case, the proposed segmentation method requires some manual intervention. We evaluated the performance for different amount of prior knowledge: 20, 40, 60 images, and finally 80 images (*i.e.* as much information about the target speaker as the other speakers of the database). Note that both the EigenContours model and the ANN were re-estimated for each experiment. 80 EigenTongues parameters were used for all speaker-independent experiments to encode ultrasound images. As expected, this number was higher than for the speaker-dependent experiments, likely due to the inter-speaker variability. The results are presented in Figure 5.

As expected, the best results (1.89 mm) were obtained when using as much as information about the target speaker as possible (*i.e.* 80 images in this experiment), but no significant differences were obtained when using only 60 images (1.94 mm). Nevertheless, for these cases, the performance of the speaker-independent system is lower than the one observed with the speaker-dependent approach.

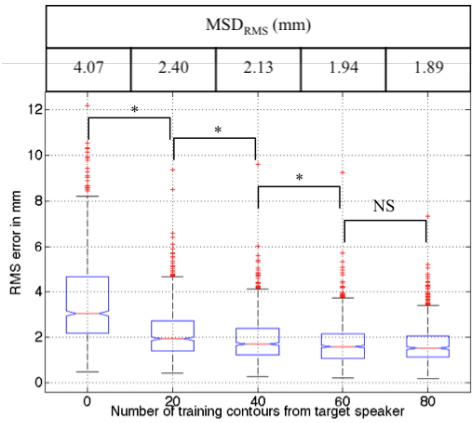

Figure 4: *Performance obtained for the speaker-independent experiments as a function of the amount of training images of the target speaker*

Surprisingly, prior information derived from other speakers degrades the mapping instead of improving it. Interestingly, the opposite effect is observed when introducing less than 40 images, since the speaker-independent system outperforms significantly the speaker-dependent one. Here, the model seems to extrapolate from other speaker's data. As a consequence, in a practical context of biofeedback therapy, the speaker-independent approach will be preferred when minimum manual intervention is required. However, the speaker-dependent approach will lead to a better performance when more than 40 images can be annotated by the system user.

4. Conclusions and perspectives

This article describes a new approach for segmenting the tongue contour in an ultrasound image. This method is developed in the context of biofeedback speech therapy, in order to 'augment' the ultrasound image by highlighting the tongue contour. The proposed method is based on statistical machine learning. It aims at exploiting all the information available in the image: indeed the areas corresponding to the tongue, but also the other visible structures (muscles, fat, speckle pattern, etc.). Experimental results demonstrate the effectiveness of this approach, even when parts of the tongue are badly imaged.

Future work will focus on the introduction of dynamic constrains (by using contextual information about previous segmented frames). A real-time implementation should be also done in order to evaluate the proposed system of ultrasound biofeedback in a practical context of speech therapy.

5. Acknowledgements

The authors would like to thank the Région Rhône-Alpes ARC6 funding agency for supporting this work through doctoral funding to Diandra Fabre. They also thank all the speakers recorded for this experiment.

6. References

[1] M. A. Epstein, "Ultrasound and the IRB," *Clinical Linguistics & Phonetics*, vol. 19, pp. 567-572, 2005.

[2] J. L. Preston and M. Leaman, "Ultrasound visual feedback for acquired apraxia of speech: A case report," *Aphasiology*, vol. 28, pp. 278-295, 2014.

[3] H. M. Lipetz and B. M. Bernhardt, "A multi-modal approach to intervention for one adolescent's frontal lisp," *Clinical linguistics & phonetics*, vol. 27, pp. 1-17, 2012.

[4] G. Modha, B. Bernhardt, R. Church, and P. Bacsfalvi, "Case study using ultrasound to treat," *International Journal of Language & Communication Disorders*, vol. 43, pp. 323-329, 2008.

[5] M. Li, C. Kambhamettu, and M. Stone, "Automatic contour tracking in ultrasound images," *Clinical Linguistics & Phonetics*, vol. 19, pp. 545-554, 2005.

[6] T. Hueber, G. Chollet, B. Denby, and M. Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," *Proc. of ISSP*, pp. 365-369, 2008.

[7] A. Roussos, A. Katsamanis, and P. Maragos, "Tongue tracking in ultrasound images with active appearance models," presented at Image Processing (ICIP), 2009 16th IEEE International Conference on, 2009.

[8] M. Loosvelt, P.-F. Villard, and M.-O. Berger, "Using a biomechanical model for tongue tracking in ultrasound images," in *Biomedical Simulation*: Springer, 2014, pp. 67-75.

[9] A. Roussos, A. Katsamanis, and P. Maragos, "Tongue tracking in ultrasound images with active appearance models," presented at Image Processing (ICIP), 2009 16th IEEE International Conference on, 2009.

[10] I. Fasel and J. Berry, "Deep belief networks for real-time extraction of tongue contours from ultrasound during speech," presented at Pattern Recognition (ICPR), 2010 20th International Conference on, 2010.

[11] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue feature extraction for an ultrasound-based silent speech interface," presented at Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, 2007.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, pp. 71-86, 1991.

[13] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," presented at INTERSPEECH, 2008.

[14] T. Hueber, G. Bailly, and B. Denby, "Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface," presented at 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), 2012.

[15] D. Fabre, T. Hueber, and P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression," presented at 15th Annual Conference of the International Speech Communication Association (Interspeech 2014), 2014.

[16] J. Porta, J. Verbeek, and B. Krose, "Active appearance-based robot localization using stereo vision," *Autonomous Robots*, vol. 18, pp. 59-80, 2005.

[17] T. Hueber, G. Chollet, B. Denby, and M. Stone, "Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application," *Proc. of ISSP*, pp. 365-369, 2008.

[18] P. Combescure, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'Acoustique*, vol. 56, pp. 34-38, 1981.

[19] F. Bocquelet, T. Hueber, P. Badin, L. Girin, and B. Yvert, "Robust articulatory speech synthesis using deep neural networks for BCI applications," presented at Interspeech 2014 (15th Annual Conference of the International Speech Communication Association), Singapour, 2014.