DETERMINATION OF THE VOCAL TRACT SPECTRUM FROM THE ARTICULATORY MOVEMENTS BASED ON THE SEARCH OF AN ARTICULATORY-ACOUSTIC DATABASE

Tokihiko Kaburagi and Masaaki Honda

Information Science Research Laboratory,
NTT Basic Research Laboratories,
3-1, Morinosato-Wakamiya, Atsugi, Kanagawa, 243-0198 Japan

ABSTRACT

This paper presents a method for determining the vocal-tract spectrum from the positions of fixed points on the articulatory organs. The method is based on the search of a database comprised of pairs of articulatory and acoustic data representing the direct relationship between the articulator position and vocal-tract spectrum. To compile the database, the electro-magnetic articulograph (EMA) system is used to measure the movements of the jaw, lips, tongue, velum, and larynx simultaneously with speech waveforms. The spectrum estimation is accomplished by selecting database samples neighboring the input articulator position and interpolating the selected samples. In addition, phoneme categorization of the input position is performed to restrict the search area of the database to portions of the same phoneme category. Experiments show that the mean estimation error is 2.24 dB and the quality of speech synthesized from the estimated spectrum can be improved by using the phoneme categorization.

1. INTRODUCTION

The electro-magnetic articulograph (EMA) system can monitor the movements of articulatory organs inside and outside the vocal tract with fine space and temporal resolutions, making it a useful tool for the study of the dynamical aspects of speech production [1,2]. However, it is very difficult to accurately determine the whole configuration of the vocal tract and examine the acoustic consequences of measured articulator movements because EMA systems are designed to detect only the positions of multiple points fixed on the articulators.

Here, we present a method to determine the vocal-tract spectrum from the positions of fixed points on the articulatory organs based on the search of an articulatory-acoustic database. The database is comprised of articulatory and acoustic data pairs that directly represent the relationship between the articulator position and the vocal-tract spectrum. The spectrum estimation is performed by finding the database samples that are coincidental with the input articulator position instead of calculating the vocal-tract area function and transfer function. On the other hand, our method requires a large amount of accurate articulatory data taken from articulatory organs which can affect the vocal-tract transfer function.

In this paper, we first describe the articulatory and acoustic measurement to construct the database. Next, the procedure for determining the vocal-tract spectrum from the articulator position is presented and finally, the accuracy of our spectrum estimation method is evaluated.

2. ARTICULATORY AND ACOUSTIC MEASUREMENT

This section describes the measurement method to assemble an articulatory and acoustic data set (Fig. 1).

To monitor the movements of the articulatory organs that might influence the acoustic characteristic of the vocal-tract, receiver coils of the EMA system (Carstens Articulograph AG100, Germany) were attached to the jaw (J), upper lip (UL), lower lip (LL), tongue (T), velum (V), and larynx (L) on the midsagittal plane and their movements were recorded at a sampling rate of 250Hz using an adaptive calibration method [3,4]. The accuracy of this calibration method is 0.106 mm for a 14×14 cm region when the receiver coil is on the midsagittal plane and is about 1 mm when the off-center misalignment is 2 mm (the tilt angles of the coil with respect to the x and y axes are less than 20 degrees for both cases). Four receiver coils were placed on the tongue from the tip to the dorsum at almost equal intervals. The larynx position was monitored by attaching the edge of a bar which could rotate smoothly on the midsagittal plane to the Adam's apple and by placing a receiver coil on the bar. The positions of two coils on the nose bridge and upper incisors were also measured to calibrate the movement of the head.

Speech waveforms were recorded at a sampling frequency of 8kHz and processed using 30-msec hamming window and second-order pre-emphasis. The center position of the hamming window was set at each sampling point of the EMA measurement to synchronize the articulatory and acoustic data. The pre-emphasis canceled the glottal and radiation characteristics. Finally, 12-order LPC analysis was performed and the values of the LSP parameters were determined as the acoustic data representing the vocal-tract spectrum.

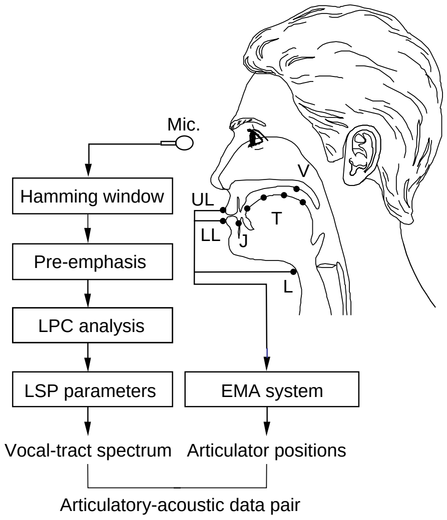

Figure 1: Procedure for constructing the articulatory-acoustic database.

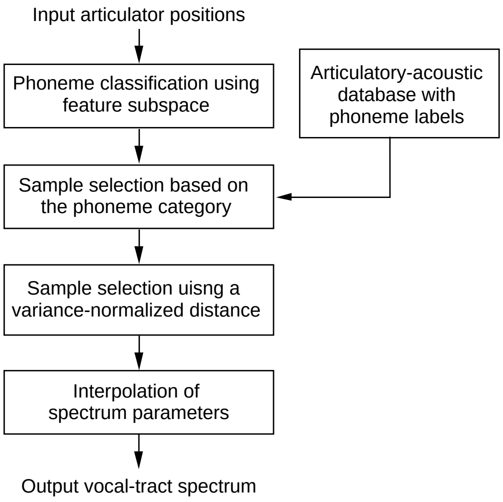
Input articulator positions
Phoneme classification using feature subspace
Articulatory-acoustic database with phoneme labels
Sample selection based on the phoneme category
Sample selection uisng a variance-normalized distance
Interpolation of spectrum parameters
Output vocal-tract spectrum

Figure 2: Procedure for determining the vocal-tract spectrum from the articulator position.

Articulatory and acoustic measurement was performed with a male Japanese subject while he made 488 utterances (92 vowel sequences, 196 non-words including voiced consonants, 110 words including voiced and unvoiced consonants, and 50 sentences). Each articulatory and acoustic data pair respectively stored the positions of nine receiver coils and the values of the LSP parameters at an instant during the utterance. In addition, the instant at which each phoneme was articulated was determined manually and a label representing the phoneme category was assigned to the database. The total number of data pairs was 79193, which corresponded to a duration of about 5.3 minutes, and the number of phoneme labels was 3247.

3. SPECTRUM ESTIMATION METHOD

Next, we describe the method for determining the vocal-tract spectrum from the input positions of the articulatory organs. A diagram of this spectrum estimation method is shown in Fig. 2.

The method is based on the selection of the articulatory-acoustic database samples for input articulator position. Before the sample selection, the phoneme category of the input articulator position is first determined and the search area of the database is restricted to portions of the same category to ensure the acoustical reliability of the resulting spectrum. Database samples coincidental with the input position are then selected using a variance-normalized distance between the input position and the database sample with the same phoneme category. The output spectrum is finally calculated using an weighted interpolation of the selected samples to maintain the continuity of the mapping.

In the following, procedures are described for the determination of the phoneme category, selection of the database samples, and determination of the output spectrum.

3.1. Phoneme Categorization

To determine the phoneme category of the input articulator position, a phoneme classification method is implemented using phoneme-specific feature subspace [5] in the articulatory domain. The feature subspace of a class p ($p = 1, 2, ..., P$) is defined as the linear transformation $\mathbf{f}_p$ that minimizes the following variance ratio

$$J(\mathbf{f}_p) = \frac{\mathbf{f}_p^t \Sigma_p \mathbf{f}_p}{\mathbf{f}_p^t \Sigma_T \mathbf{f}_p} \tag{1}$$

for Σ_p, the covariance matrix of the class p, and Σ_T, that for the total data. $\mathbf{f}_p$ is determined by solving an eigenvalue problem

$$\Sigma_p F_p = \Sigma_T F_p \Lambda_p, \tag{2}$$

where $F_p = (\mathbf{f}_{p1}, \mathbf{f}_{p2}, ..., \mathbf{f}_{pL})$ is the eigenvector matrix and $\Lambda_p = diag(\lambda_{p1}, \lambda_{p2}, ..., \lambda_{pL})$ represents the matrix in which the eigenvalues are stored in ascending order. L is the dimension of the articulatory data. The feature subspace is finally determined as $\tilde{F}_p = (\mathbf{f}_{p1}, \mathbf{f}_{p2}, ..., \mathbf{f}_{pL_p})$ where L_p is the dimension of the subspace ($L_p \leq L$).

The phoneme category of the input position $\mathbf{x}$ is determined as the phoneme class for which the following projection norm C_p is the smallest:

$$C_p = \sum_{l=1}^{L_p} \{\mathbf{f}_{pl}^t(\mathbf{x} - \bar{\mathbf{x}}_p)\}^2 / \lambda_{pl}, \tag{3}$$

where $\bar{\mathbf{x}}_p$ is the mean vector. This norm is weighted by the inverse of the eigenvalue because the axis in the subspace represents the phoneme-specific invariant feature when its eigenvalue, which equals the variance ratio J, is small. The projection norm takes a small value when the input articulator position matches this feature. As a result, phoneme classification can be performed.

3.2. Database Sample Selection

The selection of the articulatory-acoustic database samples is first performed based on the phoneme category. If ith phoneme label assigned to the database is the same as the phoneme category of the input articulatory position, the database samples within the time interval $t_{i-1} \leq t \leq t_{i+1}$ are selected, where t_{i-1} and t_{i+1} are the instants at which the preceding and following phoneme labels are assigned.

The database samples coincidental with the input position are then collected using a variance-normalized distance

$$e_i = (\mathbf{x} - \mathbf{x}_i)^t W (\mathbf{x} - \mathbf{x}_i) \qquad (4)$$

between the input position $\mathbf{x}$ and the database sample $\mathbf{x}_i$ with the same phoneme category. Here, each component of the weighting matrix $W = diag(w_1, w_2, ..., w_L)$ is given as $w_l \propto c_l^{-0.5} (\sum w_l = 1)$ using c_l, variance of the articulator position in the database. The neighboring database samples, $\mathbf{x}_j$ and $\mathbf{y}_j$ for $j = 1, 2, ..., M$ where M is the sample number, are finally selected for those the distance e_j are smaller than those for the remaining database samples.

3.3. Spectral Interpolation

Finally, to determine the values of the vocal-tract spectrum parameters $\mathbf{y}$, weighted interpolation of the selected database samples is calculated as

$$\mathbf{y} = \sum_{j=1}^{M} v_j \mathbf{y}_j. \qquad (5)$$

The weighting coefficient v_j is given as $v_j \propto e_j^{-2} (\sum v_j = 1)$ so that the database sample closer to the input position is weighted more heavily.

4. EXPERIMENT

Experiments were conducted to determine the accuracy of our spectrum estimation method. The estimation error was evaluated as a function of the number of the neighboring samples, which were interpolated to calculate the vocal-tract spectrum, and the type of the utterance. Next, the effect of phoneme categorization was investigated by evaluating the quality of the speech synthesized using the estimated vocal-tract spectrum.

4.1. Accuracy of Spectrum Estimation

Spectrum estimation was performed using the articulatory and acoustic data set described in the second section. One out of 488 utterances was selected as the test utterance and the data for the remaining utterances were used as the articulatory-acoustic database in the spectrum estimation method. The vocal-tract spectrum was determined at each instant of the articulatory movement of the test utterance. The error between the estimated and actual spectra was calculated using the 30-order cepstrum distance while changing the test utterance over the entire data set. The estimation error averaged for each type of utterance is shown in Fig. 3 as a function of the number of neighboring samples. Phoneme categorization was not used in this experiment. Estimation results are also shown in Fig. 4 by comparing the original and estimated spectra as well as the phoneme labels, articulator movements, and speech waveforms.

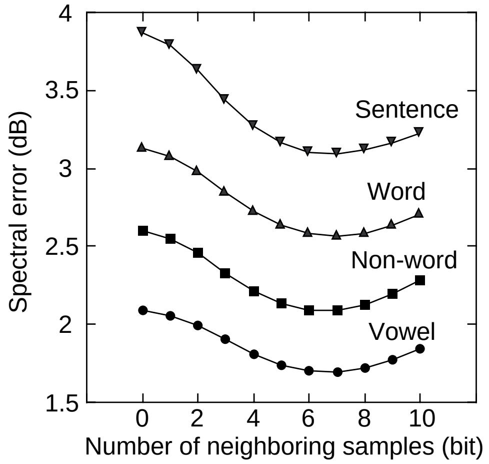

Figure 3: Relationship between the number of neighboring database samples and the spectral estimation error in each type of utterance.

Figure 3 shows that the spectral error is influenced by the number of neighboring samples, and its minimum is obtained at a sample number of 128. This indicates that the sample interpolation is effective in reducing estimation error by achieving continuous mapping. The estimation errors for vowel sequences, non-words, words, and sentences are 1.69, 2.08, 2.56, and 3.09 dB, respectively, at a sample number of 128. The mean and standard deviation for all of the utterances are 2.24 and 1.31 dB, showing good agreement between the estimated and actual vocal-tract spectra.

4.2. Effect of Phoneme Categorization

When the sample selection based on the phoneme category was used, the mean spectral estimation error (2.27 dB) was almost the same as the error when the categorization was not used (2.24 dB). Next, speech waveforms were synthesized using the estimated spectrum and excitation signals extracted from the original speech based on a multi-pulse and noise source model [6] to subjectively evaluate the speech quality. In this experiment, pair-wise speech

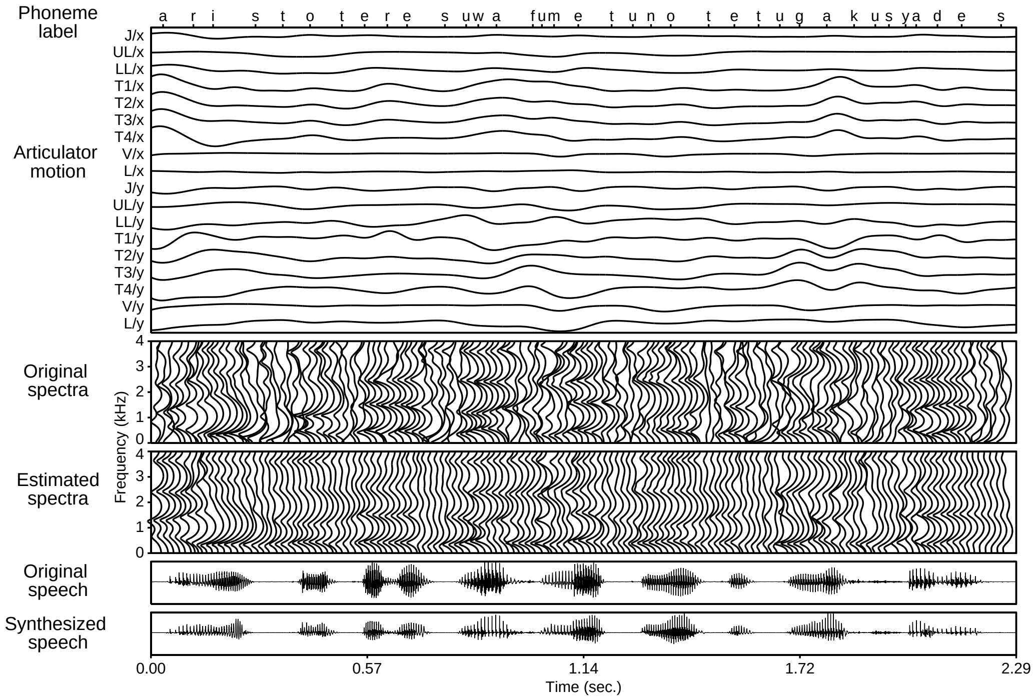

Figure 4: Results of spectral estimation and speech synthesis from articulatory movements.

stimuli synthesized with and without phoneme categorization were presented to ten subjects for each of ten sentences through headphones. The result was that 84% of speech stimuli synthesized with phoneme categorization were preferred in quality to those without categorization, indicating that categorization is effective in improving speech quality.

It was also found that the score of the proposed phoneme classification method for a close data set (86.2%) was much higher than that of the nearest neighbor method using the variance-normalized distance (46.7%) for phoneme classes including 5 vowels, 2 semi-vowels, and 15 consonants. In addition, phoneme categorization reduced the distance calculation in the database search to one fifth. The phoneme classification results indicate that phoneme categorization can reduce the spectrum estimation error caused by selecting database samples of different phoneme categories, even though this is not reflected in the cepstrum distance.

Speech samples synthesized with phoneme categorization are included in the CD-ROM [SOUND 0425_01.WAV] [SOUND 0425_02.WAV] [SOUND 0425_03.WAV].

5. CONCLUSION

It is concluded from the experimental results that the proposed method is useful for determining the vocal-tract spectrum and synthesizing speech waveforms from articulatory movements. By combining the method with the articulatory movement model [7], the proposed method can be applied to articulatory-based speech synthesis.

REFERENCES

[1] Schönle, P.W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," Brain and Language **31**, 26-35.

[2] Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., and Jackson, M.T.T. (1992). "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078-3096.

[3] Kaburagi, T., and Honda, M. (1994) "Determination of sagittal tongue shape from the positions of points on the tongue surface," J. Acoust. Soc. Am. **96**, 1356-1366.

[4] Kaburagi, T., and Honda, M. (1997) "Calibration methods of voltage-to-distance function for an electromagnetic articulometer (EMA) system," J. Acoust. Soc. Am. **101**, 2391-2394.

[5] Honda, M., and Kaburagi, T. (1996) "Statistical analysis of a phonemic target in articulatory movements," ASA and ASJ Third Joint Meeting 1pSC4.

[6] Honda, M. (1989) "Speech analysis-synthesis using phase-equalized excitation ," Technical report of IEICE, SP89-124 (in Japanese).

[7] Kaburagi, T., and Honda, M. (1996) "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes," J. Acoust. Soc. Am. **99**, 3154-3170.