SEECCM 2009
2nd South-East European Conference on Computational Mechanics
An IACM-ECCOMAS Special Interest Conference
M. Papadrakakis, M. Kojic, V. Papadopoulos (eds.)
Rhodes, Greece, 22–24 June 2009

ANALYSIS OF TONGUE SHAPE AND MOTION IN SPEECH PRODUCTION USING STATISTICAL MODELING

Maria João M. Vasconcelos[1], Sandra M. Ventura[2], João Manuel R. S. Tavares[1*] and Diamantino Rui S. Freitas[3]

[1] Faculty of Engineering, University of Porto
Lab. Optics and Experimental Mechanics, Inst. Mechanical Eng. and Industrial Management
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
{maria.vasconcelos, tavares}@fe.up.pt

[2] Radiology, School of Allied Health Science – IPP
Faculty of Engineering, University of Porto
Rua Valente Perfeito 322, 4400-330 Vila Nova de Gaia, Portugal
smr@estsp.ipp.pt

[3] Department of Electrical and Computer Engineering,
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
dfreitas@fe.up.pt

Keywords: Biomedical Informatics, Modelling, Statistical Deformable Models, Speech Production, Tongue Shape, Morphological Study.

Abstract. *The mechanisms of speech production are complex and have been raising attention from researchers of both medical and computer vision fields. In the speech production mechanism, the articulator's study is a complex issue, since they have a high level of freedom along this process, namely the tongue, which instigates a problem in its control and observation. In this work it is automatically characterized the tongues shape during the articulation of the oral vowels of Portuguese European by using statistical modeling on MR-images. A point distribution model is built from a set of images collected during artificially sustained articulations of Portuguese European sounds, which can extract the main characteristics of the motion of the tongue. The model built in this work allows understanding more clearly the dynamic speech events involved during sustained articulations. The tongue shape model built can also be useful for speech rehabilitation purposes, specifically to recognize the compensatory movements of the articulators during speech production.*

1 INTRODUCTION

The mechanisms of speech production are complex and have been raising attention from researchers of both medical and computer vision fields. Techniques of medical imaging are being continuously developed to allow a better comprehension of the vocal tract morphology. The results of segmentation and interpretation of objects represented in images through the use of statistical methods has revealed to be interesting and efficient in several areas. Some examples of it can be found in: industry, in industrial inspection [1]; security, in face recognition and automated surveillance of pedestrians [2]; medicine, in the localization and characterization of bones and organs in medical images [3].

The tongue is a large muscular organ covered by mucous membrane, located in the floor of the mouth which is attached by muscles to the hyoid bone, mandible, styloid processes, and pharynx. Besides the key role in gustation, mastication and swallowing, the tongue has an important function in speech production (because it is the articulator with more mobility and flexibility). Its main mass is composed by a set of muscles, which permits the elongation and constriction of the whole tongue or of specific parts allowing the articulation of the sounds. The tongue's structure presents a tip (which usually rests against the incisors) and margin, body, dorsum (with a convex shape which contacts with the palate), inferior surface and root (Figure 1).

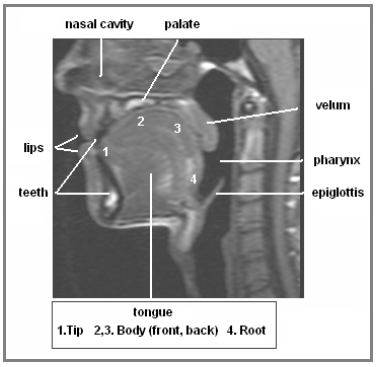

Figure 1: MR sagittal slice demonstrating the vocal tract organs.

The anterior and lateral margins of the tongue are unrestricted, while the inferior surface is connected to the floor of the mouth and to the hyoid bone by the lingual *frenulum* that blocks their movement.

In the speech production mechanism, the study of articulators is a complex issue, since they have a high level of freedom along this process, namely the tongue, which instigates a problem in its control and observation [4]. The use of magnetic resonance images (MRI) permits studying the entire vocal tract with enough safety with the major advantage of being non-invasive [5, 6]. Image quality allied with high soft-tissues resolution are other key advantages of using MRI as allow the analyses the entire vocal tract through the calculi of several descriptive parameters. This technique also allows the collection of a set of static and dynamic images that can represent the entire vocal tract along any orientation.

In this work it is characterized automatically the tongue shape during the articulation of different sounds (oral vowels) by using statistical modeling on MR-images; more specifically, using point distribution models (PDM). Thus, a point distribution model is built from a set of images collected during artificially sustained articulations of Portuguese European sounds, which can extract the main characteristics of the motion of the tongue.

Most medical studies involving PDMs using MRI are related with the localization and characterization of bones and organs. In the field of speech production this is the first work that applies PDMs to characterize tongue shape and motion. The knowledge of speech organs is essential for clinical purposes and also for a better understanding of acoustic theory in speech production.

2 METHODS

2.1 MRI protocol

Image acquisition was performed using a Siemens Magneton Symphony 1.5T system and a head array coil, with subjects lying in supine position. The speech corpus consisted of a set of images collected during sustained articulations of nine European Portuguese sounds. The acquisition of sagittal slices, T1-weighted, was done using Turbo Spin Echo Sequences, with a recording duration approximately 10 sec. and a 150 mm sized field-of-view. This static study was designed to obtain the morphologic data of most of the range of the articulator's positions aiming the imaging characterization of Portuguese sounds. The sagittal data are particularly useful in the study of the whole vocal tract anatomy, demonstrating the main aspects of the shape and positions of some articulators, e.g. tongue, lips and velum.

2.2 PDM

PDM was used in the construction of Active Shape Models (ASMs) and allowed obtaining a model that represented the mean shape of an object, as well as the admissible variations to its shape, starting with a set of images from the object in study [7]. In this work we modeled the tongue shape using the training images presented in Figure 2.

Each shape from the training set was represented by a set of labeled landmark points, which usually represent important zones of the boundary or significant internal locations of the object (Figure 3). The manual process of labeling an object is normally the simplest one; however, this considers the premise that the user has a technical knowledge about the object involved in order to choose the best locations for the landmarks and consequently, be able to mark it correctly in each image of the training set. In the labeling process, sixteen points were defined to characterize the tongue shape:

- Two points in the lingual *frenulum* (anterior and posterior);
- One point in the tongue's tip;
- One point in the tongue's root;
- Seven points along tongue's body;
- Five points along the inferior surface of the tongue.

Full details of the construction of PDMs can be found in [8], but the following gives a brief description. Considering a set of example vectors $\{x_i \in \Re^n\}$, where correspondence is established between the values at each index of x_i, each vector can be rewritten as:

$$x_i = \bar{x} + P_s b_s, \tag{1}$$

where x represents the n landmark points of the new shape of the modeled object, (x_k, y_k) is the position of landmark point k, $\bar{x}$ is the mean position of landmark points, $P_s = (p_{s1} \quad p_{s2} \quad \ldots \quad p_{st})$ is the matrix of the first t modes of variation, p_{si} correspond to the most significant eigenvectors in a Principal Component Analysis of the position variables, and $b_s = (b_{s1} \quad b_{s2} \quad \ldots \quad b_{st})^T$ is a vector of weights for each variation mode of the shape. Each

eigenvector describes the way in which linearly correlated x_{ij} move together over the set, referred to as mode of variation. New examples, not included in the training set, can be generated by manipulating the elements of b. To model objects in two-dimensions, $\{x_i\}$ is constructed using the co-ordinates of descriptive features of each object and the features must correspond to the same landmarks on each object. Given co-ordinates (x_{ij}, y_{ij}) at each feature j of object i, the shape vector is:

$$x_i = (x_{i0}, x_{i1}, \ldots, x_{in-1}, y_{i0}, y_{i1}, \ldots, y_{in-1})^T .$$

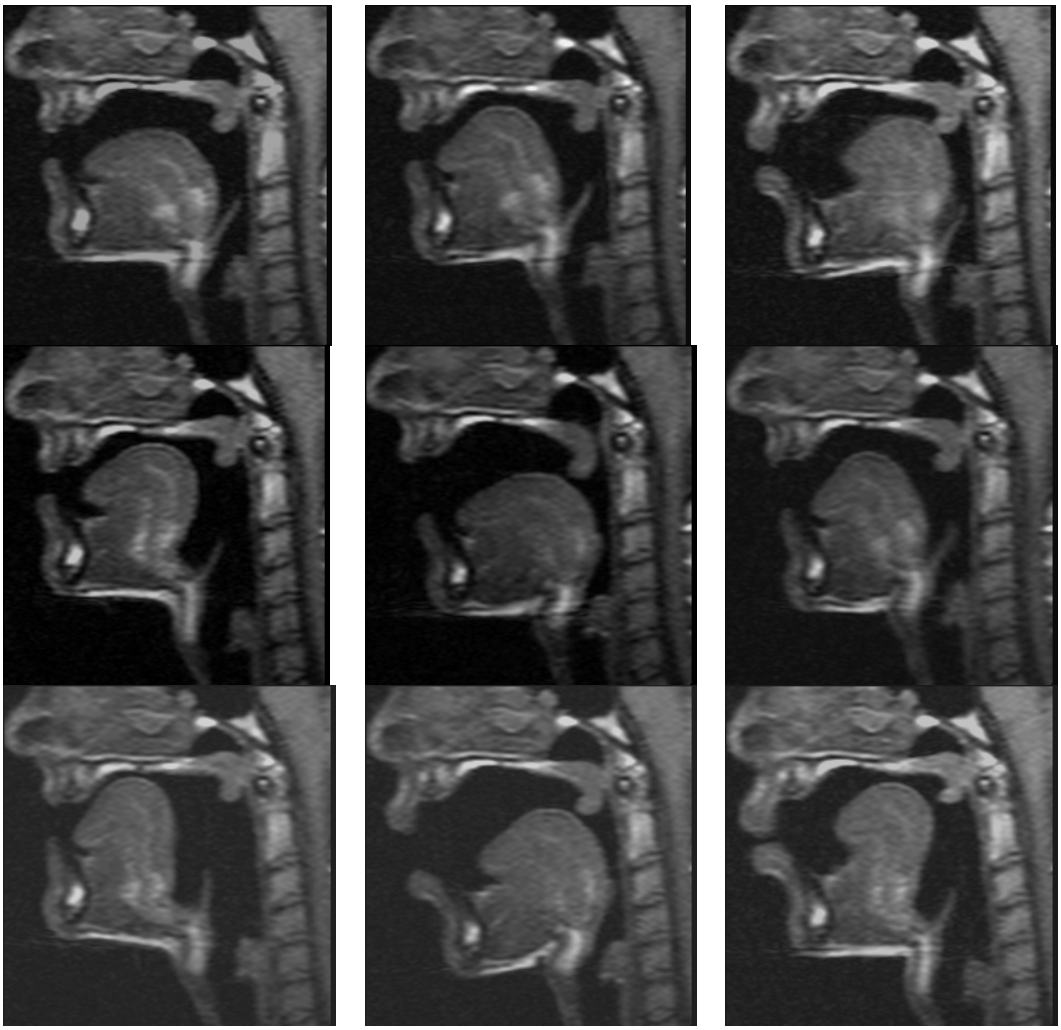

Figure 2: Training images used to build the tongue shape model, showing the different shapes and positions of the lips and the tongue during sustained articulation of the oral vowels.

The combination of PDM and the grey level profiles for each landmark of an object can be used to segment this object in new images through the ASMs. The referred technique is an iterative optimization scheme for PDMs allowing initial estimates of pose, scale and shape of an object to be refined in a new image [9]. The used approach is summarized on the following steps: 1) at each landmark point of the model calculate the necessary movement to displace that point to a better position; 2) calculate changes in the overall position, orientation and scale of the model which best satisfy the displacements; 3) finally, through calculating the re-

quired adjustments to the shape parameters, use residual differences to deform the shape of the model.

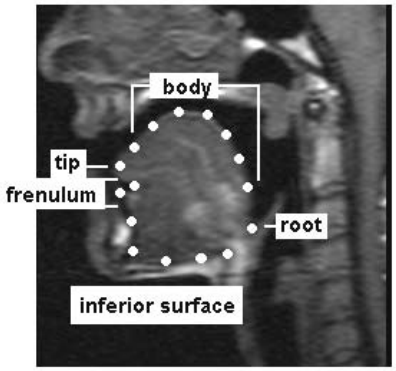

Figure 3: Landmark points considered to build the tongue shape model.

In [10] the authors presented an improved active shape model which uses multiresolution. So, the proposed method first constructs a multiresolution pyramid of the input images by applying a Gaussian mask, and then studies the grey level profiles on the various levels of the pyramid built, making faster and more robust active models.

3 RESULTS

In the sound productions of Portuguese vowels and semi-vowels, the articulators remain sufficiently spaced out allowing air flow to pass freely and almost without obstacles. The main difference between the vocalic sounds (vowels) outcomes from the position of the lips and tongue. Vowels are classified in four different classes: open, close, mid and central, according to the position of the tongue, lip's projection and the mouth aperture. The tonic system of Portuguese European is composed by nine oral vowels:

- Open front unrounded vowel: [a] (from the Portuguese word c*a*sa);
- Mid central unrounded vowel: [ɐ] (from the Portuguese word c*a*da);
- Open-mid front unrounded vowel : [ɛ] (from the Portuguese word p*é*);
- Close-mid front unrounded vowel: [e] (from the Portuguese word m*e*do);
- Open-mid back rounded vowel: [ɔ] (from the Portuguese word p*ó*);
- Close-mid back rounded vowel: [o] (from the Portuguese word f*o*rça);
- Close front unrounded vowel: [i] (from the Portuguese word v*i*);
- Close back rounded vowel: [u] (from the Portuguese word t*u*);
- Mid central unrounded vowel: [ɨ] (from the Portuguese word sed*e*).

So, the tongue moves from front-high positions to a central-low position on the oral cavity for the vowels [i, e, ɛ, a] and from this position to back-high positions for the vowels [ɐ, ɔ, o, u], respectively.

We developed an application in MATLAB to build ASMs, using the *Active Shape Models software* [11] as basis. For the construction of the model of the tongue shape it was used a training set of MRI images collected during artificially sustained articulations of Portuguese

sounds, Figure 2. The images collected are from one young male subject and without speech disorders.

From Table 1 we can observe that the first three modes of the shape model built could explain 90% of all shape variance of the tongue. The first five modes explain 95% of all shape variance and with only seven modes of variation it is possible to explain 99% of all shape variance of the tongue.

Table 1: First seven modes of variation of the model obtained and their retained percents.

Mode	**Retained %**	**Cumulative Retained %**
λ_1	56.453%	56.453%
λ_2	23.362%	79.815%
λ_3	10.623%	90.438%
λ_4	3.331%	93.769%
λ_5	2.454%	96.223%
λ_6	1.787%	98.010%
λ_7	1.378%	99.388%

The effects of varying the first four modes of variation are visible in Figure 4. The first mode is associated to movements of the whole tongue along the vertical to horizontal direction. In the second mode of variation, it is possible to observe the rise of the inferior surface and of the tongue body towards the palate. The third mode of variation translates the lowering of the tongue's tip and the advance of the tongue body simultaneously. The fourth mode of variation translates the rise and backward of the tongue dorsum. The fifth mode is related with the vertical rise of the tongue body towards the palate. The sixth mode translates the backward of the tongue's tip and finally the seventh mode is related with the diagonal movement of the whole tongue from high to lower positions.

4 CONCLUSIONS

In this work we applied Point Distribution Models in Magnetic Resonance Images to analyze the tongue's shape in the articulation of some Portuguese European sounds.

So far, most medical studies that involve PDMs using MRI are usually related with the localization and characterization of bones and organs in medical images. In the field of speech production, this is the first work that applies PDMs to characterize the tongue's shape. The sounds used to build tongue shape models represent all the oral vowel sounds of the Portuguese European.

We can conclude that the model built in this work allows understanding more clearly the dynamic speech events involved during sustained articulations. The tongue shape model can also be useful for speech rehabilitation purposes; namely, to recognize the compensatory movements of the articulators during speech production.

The data collected and analyzed can contribute to the construction of 3D articulatory models for speech synthesis. It can also be used for several studies in articulatory phonetics and in the vocal tract modeling for speech synthesis, with applications to speech pathology, linguistics and artificial speech.

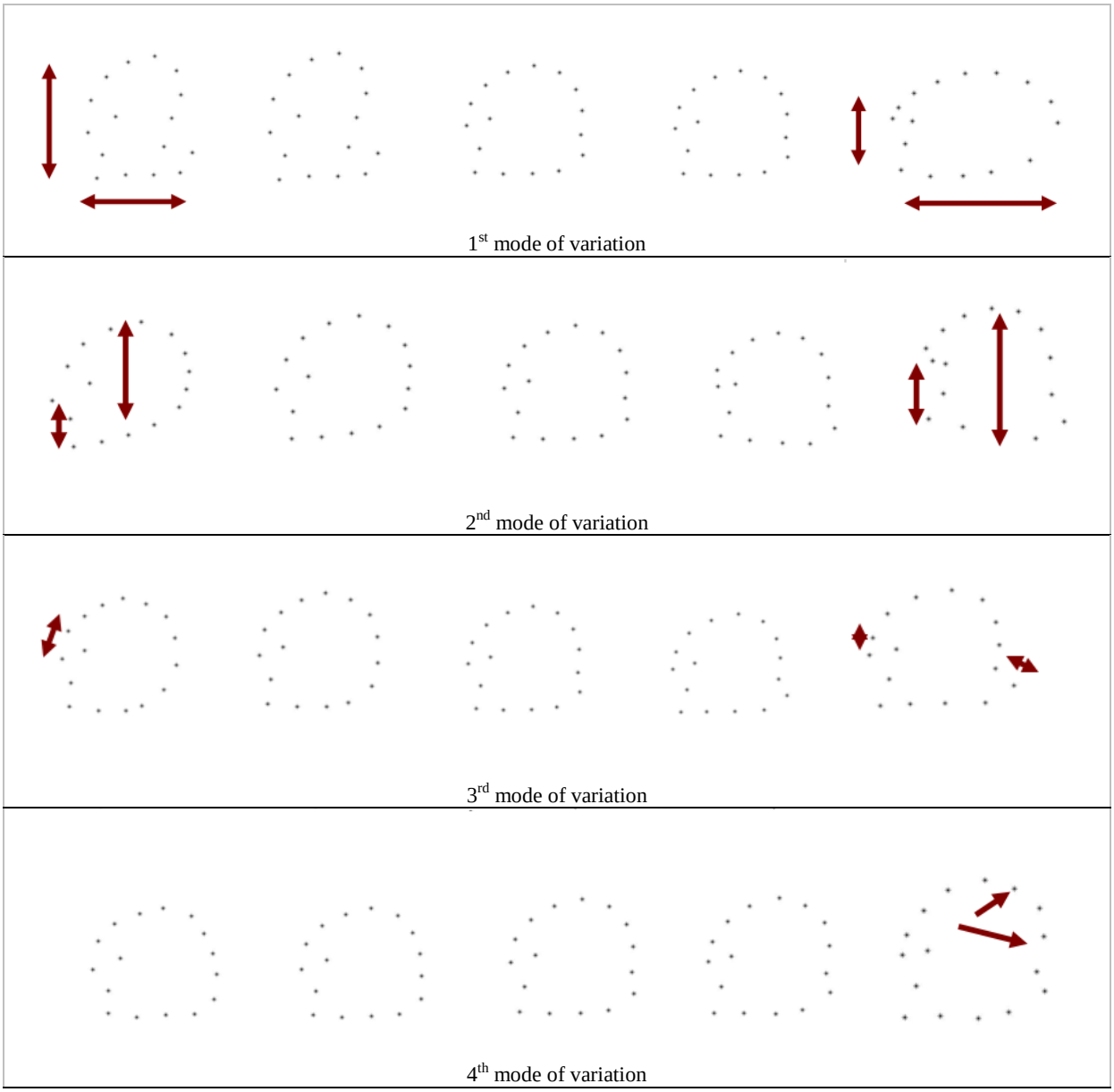

Figure 4: Effects of varying each of the first five modes of variation of the tongue model $(\pm 2sd)$.

5 ACKNOWLEGDMENTS

The first author would like to thank the support of the PhD grant SFRH/BD/28817/2006 from FCT – *Fundação para a Ciência e Tecnologia* from Portugal.

Images were acquired at the Radiology Department of the Hospital S. João, Porto, with the collaboration of Isabel Ramos (Professor of Faculdade de Medicina da Universidade do Porto and Department Director) and the technical staff, which is gratefully acknowledged.

REFERENCES

[1] F.-C. Tien, C.-H. Yeh, et al., Automated visual inspection for microdrills in printed circuit board production, *International Journal of Production Research*, Vol. 42, No. 12, pp. 2477-2495, 2004.

[2] A. Lanitis, C. J. Taylor, et al., An Automatic Face Identification System Using Flexible Appearance Models, *Image and Vision Computing*, Vol. 13, No. 5, pp. 392-401, 1995.

[3] J. Rijsdam, An automatic left-ventricular search and fitting method using a three-dimensional point distribution model, MSc thesis, Leiden University, Leiden, 1999.

[4] S.R. Ventura, D.R. Freitas, J.M.R.S. Tavares, Application of MRI and Biomedical Engineering in Speech Production Study, *Computer Methods in Biomechanics and Biomedical Engineering* (in press), 2008.

[5] M. S. Avila-Garcia, J. N. Carter, et al., Extracting Tongue Shape Dynamics from Magnetic Resonance Image Sequences, *Transactions on Engineering, Computing and Technology*, Vol. 2, December, 2004.

[6] O. Engwall, A revisit to the Application of MRI to the Analysis of Speech Production - Testing our assumptions, *6th International Seminar on Speech Production*, Sydney, Australia, 2003.

[7] M.J.M. Vasconcelos, J.M.R.S. Tavares, Methods to Automatically Built Point Distribution Models for Objects like Hand Palms and Faces Represented in Images, *Computer Modeling in Engineering & Sciences (CMES)*, Tech Science Press, ISSN: 1526-1492 (print) - 1526-1506 (online), vol. 36, no. 3, pp. 213-241, 2008.

[8] T. F. Cootes, C. J. Taylor, et al., Training Models of Shape from Sets of Examples, *Proceedings of the British Machine Vision Conference*, Leeds, 1992.

[9] T. F. Cootes and C. J. Taylor, Active Shape Models - 'Smart Snakes', *Proceedings of the British Machine Vision Conference*, Leeds, 1992.

[10] T. F. Cootes, C. J. Taylor, et al., Active Shape Models: Evaluation of a Multi-Resolution Method for Improving Image Search, *British Machine Vision Conference*, York, England, BMVA, 1994.

[11] G. Hamarneh, ASM (MATLAB), available for download from http://www.cs.sfu.ca/~hamarneh/software/asm/index.html, last accessed in February 2009.