INTERSPEECH 2009 BRIGHTON

Articulatory Modeling Based on Semi-polar Coordinates and Guided PCA Technique

Jun Cai[1], *Yves Laprie*[1], *Julie Busset*[1], *Fabrice Hirsch*[2]

[1] Groupe Parole, LORIA-CNRS & INRIA, BP 239, 54600 Vandoeuvre-lès-Nancy, France
[2] Institut de Phonétique de Strasbourg, 2, rue Descartes, 67084 Strasbourg, France
{Jun.Cai, Yves.Laprie, Julie.Busset@loria.fr}, fabrice_hirsch@yahoo.fr

Abstract

Research on 2-dimensional static articulatory modeling has been performed by using the semi-polar system and the guided PCA analysis of lateral X-ray images of vocal tract. The density of the grid lines in the semi-polar system has been increased to have a better descriptive precision. New parameters have been introduced to describe the movements of tongue apex. An extra feature, the tongue root, has been extracted as one of the elementary factors in order to improve the precision of tongue model. New methods still remain to be developed for describing the movements of tongue apex.

Index Terms: articulatory modeling, semi-polar coordinate system, guided PCA

1. Introduction

Articulatory models are usually used to transform articulatory parameters to an estimated geometric shape of the vocal tract from which the cross-sectional area function of the vocal tract can be specified and acoustic characteristics can be determined [1]. In creating such models, the concern is mainly geometric accuracy. Various articulatory models have been developed which can be classified into static and dynamic categories [2]. Also, the models can be implemented in either 2- dimensional or 3-dimensional space. In this paper, we present a research on 2-dimensional static articulatory modeling based on statistical analysis of lateral X-ray images of vocal tract. The methods for measuring articulatory parameters have been researched, as well as the methods for both feature extraction and articulatory modeling.

A technique for articulatory modeling based on statistical analysis of midsagittal vocal tract X-ray images was originally developed by S. Maeda [3-5] and was extended by others. Maeda proposed a semi-polar coordinate system for measuring the midsagittal outlines of vocal tract. Based on it, an articulatory model of tongue which was composed of linear factors was determined by a statistical analysis of tongue shapes and jaw-opening measured on the images. The combination of these linear factors could adequately describe the tongue shapes which had been observed during the utterances of 12 French vowels in continuous speech sentences and in certain disyllables.

We try to extend Maeda's techniques for articulatory modeling. A modified semi-polar coordinate system has been developed in order to increase the precision of the models. Two parameters have been introduced to represent the position of the tongue apex. A software toolbox, XArticulators 2.0, has been developed, aiming at measuring the shapes of vocal tract automatically. The guided PCA technique has been used to extract features and to build linear articulatory models. Furthermore, non-linear articulatory models have been built by using ANNs and the precision of the linear models has been compared with that of the non-linear models.

The X-ray image sequence is described in Sec. 2. In Sec. 3, the design of the semi-polar system is reviewed. The proposed improvements to articulatory measurement are presented in detail. The design of the software toolbox is also described. Sec. 4 presents the guided PCA technique for extracting features and for building linear articulatory models. In Sec. 5, the performance of the linear articulatory models are evaluated and compared with the ANN non-linear models in terms of modeling precision. It is concluded in Sec. 6 that though the proposed improvements are effective to represent the movements of vocal tract, some new methods still remain to be developed for describing the movements of tongue apex.

2. The X-ray speech production corpus

Two French speech sequences with fricative consonant [s] and stop consonants [k] and [t] were designed with the objective to investigate the coarticulation in speech. Since the tongue plays the major role in pronouncing these three consonants, the two speech sequences can be used to analyze the coarticulation due to the movement and the kinetic constraints of the tongue. Both of the sequences were uttered twice by an adult native male speaker (Speaker FH), at a normal speech rate for one time while at a faster rate for the other. Lateral X-ray images were recorded to track the midsagittal profile of the vocal tract movements of the speaker whose head was maintained in a fixed position relative to the camera. The X-ray images were taken with a resolution of 0.05cm/pixel and with a frequency of 50Hz, forming an image sequence of 672 consecutive frames. The vocal tract outlines have been marked manually in yellow contours by using XArticulators 2.0, while the lip contours have been generated quasi-automatically.

3. A modified semi-polar system for articulatory measurement

To be measured, the vocal tract is usually divided into three sections: the rear section of the laryngeal extreme, the principal section corresponding to the pharynx and the posterior part of buccal cavity, and the frontal section of the anterior part of buccal cavity. A reference system is required for the measurement. The position of the lower jaw should be also measured for each frame.

3.1. The semi-polar system

The semi-polar coordinate system, which acts as the reference system for measuring and reconstructing the outlines of vocal tract, consists of a polar region for the principal section, and two linear regions for the rear and frontal sections,

respectively. Because of the variety of speakers, the semi-polar system needs to be adapted before being applied to a particular speaker. Also, it must be positioned properly. Though some researchers proposed to determine the coordinate origin in relative to some anatomical landmarks that are visible on the images [6], the positioning of the coordinate origin is somewhat arbitrary. However, since the tongue is the major articulator being investigated, the tongue contour must always intersect with each grid line once and only once, so that there is no hole in the vector representations for all the tongue outlines in a data set. We propose to locate the coordinate origin beneath and close to the second molar tooth in the image when the vocal tract is in a natural state before speech is uttered. The position of semi-polar system must be fixed to some motionless part around the vocal tract. The rigid maxilla is usually chosen to act as the fixed reference. In XArticulators 2.0, a software module has been implemented based on maximizing the correlation between pixels in consecutive X-ray images to track the area of rigid maxilla so as to maintain the fixed position of the coordinate origin to it.

3.2. Representation of vocal tract profiles

The marked contours of the internal vocal tract are measured by using the semi-polar coordinate system, as shown by the red points in Figure 1. The coordinates at these intersection points form the representation vector of the internal vocal tract, and the dimension of the representation vector depends on the number of grid lines. The intersection points are searched automatically.

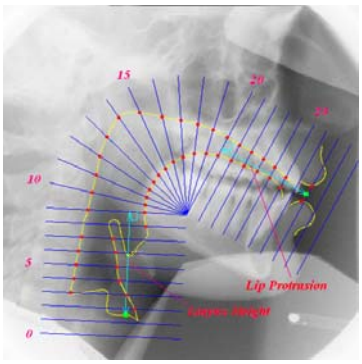

Figure 1: *Semi-polar system and measured parameters*

For measuring the jaw opening, a fully automatic procedure has been developed in XArticulators 2.0 to track the movements of the upper and lower incisors in consecutive X-ray images based on maximizing the correlation between pixels of the incisor areas in consecutive image frames. The lips are flesh parts; they can change their shapes in speech production. Therefore, the correlation-based technique which attempts to maximize the correlation between pixels of the lip areas in consecutive frames is not valid for tracking the lip movements through the whole image sequence. This correlation-based technique is useful, however, to track the movements of lips in a short time span because the shapes and positions of both lips normally do not change significantly within a short period. Therefore, in XArticulators 2.0, for every five consecutive frames in the sequence, the contours of both lips in the first frame are marked manually. Then, the correlation-based technique is used to track the movement of the lips in the following four frames and thus the lip contours are generated automatically for these four frames. This short-span quasi-automatic tracking works well for many image frames, but sometimes the generated lip contours should be corrected manually because the automatic method fails to track the abrupt changes of lip shapes.

In the regions of lip-opening and larynx, the semi-polar system is not useful for measuring the data because the articulators in these regions tend to move perpendicularly to the grid lines. In our research, the lip protrusion and the larynx height are measured as shown in Figure 1. The jaw parameter, *J*, is defined as the projection of the distance between the upper and lower central incisors on the last grid line. This definition is plausible because the grid lines are usually tuned to be perpendicular to the propagation direction of the acoustic wave, and in most cases the last grid line is almost the projecting line which maximizes the contribution of the jaw parameter to the variance of the tongue contour data.

3.3. The modified semi-polar system

The semi-polar system fails to take into account the subtle yet wide variety of possible movements of the tongue apex in forming the sounds of speech, especially in forming consonants. By using the semi-polar system, the forward-backward movements can not be described effectively, neither can the convexly or concavely bending of the top front surface just behind the tongue apex.

There is another problem which links to factor analysis of the articulatory data. Usually, vectors for factor analysis must have fixed length. This rule is such a strong constraint that the measurements of the tongue apex in some image frames must be discarded unfortunately to maintain the homogeneity of the measured data. Using the semi-polar system in Figure 1, the tongue contours in all the frames intersect with the 23rd grid line while the tongue contours in some frames do not even intersect with the 24th line. So, the last measured coordinate of the tongue contour is chosen on the 23rd line; the coordinates on the 24th and 25th lines are discarded. That implies that, for many image frames, the geometric information of the tongue apex can not be described correctly by using the representation vector.

To improve the accuracy of the measurement and the description ability of the articulatory models, a modified semi-polar system is proposed, as shown in Figure 2. Firstly, the density of the grid lines in the two linear regions is increased in order to have a better precision of the geometric description. Furthermore, the position of the tongue apex had been measured and thus been included in the observation vector. The tongue apex is defined as the farthest point on the tongue contour to the right. The position of the tongue apex is defined as an x-y tuple (x_t, y_t). Here, x_t and y_t are the coordinates of the tongue apex measured along and perpendicular to the last grid line, respectively.

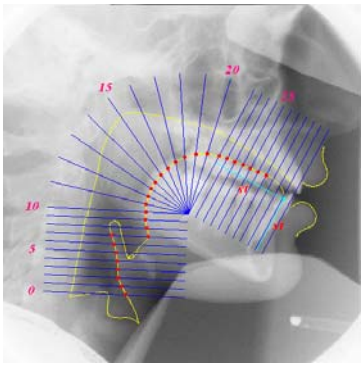

Figure 2: *The modified semi-polar system for measuring data for Speaker FH*

With the measured data of the tongue apex and of the jaw opening, the observation parameters of the tongue contour constitute a *24*-dimensional observation vector:

$$P_{tongue} = \left[C_7, C_8, C_9, \cdots, C_{26}, C_{27}, x_t, y_t, J \right]^T \quad (1)$$

where C_i (i=*7*, *8*, …, *27*) is the coordinate of the intersection point on the ith grid line. Similarly, the observation vector of the larynx outline is defined as a *10*-dimensional vector:

$$P_{larynx} = \left[C_0, C_1, C_2, \cdots, C_7, H, J \right]^T \quad (2)$$

where H is the larynx height.

4. Feature identification and linear articulatory modeling

The articulatory representation of speech is usually realized as low dimensional articulatory models. Normally, the vocal tract outlines are regarded as linear combinations of the effects of different articulatory features. The articulatory models, therefore, are realized by linear transformations from the features to the observation vectors. Usually, a regression analysis is performed on the measured data to determine these linear transformations.

4.1. Articulatory features selection

The features usually can not be determined automatically from the data by a regression analysis. To extract features from the measured data, we must have a basic idea in advance about how to represent the vocal tract as features, based on the physiology of vocal tract and the knowledge of speech production. A jaw-based representation was proposed by Lindblom et al. [7] and was applied successfully by different researchers in articulatory modeling [3-5]. In this representation, the vocal tract shapes are assumed to be functions of the features such as the jaw, the shape and the position of the tongue body, the position of the tongue tip, the lip height and width, and the larynx height. Upon determining the feature set, factor analysis can be used to extract the features from the data set, as well as to determine the linear transformation [3-5, 8].

4.2. Factor analysis for articulatory modeling

Typically, factor analysis is used to uncover the latent structure of a set of variables; that is, it is a "non-dependent" procedure in terms of that no dependent variable is assumed in advance. Several different factor analysis techniques, such as the principal component analysis (PCA) [3, 9] and the PARAFAC [10], have been used to extract elementary features from the measured data of vocal tract. Though both methods can provide a unique solution for a given set of data, they do not guarantee that every factor can be interpreted by articulatory terms; the extracted factors are inferred rather than observed. The analysis usually can not result in an explanatory articulatory model.

4.3. Articulatory modeling based on guided PCA

To deal with the problem with PCA and PARAFAC, the guided PCA technique has been proposed to help us to extract a set of interpretable factors [11]. The guided PCA is a procedure of arbitrary orthogonal factor analysis followed by PCA. The idea is that the measured data are not directly subjected to PCA but to an arbitrary factor analysis at first which can be considered as an extraction-subtraction procedure of known parameters. The state of the most important parameter is extracted by means of linear regression and then subtracted from the data set. Then, another less important variable is extracted and its effect is subtracted in the same way. This procedure is repeated until all relevant factors corresponding to elementary features are extracted one-by-one. At the end, the residual can be subjected to a PCA analysis to extract the variance maximally with a minimal number of factors.

We take the analysis of the tongue dorsum outline as the example to describe the guided PCA technique. According to a theoretic model of articulation [5], the average midsagittal outlines of the tongue dorsum are mainly determined by the postures of the following elementary articulators:
the parameter of jaw opening;
the state of tongue body front-back position;
the tongue dorsal flat-arched shape;
the state of tongue apex.

Usually, the jaw opening is the most important parameter which should be extracted as the first factor. With the measured data of (*p-1*) variables of tongue outlines, we can easily estimate the correlations between the (*p-1*) variables and the parameter of jaw, forming the correlation matrix $\boldsymbol{R}$ of order $p \times p$. Factor loadings for the parameter of jaw are nothing but the correlation coefficients between the jaw parameter and the variables of tongue outline. This set of factor loadings for the jaw parameter can be denoted as a vector $\boldsymbol{a}_1$:

$$\boldsymbol{a}_1 = \left[a_{11}, a_{21}, \cdots, a_{(p-1)1}, 1 \right]^T \quad (3)$$

where a_{i1} (i = *1*, *2*, …, *p-1*) is the correlation coefficient between the ith tongue variable and the jaw parameter. The influence of the jaw parameter is extracted by the vector $\boldsymbol{a}_1$ and then can be eliminated from the correlation matrix $\boldsymbol{R}$ by subtraction as follows:

$$\boldsymbol{R}_1 = \boldsymbol{R} - \boldsymbol{a}_1 \boldsymbol{a}_1^T \quad (4)$$

where $\boldsymbol{R}_1$ represents the residual variance-covariance matrix. Since the influence of the jaw parameter has been eliminated, the values of the elements in the last row and the last column of $\boldsymbol{R}_1$ are zeros. Thus, $\boldsymbol{R}_1$ is in fact a matrix of order (*p-1*)×(*p-1*). $\boldsymbol{R}_1$ containing no more influence of the jaw parameter implies that any measured variable in the pharyngeal region can be regarded as a factor of tongue body position, because the outline of the tongue in this region is mainly determined by the postures of the jaw and the tongue body position. The variable in this region which has the maximum proportion of variance of the measured data is chosen as the factor of tongue body position, and therefore the loadings of this factor can be derived from $\boldsymbol{R}_1$. The influence of the factor of tongue body position is eliminated from $\boldsymbol{R}_1$ to form a (*p-2*)×(*p-2*) residual matrix $\boldsymbol{R}_2$, which afterwards can be used to determine the weights of the factor of tongue dorsal shape, by taking the variable with the maximum proportion of variance in the tongue dorsal region as this particular factor. The weights of the factor of tongue apex can also be calculated after eliminating the influence of the factor of tongue dorsal shape. One more factor, such as the parameter of tongue root, could be extracted in the same way to improve the descriptive ability of the features. Also, some unknown factors, if they exist, can be determined by subjecting the final residual matrix to a PCA. In this way, the factors corresponding to the elementary postures are sure to be extracted and an accurate linear model can be built up.

5. Experimental results

The articulatory data for the speaker FH are analyzed by using the guided PCA. For modeling the tongue outline, the parameter of jaw opening is extracted as the first factor, contributing 31.85% of the variance. Three other factors are extracted one-by-one. These four factors and their extracted proportions of the variance are listed in Table 1. The four factors, being taken together, account for 93.83% of the variance of the observed tongue data for the speaker FH. Apparently, an articulatory model based only on these four factors can not describe the measured movements of the tongue outline in a satisfied precision

Table 1. Extracted factors for the tongue outline

Factor Index	Factor Name	Parameter	Proportion of Variance
1	jaw-opening	J	31.85
2	tongue dorsal shape	C_{20}	30.66
3	tongue body position	C_{14}	26.68
4	tongue apex	C_{26}	4.64
5	tongue root	C_8	3.03

To build up a more precise linear model of tongue movement, one more explainable factor is extracted from the residual matrix. This factor corresponded to the parameter C_8, therefore is regarded as the factor of tongue root. 3.03% of the variance is contributed by this factor. The total proportion of the variance due to all five factors amount to 98.86%; only 1.14% of the variance can be attributed to unknown sources. Thus, the linear articulatory model based on these five factors is precise enough to describe the variation of tongue outline. Further PCA of the residual matrix is not necessary. For all image frames, the root mean squared (RMS) reconstruction error relative to the initial marked tongue contours is 2.84%. For vowels, the reconstructed tongue contours simulate the marked ones in a high precision. For consonants, however, there is a significant reconstruction error in the apical region, while in other regions the reconstructed tongue contours matches the marked ones in a high accuracy. Figure 3 shows an example of reconstruction results for an apical consonant.

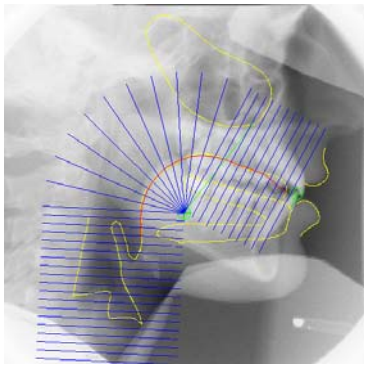

Figure 3: *The simulated contour (red) vs. the marked contour (yellow) for an apical consonant*

Furthermore, we have built an ANN-based non-linear articulatory model based on the factors extracted by the guided PCA and have compared the performance of the non-linear model with that of the linear one. Multilayer perceptrons have been trained on the measured data by using the MATLAB Neural Network Toolbox. Results show that the linear articulatory model generated by the guided PCA performs as well as the ANN-based non-linear articulatory models in terms of reconstruction accuracy.

6. Conclusions

Increasing the number of grid lines in the two linear regions of the semi-polar system is an effective way to increase the descriptive precision of the vocal tract outlines. Introducing new parameters of tongue apex is a practical way to describe the position of the tongue tip, but it fails to describe the variation in the frontal part of tongue. Some new parameters and perhaps a totally new coordinate system must be designed for representing and measuring the outline in the tongue apical region more precisely. The linear models generated by using the guided PCA perform well in simulating the vocal tract outlines and therefore the guided PCA is an effective way to extract articulatory features and to build articulatory models. Extracting a few more explainable factors (e.g. the tongue root) is a feasible way to build up highly precise articulatory models.

7. Acknowledgements

This work was supported by the research fund of European Commission, under Project “Audiovisual to Articulatory Speech Inversion (ASPI)” (Project No. 2005-021324). Gratitude is due for Dr. Shinji Maeda for his valuable advice and helpful supports.

8. References

[1] Schroeter, J., and Sondhi, M. M., "Techniques for Estimating Vocal-tract Shapes from the Speech Signal", IEEE Trans. Speech and Audio Proc., 2(1), Part 2:133-150, 1994.

[2] Dusan S. V., "Statistical Estimation of Articulatory Trajectories from the Speech Signal Using Dynamical and Phonological Constraints", Ph. D. Thesis, Dept. of Electrical and Computer Eng., Univ. of Waterloo, 2000.

[3] Maeda, S., "Une Analyse Statistique Sur Les Positions De La Langue: Etude Preliminaire Sur Les Voyelles Francaises", 9èmes Journees d’Etude Sur La Parole:192-199, 1978.

[4] Maeda, S., "A Digital Simulation Method of the Vocal-Tract System", Speech Communication, 1(3-4):199-229, 1982.

[5] Maeda, S., "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 ([Ed]), Speech Production and Speech Modelling, 131-149, Kluwer Academic Publishers, 1990.

[6] Jackson, M. T.-T., and McGowan, R. S., "Predicting Midsagittal Pharyngeal Dimensions from Measures of Anterior Tongue Position in Swedish Vowels: Statistical Considerations", J. Acoust. Soc. Am., 123(1):336-346, 2008.

[7] Lindblom, B., et al., "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation", J. Acoust. Soc. Am., 62(S1): S15-S15, 1977.

[8] Engwall, O., "A 3D Vocal Tract Model for Articulatory and Visual Speech Synthesis", In Proc. of Fonetik 98, The Swedish Phonetics Conference:196-199, 1998.

[9] Slud, E., et al., "Principal Components Representation of the Two-Dimensional Coronal Tongue Surface", Phonetica, 59(2-3):108-133, 2002.

[10] Nix, D. A., et al., "Two Cross-linguistic Factors Underlying Tongue Shapes for Vowels", J. Acoust. Soc. Am., 99(6):3707-3717, 1996.

[11] Maeda, S., "Face Models Based on A Guided PCA of Motion-capture Data: Speaker Dependent Variability in /s/-/ / Contrast Production", ZAS Papers in Linguistics, 40:95-108, 2005.