DETERMINATION OF ARTICULATORY POSITIONS FROM SPEECH ACOUSTICS BY APPLYING DYNAMIC ARTICULATORY CONSTRAINTS

Shin SUZUKI, Takesi OKADOME, and Masaaki HONDA
NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atugi-Shi, Kanagawa, 243-0198 Japan
e-mail: shin@idea.brl.ntt.co.jp

ABSTRACT

A method for determining articulatory parameters from speech acoustics is presented. The method is based on a search of an articulatory-acoustic codebook which is designed from simultaneous observation data of articulatory motions and speech acoustics. The codebook search employs dynamic constraints on acoustic behavior as well as articulatory behavior. There are two constrains. One of the constraints is use of spectral segments in the codebook search and the other is use of the smoothness of articulatory trajectories in the articulatory parameter path search. The articulatory parameters are determined by selecting the articulatory code vector in the codebook which minimizes the weighted distance measure of segmental spectral distance and squared distance between succeeding articulatory parameters. An experiment was conducted to evaluate the efficiency of both constraints in determining of articulatory parameters by comparing the estimated and the observed articulatory parameters. The results show that an rms error between the estimated and observed articulatory parameter was about 2.0 mm on average, and the articulatory features for vowels and consonants are recovered well.

1. INTRODUCTION

Acoustic-to-articulatory inverse problem for determining articulatory parameters from speech acoustics is characterized by one-to-many mapping (Atal et al., 1978). In order to uniquely determine an articulatory parameter from speech acoustics, additional constraints on articulatory configuration and its dynamic behavior are required.

Schroeter and Sondhi (1992, 1994) presented a method of inverse mapping based on an articulatory-acoustic codebook search. In this method, the codebook was designed from the articulatory-acoustic pair data which is computed by using the geometrical articulatory and vocal tract acoustic models and by selecting uniform samples in an articulatory space because of a lack of quantitative knowledge of articulatory distribution. They also use continuity constraint on articulatory trajectories in determining articulatory parameters from speech acoustics. The dynamic constraint was, however, limited to the temporal smoothness of articulatory motions because of a lack of quantitative knowledge about the actual articulatory temporal behavior.

In this study, we construct an articulatory-acoustic codebook based on simultaneous observations of articulatory motions and speech acoustics. The use of observed data gives static constraint on the articulatory configuration (Hogden et al., 1996). We also introduce two dynamic constraints on the codebook search for determining articulatory parameters: One constraint is a codebook search using spectral segment matching between input speech acoustics and the code vectors in the codebook. Spectral segment behavior is a consequence of the temporal articulatory behavior; thus, spectral segment matching implicitly gives dynamic constraint on temporal articulatory behavior. The continuity of articulatory trajectories is also used in the codebook search as an explicit dynamic constraint on the articulatory behavior.

2. DETERMINATION PROCEDURE

The procedure for determining articulatory parameters from speech acoustics is shown in Figure 1. The procedure consists of the following three parts:

2.1. Articulatory-acoustic codebook

The articulatory-acoustic codebook is designed from simultaneous observations of articulatory motions and speech acoustics, which are taken from continuous speech utterances. Articulatory parameters are represented by the vertical and horizontal positions of multiple points of articulatory organs. Spectral parameters are obtained from recorded speech signals. The codebook contains pair data of spectral segments and articulatory parameters as shown in Figure 1. Each articulatory parameter is associated with a spectral segment. The interval time of the segment spans a fixed period before and after articulatory timing. In this study, non-clustering of the observed data was used in the codebook design and all of the training samples were used as code vectors.

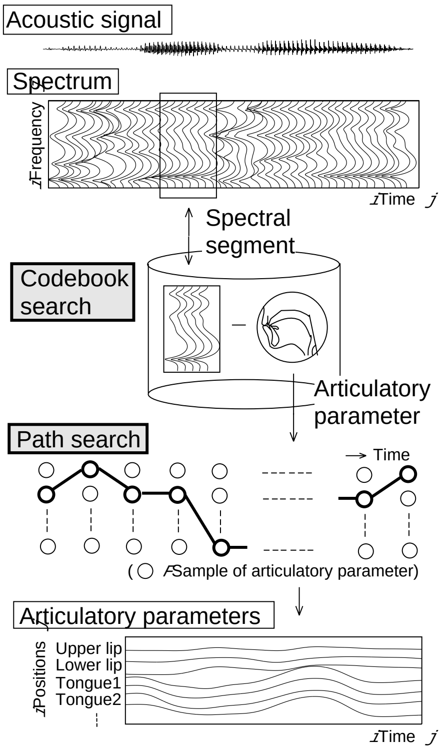

Fig.1. Procedure for articulatory parameter determination.

2.2. Codebook search using spectral segment matching

To estimate the articulatory parameters, we first transform a given input speech signal into a spectrum sequence. Second, each spectral segment of the sequence is compared with the code spectral segments in the codebook in every frame, and the spectral segment distance is computed. Then, a fixed number of candidates of articulatory parameters are selected from the articulatory-acoustic pair code vectors in order of increasing spectral segment distance. Thus we obtain candidate sequences of articulatory parameters.

2.3. Path search among the candidates of articulatory parameters

Among the candidate sequences selected in the codebook search, we search for an articulatory parameter path which minimizes the weighted distance-measure defined by

$$\min_{(c(t),\, x(t))} \sum_t \{D_s(c_s(t),\ c(t)) + w \cdot D_x(x(t-1),\ x(t))\},$$

where $(c(t),\ x(t))$ is a candidate code vector, $(c_s(t))$ is the spectral segment of an input speech signal, D_s is the average spectral distance on the spectral segment, D_x is the squared distance, w is a weight, and t is the time. Minimizing the squared distance D_x between succeeding candidates of articulatory parameters introduces smooth trajectories of articulatory motions. The weighted distance measure reflects both the acoustical distance and the smoothness of articulatory motions of the sequences.

The path sequence of the articulatory parameters which will minimize the weighted distance measure is obtained by finding the optimum solution using a dynamic programming method.

A codebook search using spectral segment matching together with an articulatory parameter path search based on trajectory smoothness reduces ambiguity in acoustic-to-articulatory mapping.

3. EXPERIMENT

3.1. Data and experimental conditions

We conducted an experiment in order to evaluate the method. In the experiment, an articulatory-acoustic codebook was constructed from simultaneous observation of articulatory motions and speech acoustics using an electro-magnetic articulograph (EMA) (Kaburagi and Honda, 1994). The procedure of simultaneous articulatory-acoustic observations is shown in Figure 2.

The articulatory data was collected using EMA at a sampling rate of 250 Hz. The articulatory data represents the vertical and horizontal positions of nine points, which were on the lower jaw, the upper and lower lips, the tongue, the velum and the Adams apple for larynx height monitoring. A speech signal was simultaneously recorded and sampled at 8 kHz. Then 30 LPC cepstral coefficients were obtained from it as speech acoustical parameters.

Simultaneous articulatory-acoustic observations were made for 354 sentences spoken by a Japanese male. The articulatory-acoustic codebook was designed from the data of 338 sentences randomly selected among the 354. The remaining data of 16 sentences were used as test data for the evaluation of the method. The codebook contains 222894 articulatory-acoustic pair data.

In the experiment, the interval time of spectral segments was tested in a range of 4 ms (single frame) to 320

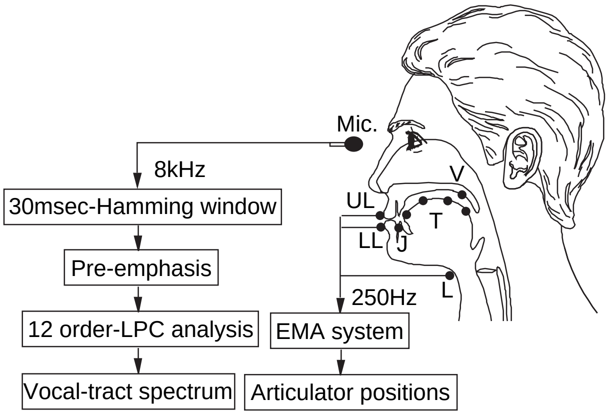

Fig.2. Simultaneous articulatory and acoustic observations.

ms. The weight value of the distance measure used in the articulatory path search was also tested in a range of 0.001 to 1. The number of candidates of articulatory parameters for the path search was fixed at 100. The estimated articulatory parameters were smoothed using FIR low-pass filter at a cut off frequency of 10 Hz in the final step.

3.2. Results

First, we tested the estimation error of the articulatory parameters in terms of the interval time of the spectral segment in the codebook search. Figure 3 shows the average rms error of the estimated articulatory parameters for the test data as a function of the segment interval for two methods: codebook searches with and without the continuity path search. Here, the segment interval of 4 ms corresponds to single frame spectral matching in the codebook search. The use of spectral segments in a codebook search is more efficient for reducing the error than is possible in the single spectral matching for both methods. The error shows the minimum value at segment intervals of 160 ms and 200 ms for each method. The segment interval of 160 ms corresponds to the interval of a triphone at the normal speed utterance. Moreover, the minimum error for the method without path search is slightly better than that for the method with the single frame spectral matching and the path search. This means that the the codebook search based on the temporal patterns of speech acoustics is efficient for reducing the ambiguity in acoustic-to-articulatory inverse mapping by employing an implicitly dynamic constraint on articulatory temporal behavior. The continuity constraint of articulatory trajectories in addition to a spectral segment codebook search provides additional improvement in the estimation accuracy.

Figure 4 shows the effect of the weight value of the weighted distance measure on the estimation error. In this case, the segment interval is at 160ms. The error shows a minimum value for the weight value of 0.01.

The estimated articulatory trajectories and the spectral pattern for the test sentences are shown in Figure 5 (on the next page) as compared with the measured ones.

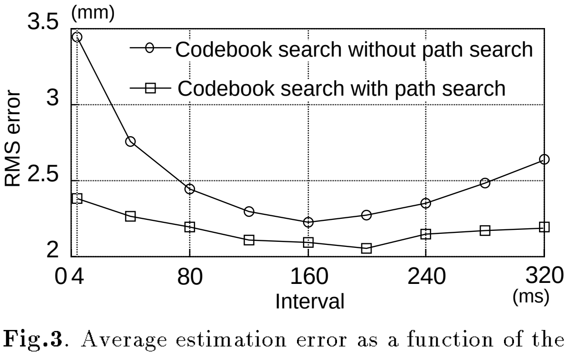

Fig.3. Average estimation error as a function of the interval time of the spectral segment.

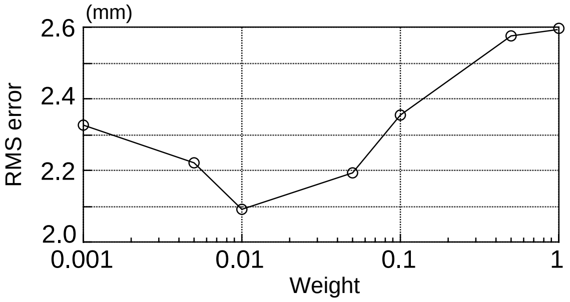

Fig.4. Average estimation error as a function of the weight value of the weighted distance measure.

The estimated articulatory trajectories capture features of actual articulatory behavior well for not only vowels but also for most consonants. The consistent features of the primary articulator for consonants such as labial closure for labial consonants and tongue tip closure for alveolar consonants are recovered well in the estimated articulatory parameters. The spectral error between the estimated and measured spectra in this example is 1.70 dB on average. This means that speech can be synthesized from the estimated articulatory parameters using articulatory-acoustic codebook with a little perceptual degradation.

Figure 6 summarizes the estimation error. In this figure, 'ALL' shows the rms error averaged over all articulatory parameters and the entire test data, and 'VOWEL' to 'NASAL' show the rms error averaged over particular primary articulatory parameters and articulatory timing instants. For example, the errors in 'Labial' and 'Alveolar' are evaluated only for the lip parameters and the tongue tip parameters, respectively. The total average error is 2.09 mm. The error is relatively small for labial, alveolar and nasal consonants, but large for velar consonants.

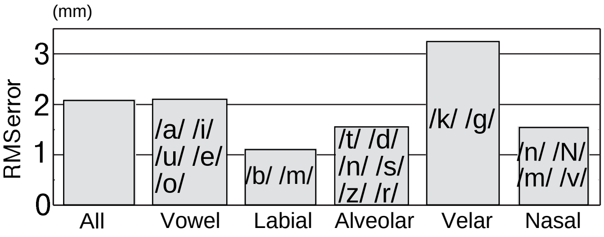

Fig.6. Average rms error of estimated articulatory parameters.

4. CONCLUSION AND DISCUSSION

We have presented a method for recovering articulatory motion from speech acoustics. A codebook search using spectral segment matching together with the continuity constraint on articulatory trajectories is efficient for

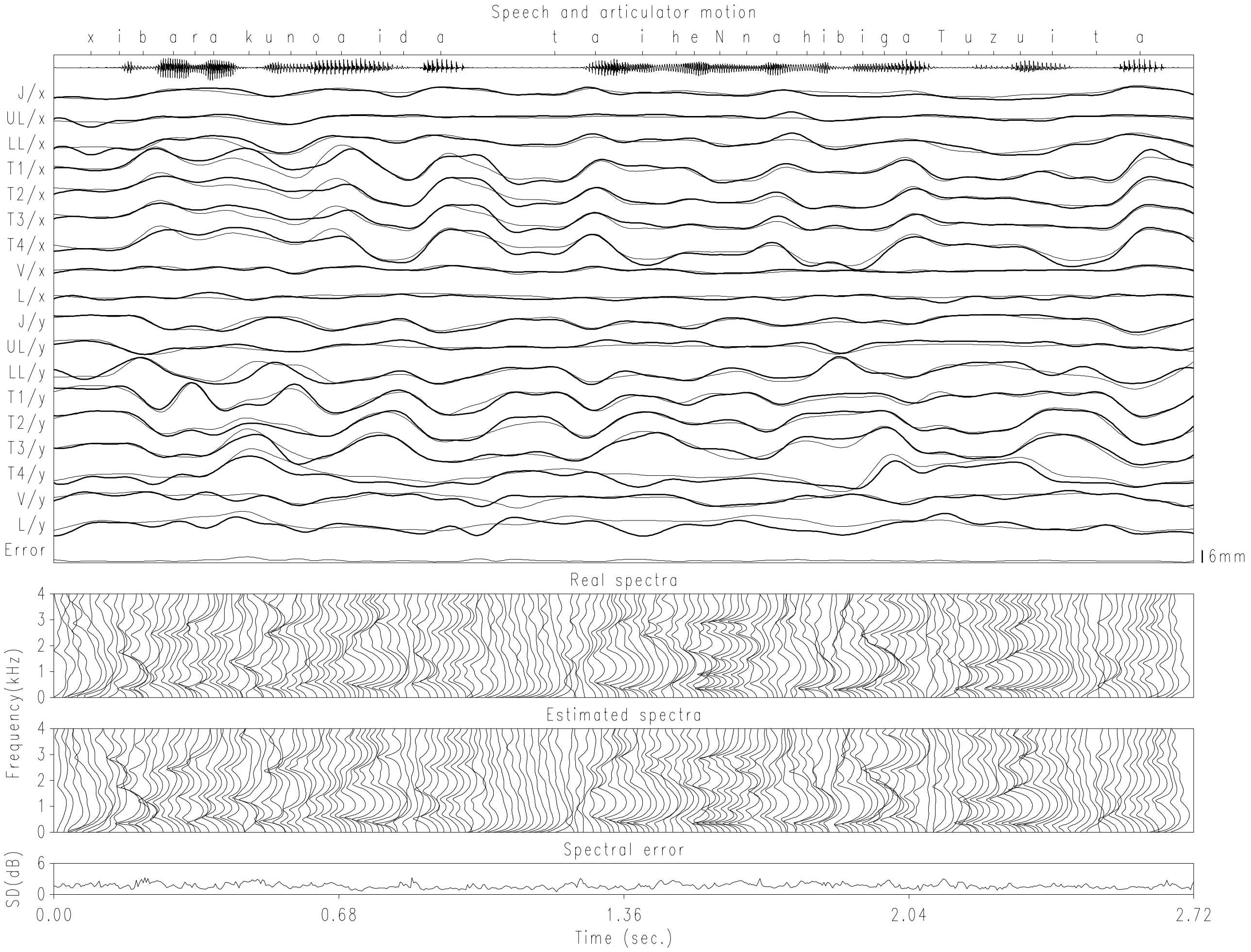

Fig.5. Comparison of the estimated articulatory parameters and spectra with the observed ones. The thick lines are the estimates and the thin lines are the measured ones.

implementing acoustic-to-articulatory inverse mapping. Spectral segment matching with an interval of 160ms, which is approximately a triphone interval, significantly improves the estimation accuracy than does single frame spectral matching.

ACKNOWLEDGMENT

The authors would like to thank Information Science Research Laboratory Executive Manager Dr Ken-ichiro Ishii and other members of the HONDA Research Group for many useful and helpful discussions.

REFERENCES

1. Atal, B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W. (1978). “Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique,” J. Acoust. Soc. Am. **63**, 1535-1555.
2. Schroeter, J. and Sondhi, M.M. (1992). “Speech coding based on physiological models of speech production,” in *Advances in Speech Signal Processing*, edited by S.Furui and M.M.Soondhi (Dekker, New York), pp.231-267.
3. Schroeter, J. and Sondhi, M.M. (1994). “Techniques for estimating vocal-tract shapes from the speech signal,” IEEE Trans. Speech Audio Process. **2**(1), 133-150.
4. Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., and Saltzman, E. (1996). “Accurate recovery of articulator positions from acoustics: New conclusions based on human data,” J. Acoust. Soc. Am. **100**, 1819-1834.
5. 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.