ATLANTIS PRESS

*Advances in Intelligent Systems Research, volume 168*

International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)

# Active Appearance Model Based Contour Extraction for MRI Images of Human Tongue

Zhi-cheng LIU[1], Qi-long SUN[2] and Jian-guo WEI[1]

[1]School of Computer Software, Tianjin University, Weijin Road, Tianjin, 300072, China

[2]School of Computer Science, Qinghai Nationalities University, Bayi Road, Xining, 810007, China

**Keywords:** Tongue Contour, MRI images, Speech production.

**Abstract.** In this article, we present the results of automatic extraction of speech articulator contours from Magnetic Resonance Imaging movie by employing the Active Appearance Model. An Active Appearance Model based framework is proposed to deal with the high nonlinear property of articulatory deformation during articulation, which demonstrates the advantage for tracking articulators shape from noisy MRI images. The extraction of the vocal tract contour was carried on MRI movies from Chinese subjects. The performance of this framework was evaluated by comparing manually labeled contours with automatically extracted ones. The average error is around 2.1 pixels.

## Introduction

Speech is one of the most important functions of human communication. However, the mechanism of speech production is far from being fully discovered. The morphological and dynamic aspects of speech organs are the essential for understanding the knowledge of speech dynamic. Advanced imaging and image processing technologies are important for this research field.

Magnetic Resonance Imaging (MRI) is able to produce high-resolution images of human articulators. This function makes MRI currently one of the most promising means for speech research and hence has been widely used in study speech production [1-3]. A set of databases of MRI image of human speech organs have been available for various purposes. A necessary procedure to use such databases, however, is a successful extraction of the desired speech organs from these images. A large variety of algorithms have been developed over the last few decades trying to handle this issue [4-6]. They mainly can be categorized as data-driven approach such as snake-like methods and model-driven approach that use the prior knowledge to complete the task. Both categories have their own pros and cons. For data-driven approach, each image has to be given an initial shape before extracting the shape, which could not be fully automatic. The model-based approach has to be trained by a training set, which has to be labeled manually beforehand.

Active Appearance Models (AAM) is one of the model-based approaches, which has been shown that it has great promising for automatically tracking objects from images. As MRI database of speech has a large number of images for recording articulatory movements, it is worthy to label a small training set for automatically extracting the shape from remaining images.

AAM was developed by Cootes et .al [7-10], which is a statistical point distribution model (PDM). AAM has demonstrated its capability for image segmentation [11]. It is able to automatically learn the parameters of the PDMs from sets of corresponding landmarks as well as incorporating the shape and boundary gray-level information. An AAM describes the image appearance and shape of object of interest by obtaining a statistical shape-appearance model from a training set. AAM minimize the difference between the synthesized image from the model and an unseen image by tuning the model parameters， when it is applied to image interpretation or segmentation. AAM has demonstrated high robust for segmentation in Cardiac MRI images and face feature extraction. The articulators such as tongue, soft palate and lips, however, are highly deformable organs than face and heart. In this research we adopt AAM as a mean for extracting tongue and palate contours from MRI image sequences as well as the contours of the profile view of upper and lower lips.

### The Dataset of This Research

In this section, we briefly introduce the procedure for obtaining the MRI data, including the speech materials and subjects' selection, the paradigm of MRI experiments.

The MRI data were acquired using the Shimadzu-Marconi ECLIPSE 1.5T Power Drive 250 installed at the Brain Activity Imaging Center, Advanced Telecommunications Research Institute (ATR-BAIC), Kyoto, Japan.

There are nine single Mandarin vowels /a o e i u ü (i)e (s)i (sh)i / . These vowels have been uttered by saying the Chinese characters "啊喔嗣衣乌淤噎思诗" with first tone(high flat tone) to ensure the stability of the sustained vowels.

As for the dynamic movements, There are 39 syllables were selected, including diphthongs (e.g. /ai ei ao/), triphthongs (e.g. /iao iou uai/), and CV syllables (e.g. /ba bi bu/).

As well known, the major drawback of MRI is its poor time resolution. To solve this shortcoming, a synchronized sampling method (SSM) developed by [2] is adopted to record the movements of the speech organs as a set of sequential images, namely, MRI movie. This method can also be used for acquiring the static 3D shape of vowels. The details of data acquisition can refer to [12].

Figure 1. An example of MRI image of human speech organs

### The AAM Based Contour Extraction

The Active Appearance Model (AAM) [7] can encode both shape and texture information of an image, which show inherent benefit over traditional appearance based methods. The AAM approach becomes popular in face recognition and representation researches recently. In this section we propose a procedure of extracting tongue movements from MR images based on AAM framework.

### Introduction of Active Appearance Models

AAM [8] learns from a training set to create a compact parameterization of the variability of an object's properties. The AAM generates a statistical appearance model for matching a combined model of shape and texture, by combining a model of shape variations with a model of texture variations. Object shape in training set is defined manually, semi-automatically or automatically labeling it with landmarks. For instance, a planar shape with n points for each training image i is represented as a 2n 1 vector containing x and y coordinates of the landmarks:

$$\mathbf{S_i} = [x_{1i}, x_{2i}, \ldots, x_{ni}, y_{1i}, y_{2i}, \ldots, y_{ni}, ]^T \quad (1)$$

For m training images, the mean shape is obtained from a 2n×m matrix. The shape examples are aligned to a common mean shape $\bar{\mathbf{s}}$ by using generalized procrustes analysis. This geometrical normalized frame represents the shape-free reference where the texture samples are extracted with associated pixel information. The mean texture $\bar{\mathbf{g}}$ is then obtained from the set of warped images. After geometrical normalization, principal component analysis (PCA) is applied to build the statistical shape and texture models. The equations of shape and texture models are:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P_s b_s}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \quad (2)$$

Where the $\mathbf{P_s}$ and $\mathbf{P_g}$ are the eigenvectors of shape and texture covariance matrices, they describe a sufficient fraction of the total shape variation. $\mathbf{b_s}$ and $\mathbf{b_g}$ are the vectors containing the coefficients for shape and texture respectively.

PCA is further applied on the concatenated shape and texture coefficients, from which process the subspace transform **T** between the new coefficients **c** and **b**. The $\mathbf{W_s}$ is a diagonal matrix that balances the energy discrepancy between the shape and texture models [8].

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix}$$

$$\mathbf{b} = \mathbf{Tc} \quad (3)$$

The appearance model control the shape and texture by the coefficients vector c according to:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q_s c}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q_g c} \quad (4)$$

Where matrices $\mathbf{Q_s}$ and $\mathbf{Q_g}$ indicate the modes of variation derived from the training set [8]. The AAM adopt prior knowledge of objects for fitting the model to unseen images.

**Framework for Tracking Contour of Articulators.**

The deformation of the articulators during movements is quite large due to its elastic property. In order to construct the PDM model of AAM, a small subset of MRI images is randomly selected as the development set of images. The contour of these images will be labeled manually. One more set of images is also selected randomly and labeled manually, which is used for evaluation. The remaining images will serve as implementation set. The framework was shown in Figure 2.

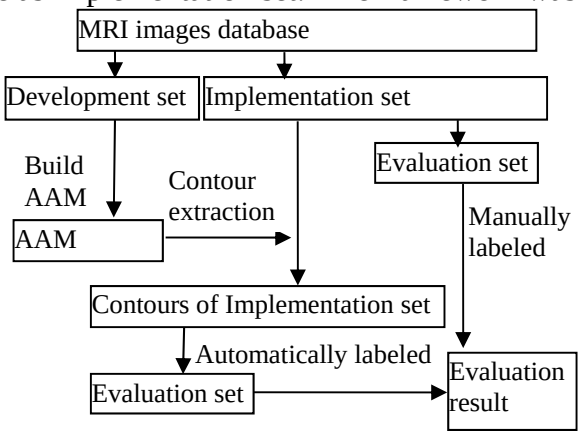


Figure 2. The framework of tracking the tongue contour by AAMs

**Experiments**

In this section, we will describe the details of the experiments of automatically extracting the tongue contours from MRI images. The experiments include the procedure of label the development set of images and evaluation set of images. The features obtained by AAM will be demonstrated also in this section.

**The Region of Interest of the Images**

The image recorded by MRI shown in figure 3 includes not only the objects of articulators but also the other region of the human head such as brain etc. As the illumination of the texture changes frame by frame, the more objects the images have, the more difficulty the contour extraction.

The Region of Interest (ROI) extraction from the raw images is a necessary step, in order to remove the unwilling effects from the other parts of the image. The original images size is 256*256 pixels, in which a region of 128*128 pixels was selected. The selected region was shown in the square with white frame in the Figure 3.

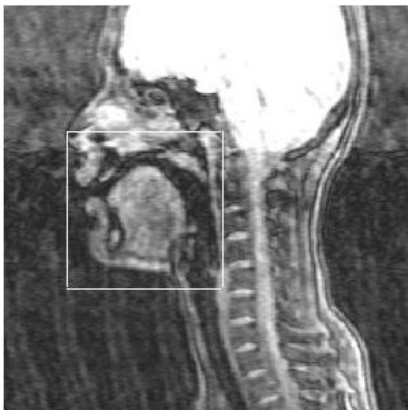

Figure 3. The ROI of the MRI image was located in the square with white frame

## Labeling the Images of Development Set

The image sequence comprises 128 frames, in which 20 images was selected randomly to serve as the development set. The other 20 images were randomly selected from remaining images serving as the evaluation. 8 landmarks were labeled on the upper lip, 18 landmarks on lower lip, 15 landmarks for palate and 21 points around tongue body for each image. There are totally 62 landmarks for one image. Figure 4 is an example of labeled image, where there is a big gap between the tongue and surface skin of the jaw because the bone of the mandible cannot be imaged in MRI. The white dots are the landmarks around each articulator, which were linked by the solid lines. As the articulators are high elastic organs, location of the landmarks varies frame by frame. In our specification, the key points located on corners or some special features that can be easily identified are strictly guaranteed. While the other points lie in between theses key points with order number and roughly equidistance.

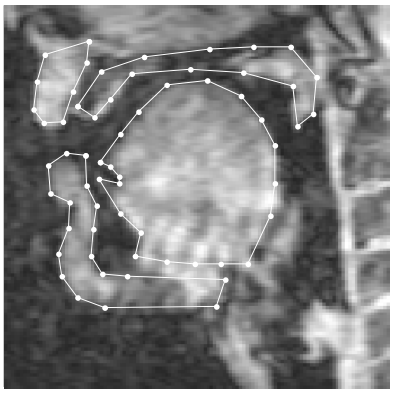

Figure 4. The labeling of tongue contour

## AAM Based Features Demonstration

The AAM of MRI images was constructed based on the labeled images of development set. The shape denoted by the landmarks and the texture information inside the region was used to build the model.

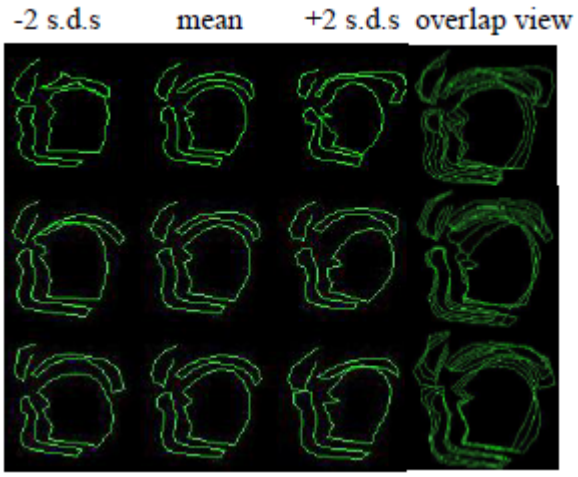


Figure 5. The first 3 modes of the tongue shapes. First row stands for the first mode, second row stands for the second mode, and the last row for the third mode

The outputs of AAM models for each new image are not only the contours but also AAM feature as the coefficient c described in equation 2 of Section 3.1. Figures 5 and 6 show the first 3 modes of shape of the AAM and shape with texture of the AAM, respectively.

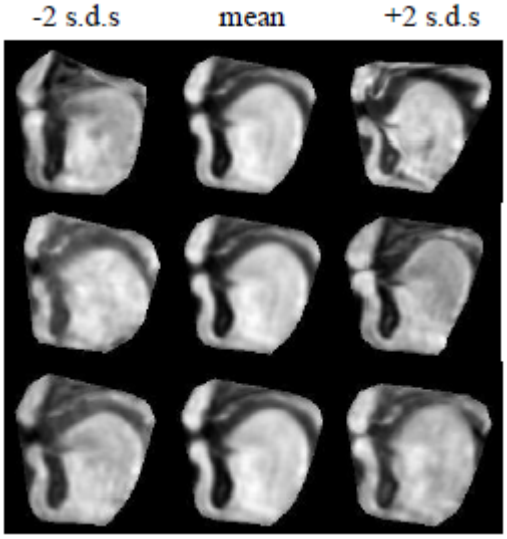


Figure 6. The first 3 modes of the AAM tongue model. First row stands for the first mode, second row stands for the second mode, and the last row for the third mode

## Performances and Evaluation

The proposed method has been applied on unknown images to extract the contours. Four of them are shown in Figure 7. The average Euclidean distances between manually labeled landmarks and automatically extracted by AAM have been calculated over all the landmarks of all the images of test set, which was about 2.1 pixels.

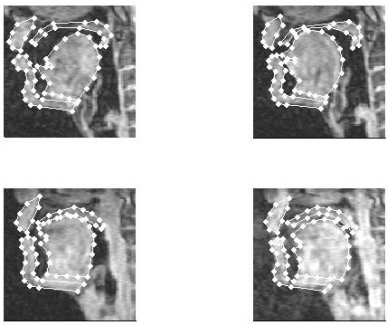

Figure 7. Examples of extracted contours by using AAM method

## Conclusions

The AAM based contour tracking approach was demonstrated, which has shown the ability for fully automatic tracking tongue shape from MRI images. Evaluation of the proposed method showed that the average error over all the landmarks of test set images is around 2.1 pixels. The AAM based tongue contour tracking method is useful for analysis of articulatory movement. Moreover, the AAM features including shape and texture information can serve as a kind of compact articulatory features that can be applied in speech production field in the future.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61175016) and 973 project (No. 2013CB329305), as well as in part of Key Fund projects of NSFC of China (No. 61233009) and projects of China Qinghai provincial science & technology department (No. 2016-ZJ-Y04).

## References

[1] Honda K. "Modeling Vocal Tract Organs Based on MRI and EMG Observations and Its Implication on Brain Function", Ann. Bull. RILP, 27, p.37-49, 1993

[2] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Ninomiya, "MRI-based speech production study using a synchronized sampling method," J. Acoust. Soc. Jpn.(E), vol. 20, pp. 375-379, 1996

[3] Narayanan S., Alwan A., Haker K. , "An articulatory study of fricative consonants using Magnetic Resonance Imaging", J. Acoust. Soc. Am. (JASA), 98(3), p.1325-1347

[4] Stone, M., (2005) "A Guide to Analyzing Tongue Motion from Ultrasound Images."Clinical Linguistics and Phonetics, Sept-Nov, 2005 Volume (19) 6-7 455-502.

[5] Yves Laprie And Berger, M.-O., "Extraction Of Tongue Contours In X-Ray Images With Minimal User Interaction", ICSLP 1996, Volume: 1, pp: 268-271, Philadelphia, PA, USA.

[6] Akgul, Y.S.; Kambhamettu, C.; Stone, M., "Automatic extraction and tracking of the tongue contours", IEEE Transactions Medical Imaging, Volume: 18, Issue: 10, PP: 1035-1045.

[7] T.F.Cootes, G.J. Edwards and C.J.Taylor. "Active Appearance Models", IEEE PAMI, Vol.23, No.6, pp.681-685, 2001.

[8] T.F. Cootes and C.J. Taylor, "Statistical models of appearance for medical image analysis and computer vision", Proc. SPIE Medical Imaging 2001

[9] T.F.Cootes and C.J.Taylor, "An Algorithm for Tuning an Active Appearance Model to New Data", Proc. British Machine Vision Conference, Vol. 3, pp.919-928, 2006

[10] M. B. Stegmann, B. K. Ersbøll, R. Larsen, FAME - A Flexible Appearance Modelling Environment, IEEE Transactions on Medical Imaging, vol. 22(10), pp. 1319-1331.

[11] Steven C. Mitchell, Johan G. Bosch, Boudewijn P. F. Lelieveldt, Rob J. van der Geest, Johan H. C. Reiber, andMilan Sonka, "3-D Active Appearance Models: Segmentation of Cardiac MR and Ultrasound Images", IEEE Transactions on Medical Imaging,VOL. 21, NO. 9, pp. 1167-1178

[12] Gaowu Wang, Jianwu Dang and Jiangping Kong. "Estimation of Vocal Tract Area Function for Mandarin Vowel Sequences Using MRI". Interspeech 2008, pp1182-1185, Brisbane Australia.