A PARAMETRIC THREE-DIMENSIONAL MODEL OF THE VOCAL-TRACT BASED ON MRI DATA

Hani Yehia *Mark Tiede*

yehia@hip.atr.co.jp tiede@hip.atr.co.jp

ATR Human Information Research Laboratories

2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

In this paper, 24 three-dimensional (3D) vocal-tract (VT) shapes extracted from MRI data are used to derive a parametric model for the vocal-tract. The method is as follows: first, each 3D VT shape is sampled using a semi-cylindrical grid whose position is determined by reference points based on VT anatomy. After that, the VT projections onto each plane of the grid are represented by their two main components obtained via principal component analysis (PCA). PCA is once again used to parametrize the sequences of coefficients that represent the sections along the tract. It was verified that the first four components can explain about 90% of the total variance of the observed shapes. Following this procedure, 3D VT shapes are approximated by linear combinations of four 3D basis functions. Finally, it is shown that the four parameters of the model can be estimated from VT midsagittal profiles.

1. INTRODUCTION

Vocal-tract (VT) models play important roles in the investigation of articulatory speech processes. One such role is the representation of VT shapes by a reduced set of parameters permitting good description of the behavior of the articulators during the speech production process. The inverse function is also important: given a set of parameters, the VT model should allow reconstruction of the tract shape from which the parameters were extracted.

Vocal-tract models currently in use[1, 2, 3] were developed from vocal-tract imaging techniques limited to two dimensions, effectively restricting the acquisition of VT shapes to midsagittal profiles. For this reason such profiles constitute the focus of attention of these models. The relation with the basic acoustic properties of the tract is implemented via the area function (i.e. the cross-sectional area along the tract). The area function works as "bridge" between geometric and acoustic characteristics of the vocal-tract: from the acoustic point of view, it is possible to obtain synthesized speech of good quality from the area function[4, 5]; whereas from the geometric point of view, it is possible to estimate the area function from the midsagittal distances between anterior and posterior tract walls[6, 7, 8, 9, 10].

There exist, however, two sources of inaccuracy in this approach. The first one is acoustic: the area function alone is not enough to model nonlinear sound generation as well as transverse modes of propagation. This is a problem for sounds other than vowels, and for frequencies above around 4 kHz. The second source of inaccuracy is morphological: the area function, although highly dependent on the midsagittal distances, is not fully determined by them.

With the advent of three-dimensional (3D) imaging techniques[11], the information necessary to derive a 3D model of the vocal-tract has become available. In this paper, a principal component analysis (PCA), similar to that described in [3] for the two-dimensional case, is carried out for the three-dimensional case. By doing so, given a reduced set of parameters, it is possible to recover (with some loss of information) the 3D VT shape rather than only the midsagittal profile as in the classical VT models[1, 2, 3]. With this approach, the area function can be computed directly, without the need of a transformation from the midsagittal profile. Furthermore, more elaborate acoustic models can be implemented with the cross-sectional shapes available.

In the model proposed here, each 3D VT shape is approximated by a linear combination of four basis functions. These basis functions are derived from the corpus of MRI data available and, once determined, allow the reconstruction of the 3D VT shape from the four weights used in the linear combination of basis functions. These weights are the parameters of the 3D VT model. In the last section, it is shown that it is possible to estimate these parameters from the partial information contained in the midsagittal profile of the vocal-tract (e.g. obtained by cineradiography[12]).

2. CORPUS

The initial corpus consists of 24 volumes obtained by Magnetic Resonance Imaging (MRI) from a male Brazilian Portuguese native speaker (HY) for the sustained vowel articulations **/a/**, **/eh/**, **/ey/**, **/iy/**, **/ao/**, **/ow/**, **/uw/**, **/an/**, **/en/**, **/in/**, **/on/**, and **/un/**. There are two volumes for each articulation, one scanned in transverse and the other in coronal orientation. The procedure for sampling MRI data into appropriate sets of points defining cross-section areas is as follows (see [13] for more details): for each volume a semi-polar grid is established on its midsagittal projection using four anatomically based reference points. After that the MRI volumes area sampled along the grid planes (perpendicular to the midsagittal plane), resulting in a 'semi-cylindrical' unwrapping of the vocal-tract. A thresholding function is then applied to obtain the air/tissue boundaries. Finally, the points defining each cross-section are resampled so that each section is given by $M = 128$ points. The sections themselves are also resampled so that the resulting 3D VT shape is defined by $N = 32$ sections. After these

operations, the processed corpus consists of $P = 24$ 3D VT shapes represented by $N = 32$ sections, each of them containing $M = 128$ points.

3. MATHEMATICAL PROCEDURE

Mathematically, each volume is represented by a matrix

$$\mathbf{C}_p = \left[\begin{array}{ccc} \mathbf{c}_{p1} & \cdots & \mathbf{c}_{pN} \end{array} \right] = \left[\begin{array}{ccc} c_{p\,1\,1} & \cdots & c_{p\,1\,N} \\ \vdots & & \vdots \\ c_{p\,2M\,1} & \cdots & c_{p\,2M\,N} \end{array} \right]; \quad (1)$$
$$p = 1, \ldots, P; \quad P = 24; \quad N = 32; \quad M = 128.$$

(Each of the columns of $\mathbf{C}_p$ has $2M = 256$ entries due to the fact that the points contained in the plane of each section are determined by two Cartesian coordinates.) Principal component analysis (PCA) is performed for each section as follows: the covariance matrix of section n is estimated by

$$\mathbf{R}_n = \frac{1}{P-1} \sum_{p=1}^{P} (\mathbf{c}_{pn} - \mu_n)(\mathbf{c}_{pn} - \mu_n)^T; \quad (2)$$
$$\mu_n = \frac{1}{P} \sum_{p=1}^{P} \mathbf{c}_{pn} \quad n = 1, \ldots, N;$$

where μ_n is the mean vector of $\mathbf{c}_{pn}$ and T denotes transpose. Singular Value Decomposition (SVD) is then used to express $\mathbf{R}_n$ as

$$\mathbf{R}_n = \mathbf{U}_n \mathbf{S}_n \mathbf{U}_n^T; \quad (3)$$

where $\mathbf{S}_n$ is a diagonal matrix whose (diagonal) entries are the eigenvalues of $\mathbf{R}_n$. The associated eigenvectors, normalized to unit Euclidean norm, are the columns of $\mathbf{U}_n$. It was verified that the first $K = 2$ eigenvectors of $\mathbf{R}_n$ account for, on average, about 90% of the total variance of $\mathbf{R}_n$, so that the shape of each cross-section can be well approximated as a linear combination of the $K = 2$ eigenvectors associated with the $K = 2$ largest eigenvalues of $\mathbf{R}_n$

$$\mathbf{c}_{pn} \simeq \mathbf{U}_{nK} \mathbf{d}_{pn} + \mu_n; \quad (4)$$

where $\mathbf{U}_{nK}$ is the matrix formed by the first two columns of $\mathbf{U}_n$, and $\mathbf{d}_{pn}$ is a $K = 2$–dimensional vector given by the projections of $\mathbf{c}_{pn}$ onto the directions defined by the eigenvectors contained in $\mathbf{U}_{nK}$

$$\mathbf{d}_{pn} = \mathbf{U}_{nK}^T (\mathbf{c}_{pn} - \mu_n), \quad \mathbf{d}_{pn} = \left[\begin{array}{c} d_{pn1} \\ d_{pn2} \end{array} \right]. \quad (5)$$

Since the shape of one section of the tract is not (in principle) independent of the shapes of the other sections, considerable correlation can be expected among the $N = 32$ vectors $\mathbf{d}_{pn}$ that represent a given volume $\mathbf{C}_p$. To take advantage of this fact, PCA is once again carried out in the following way: the vectors that parametrize the volumes of the corpus are grouped to form $KN = 64$–dimensional vectors

$$\mathbf{D}_p = [d_{p11} d_{p12} \ldots d_{pN1} d_{pN2}]^T; \quad p = 1, \ldots, P; \quad (6)$$

whose covariance matrix is estimated by

$$\mathbf{R}_D = \frac{1}{P-1} \sum_{p=1}^{P} (\mathbf{D}_p - \mu_D)(\mathbf{D}_p - \mu_D)^T; \quad (7)$$

$$\mu_D = \frac{1}{P} \sum_{p=1}^{P} \mathbf{D}_p; \quad (8)$$

where μ_D is the mean vector of $\mathbf{D}_p$. SVD is now used to express $\mathbf{R}_D$ as

$$\mathbf{R}_D = \mathbf{V}\mathbf{W}\mathbf{V}^T; \quad (9)$$

where $\mathbf{W}$ is a diagonal matrix whose (diagonal) entries are the eigenvalues of $\mathbf{R}_D$; and the columns of $\mathbf{W}$ contain the associated normalized eigenvectors. For the analyzed corpus, it was verified that the first $L = 4$ eigenvectors of $\mathbf{R}_D$ account for about 90% of the total variance of $\mathbf{R}_D$. Thus, each vocal-tract volume can be well approximated as a linear combination of these $L = 4$ eigenvectors

$$\mathbf{D}_p \simeq \mathbf{V}_L \mathbf{e}_p + \mu_D; \quad (10)$$

where $\mathbf{V}_L$ is a matrix whose columns contain the $L = 4$ eigenvectors associated with the $L = 4$ largest eigenvalues of $\mathbf{R}_D$; and $\mathbf{e}_p$ is a vector whose components are the projections of $\mathbf{D}_p$ onto these eigenvectors

$$\mathbf{e}_p = \mathbf{V}_L^T (\mathbf{D}_p - \mu_D). \quad (11)$$

Each vocal-tract volume $\mathbf{C}_p$ is now parametrized by the $L = 4$ coefficients of vector $\mathbf{e}_p$. These parameters, together with the matrices of eigenvectors $\mathbf{V}_L$ and $\mathbf{U}_n$, and the mean vectors μ_D and μ_n, can be used to recover the vocal-tract volume by means of equations (10), (6) and (4).

As a final point, it was observed that, for a given volume $\mathbf{C}_p$ it is possible to estimate the vector of parameters $\mathbf{e}_p$ as a linear transformation of a vector of midsagittal points

$$\mathbf{e}_p \simeq \mathbf{A}\mathbf{y}_p; \quad (12)$$

where $\mathbf{y}_p$ is a vector of points along the midsagittal profile of volume $\mathbf{C}_p$, and $\mathbf{A}$ is the linear transformation that minimizes the mean squared error

$$\frac{1}{LP} \sum_{p=1}^{P} (\mathbf{e}_p - \mathbf{A}\mathbf{y}_p)^T (\mathbf{e}_p - \mathbf{A}\mathbf{y}_p); \quad (13)$$

which is given by

$$\mathbf{A} = \mathbf{E}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}; \quad (14)$$
$$\mathbf{E} = [\mathbf{e}_1 \ldots \mathbf{e}_P]; \quad \mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_P]. \quad (15)$$

4. DISCUSSION AND RESULTS

In this section some aspects of the vocal-tract model described are briefly analyzed. The first point is the amount of information contained in the K eigenvectors used as basis vectors (basis functions in the continuous case) in the representation of each section of the vocal-tract. The bottom panel of Fig. 1 shows the cumulative percentage of the variance of the sections along the tract explained by the first four eigenvectors of the corresponding covariance matrix ($\mathbf{R}_n$). The solid black line shows the case of $K = 2$

adopted in the model. The top panel of Fig. 1 shows the mean variance of each section. Note that the regions where it is necessary to have more components to explain a high percentage of the total variance are regions where the total variance is comparatively small.

Considering now the three-dimensional case, the four eigenvectors contained in the columns of $\mathbf{V}_L$ can be seen as weights which, when applied to the basis vectors of the sections along the tract, represent four three-dimensional basis functions in a discrete form. The influence of the first two basis functions on the shape of the vocal-tract is shown in Fig. 2. These two basis functions account for about 80% of the total variance of the 3D VT shapes contained in the corpus, whit the other two basis functions accounting for an additional 10%.

Figure 3 shows an example where the cross-sectional areas of the coronally scanned volume of vowel /iy/are estimated from points along the midsagittal profile using equations (12) and (14). The black lines represent the original cross-sectional shapes extracted from MRI data, while the gray lines represent the shapes estimated from the midsagittal profile.

Figure 4 shows the area functions obtained from the original cross-sectional shapes extracted from MRI(black lines) compared with those estimated from points in the midsagittal profile (gray lines).

5. CONCLUSION AND COMMENTS

A PCA based method to represent 3D VT shapes as a linear combination of four basis functions was presented. The fixed set of basis functions was derived from a corpus of 24 MRI 3D scans. The weights of the linear combination are the parameters of the model. It was also shown that these parameters can be estimated, with acceptable accuracy, from the partial information contained in the midsagittal profile.

An aspect of the model currently being explored is the possibility of estimating the 3D model parameters from the position of a set of pellets (e.g. obtained by X-ray microbeam[14, 15] or EMMA[16]), midsagittally placed on the lips and articulators of the oral cavity, combined with the acoustic constraints determined by the formant frequencies extracted from the speech signal[17].

The acoustic modeling and articulatory characterization of the 3D VT model proposed here are important points to be explored in the future.

REFERENCES

[1] C. Coker and O. Fujimura. Model for specification of the vocal tract area function. *The Journal of the Acoustical Society of America*, 40:1271, 1966.

[2] P. Mermelstein. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, 1973.

[3] S. Maeda. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 131–149. Kluwer Academic Publishers, 1990.

[4] S. Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3–4):199–229, 1982.

[5] M. M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):955–967, 1987.

[6] J. M. Heinz and K. N. Stevens. On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *The Journal of the Acoustical Society of America*, 36:1037, 1964.

[7] S. Maeda. On the conversion of x-ray data into formant frequencies. Technical report, Bell Laboratories, Murray Hill, N.J., 1972.

[8] J. Sundberg, C. Johanson, H. Widbrand, and C. Ytterbergh. From sagittal distance to area: a study of transverse, vocal-tract cross-sectional area. *Phonetica*, 44:76–90, 1987.

[9] P. Perrier, L. J. Boe, and R. Sock. Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modelling the transition with two sets of coefficients. *Journal of Speech and Hearing Research*, 35:53–67, 1992.

[10] D. Beautemps, P. Badin, and R. Laboissière. Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. *Speech Communication*, 16:27–47, 1995.

[11] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 90(2):799–828, 1991.

[12] A. Bothorel, P. Simon, F. Wioland, and J. P. Zerling. *Cinéradiographie de voyelles et consonnes du Français.* Institut de Phonétique de Strasburg, 1986.

[13] M. K. Tiede and H. Yehia. A shape-based approach to vocal tract area function estimation. In *Proceedings of the 1996 Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, 1996.

[14] O. Fujimura, S. Kiritani, and H. Ishida. Computer controlled radiography for observation of movements of articulatory and other human organs. *Computers in Biology and Medicine*, 3:371–384, 1973.

[15] J. R. Westburry. The significance and measurement of head position during speech production experiments using the x-ray microbeam system. *The Journal of the Acoustical Society of America*, 89(4):1782–1791, 1991.

[16] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078–3096, 1992.

[17] H. Yehia, M. K. Tiede, E. Vatikiotis-Bateson, and F. Itakura. Applying morphological constraints to estimate three-dimensional vocal-tract shapes from partial profile and acoustic information. In *Proceedings of the 1996 Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, 1996.

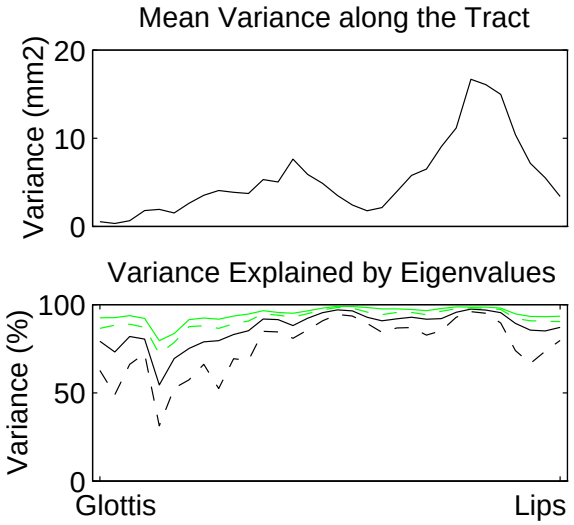

Figure 1. Top: Mean variance (in mm^2) of the sections along the vocal-tract. Bottom: Percentage of the variance of each section explained by the first (dashed black line), first two (solid black line), first three (dashed gray line), and first four (solid gray line) eigenvectors of the corresponding covariance matrix.

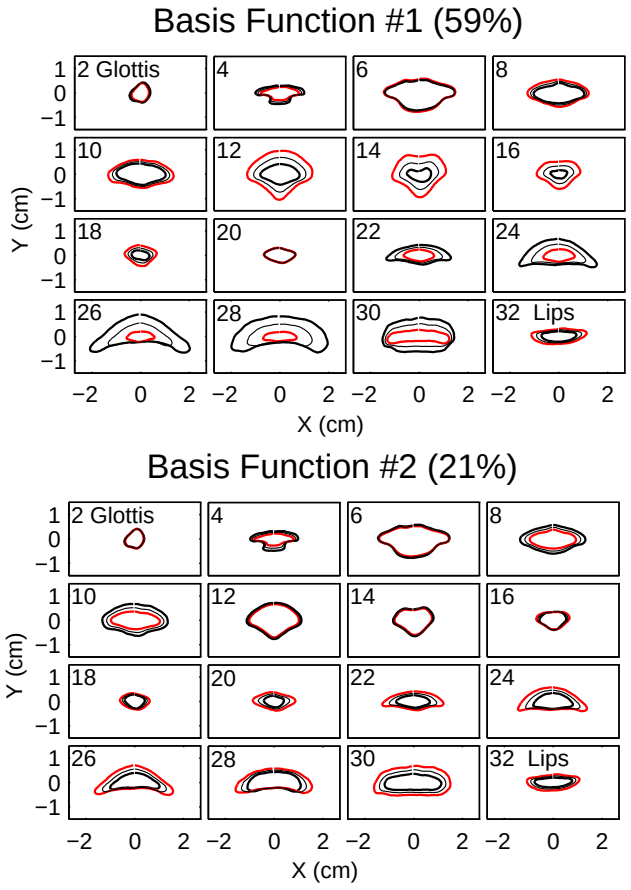

Figure 2. Influence of the first two basis functions on the shape of the vocal-tract. The thin black lines show the mean shape of the tract, overlain by thick black/gray lines showing plus/minus one standard deviation of the distribution observed in the corpus.

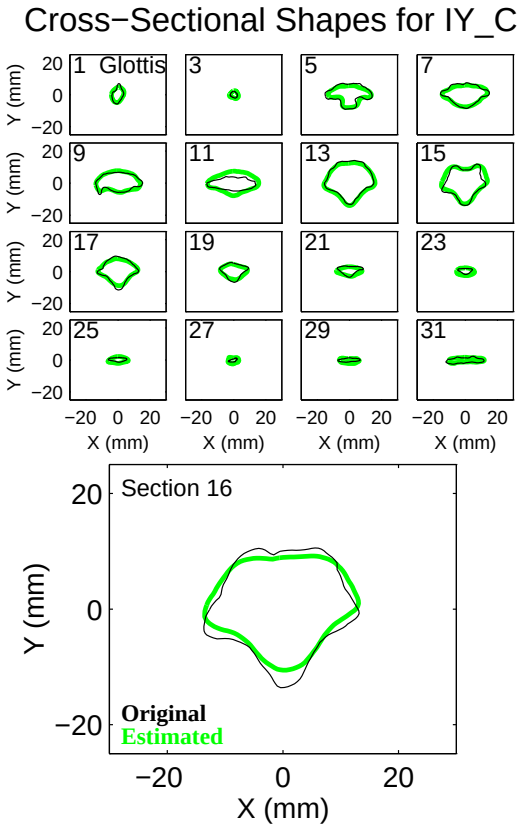

Figure 3. Cross-sectional shapes for the vowel /iy/coronally scanned. Black lines: extracted from MRI. Gray lines: estimated from midsagittal profile.

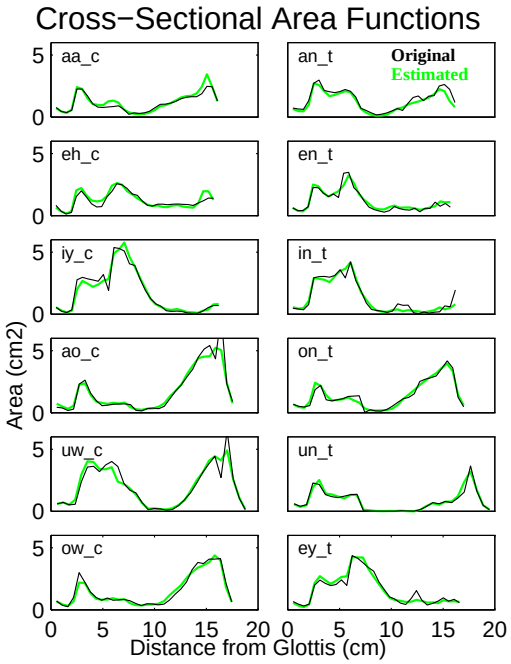

Figure 4. Area functions of some analyzed volumes. Black lines: obtained from MRI derived shapes. Gray lines: obtained from 3D VT model parameters which were estimated from midsagittal profiles.