MULTI-SPEAKER ARTICULATORY RECONSTRUCTION BASED ON AN EIGEN ARTICULATORY HMM

Sadao HIROYA and Takemi MOCHIDA

NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-0198 Japan
e-mail: {hiroya, mochida}@idea.brl.ntt.co.jp

ABSTRACT

We present a multi-speaker articulatory reconstruction method based on speaker-independent articulatory features and speaker-dependent features. These features are separated by using a multi-speaker articulatory database. This separation method consists of normalizing palate positions among the speakers and separating multi-speaker articulatory data into a speaker-independent eigen articulatory HMM and a speaker-adaptive matrix by using speaker-adaptive training (SAT). With the proposed method, the average RMS errors of the measured and reconstructed articulatory parameters were 1.35 mm. This result shows that the proposed method makes it possible to control speaker idiosyncrasies in the articulatory parameter domain.

1. INTRODUCTION

Speech signal consists of speaker-independent phonological features and speaker-dependent features, including vocal-tract shape and length, speaking style, gender, and so on. In previous studies, these features have been separated in the speech spectrum domain by using a multi-speaker database [1, 2]. However, the separation was insufficient because of the effect of the complicated speech spectrum.

On the other hand, separating these features in the articulatory parameter domain is expected to be better than in the speech spectrum domain, because the articulatory parameters for a given phoneme are less variable than a speech spectrum. However, the measured articulatory parameters do not have a common axis among speakers, while the speech spectrum has a frequency scale. Hashi *et al.* [3] proposed a normalization method for point-parameterized articulatory data, but this method cannot reduce cross-speaker variance in the horizontal axis of the data and does not take into account the dynamic features of articulatory parameters.

In this study, we present a method of separating multi-speaker articulatory parameters into speaker-independent articulatory features and speaker-dependent features based on the normalization of the measured articulatory parameters among the speakers. This normalization process consists of two parts: normalization of palate positions among the speakers and the construction of a speaker-independent hidden Markov model (HMM) of articulatory parameters from multi-speaker articulatory data. To extract the speaker-independent features, we use a speaker-adaptive training (SAT) paradigm [1]. The SAT paradigm makes it possible to separate multi-speaker articulatory parameters into a speaker-independent eigen articulatory HMM and speaker-adaptive matrix. Then, the speaker-adapted articulatory HMM is obtained from these models and the articulatory parameters are reconstructed from the speaker-adapted HMM. We evaluate this method in terms of the RMS error between the measured and reconstructed articulatory parameters. Finally, we discuss the speech spectrum estimated from the reconstructed articulatory parameters.

2. DATA COLLECTION

Articulatory movements and speech acoustics data were obtained from simultaneous observations using the EMA system [4] and acoustic recordings of continuous speech utterances. The articulatory data were collected at a sampling rate of 250 Hz. The articulatory parameters were represented by the vertical and horizontal positions of six coils, which were placed on the upper and lower incisor, the upper and lower lips, and the tongue (three positions). At that time, the palate positions were measured. Speech signal was recorded at a sampling rate of 16 kHz. Twenty-five mel-cepstrum coefficients without the 0-th coefficient were obtained as acoustic parameters using a 32-ms Blackman window with a 4-ms frame. We made articulatory-acoustic recordings by 375 sentences (273,738 frames/subject, about 18 minutes) spoken at normal speed by five Japanese male subjects (speaker A to E) and designed an articulatory-acoustic pair codebook for respective speakers.

3. TRAINING PROCEDURE

The method for constructing the eigen articulatory HMM from multi-speaker articulatory data was performed as follows (Fig. 1). First, the articulatory data among the speakers was normalized by rotating the palate positions. Then, we constructed the average articulatory HMM using the above

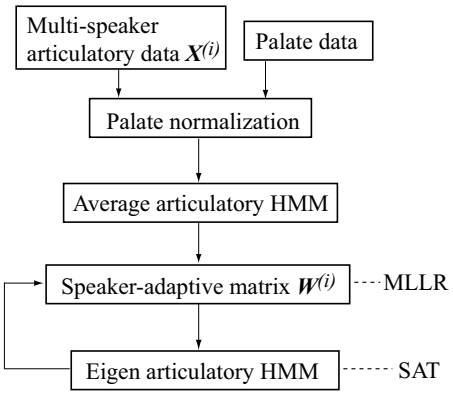

Fig. 1. Training procedure for eigen articulatory HMM.

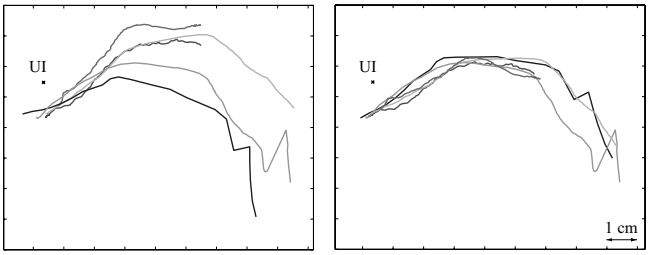

Fig. 2. Palate positions before and after normalization.

multi-speaker data. In the MLLR, the speaker-adaptive matrix was re-estimated in accordance with a standard EM algorithm. Then, in the SAT paradigm, the mean vectors and the covariance matrices of the Gaussian pdfs were re-estimated using the updated values of the speaker-adaptive matrices based on an extended EM algorithm. This re-estimation process was repeated until the increase of the likelihood converged.

3.1. PALATE NORMALIZATION

When the articulatory data was measured, a face direction varied among the speakers. Therefore, it was necessary to normalize the palate positions among the speakers. This procedure was conducted by rotating the palate positions for the position of upper incisor (UI). Rotate angle was determined by minimizing the error of palate positions among the speakers. Fig. 2 shows palate positions before and after normalization for five speakers. The articulatory data for respective speakers were normalized according to the obtained rotate angle.

3.2. AVERAGE ARTICULATORY HMM

The HMMs of articulatory parameters, called the articulatory HMM [5], has a sequence of states for each phoneme and generates an articulatory parameter vector in a probabilistic form for a given phoneme sequence. We estimated the initial parameters $\lambda_{ave} = \{\bar{\boldsymbol{x}}_m, \boldsymbol{\sigma}_m, a_{mn}\}$ of the HMM model so that the resulting model maximizes the likelihood of the training articulatory parameter sequences. Here, the $\bar{\boldsymbol{x}}_m$ and $\boldsymbol{\sigma}_m$ are the mean and covariance of the articulatory parameter vector at state m, and a_{mn} is the probability of the transition from state m to state n. Consider a training database that consists of articulatory parameters collected from I speakers, with each speaker i, contributing a transcribed observation sequence $\boldsymbol{X}^{(i)} = [\boldsymbol{x}_1^{(i)}, \cdots, \boldsymbol{x}_t^{(i)}, \cdots, \boldsymbol{x}_{T_i}^{(i)}]$ of length T_i. Here, we assume that articulatory parameter vector $\boldsymbol{x}_t^{(i)}$ consist of static parameters and their velocity and acceleration (dynamic). The initial model λ_{ave}, called the average articulatory HMM, is derived as

$$\lambda_{ave} = \underset{\lambda}{\operatorname{argmax}}\, P(\boldsymbol{X}|\lambda) = \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{I} P(\boldsymbol{X}^{(i)}|\lambda), \quad (1)$$

where $P(\boldsymbol{X}|\lambda)$ is the output probability of the observation sequence $\boldsymbol{X}^{(i)}$ given the existing set of models λ.

3.3. MLLR

In MLLR-based speaker adaptation [6], the adapted mean vector $\bar{\boldsymbol{x}}_m^{(i)}$ of state m of speaker i is estimated by

$$\bar{\boldsymbol{x}}_m^{(i)} = \boldsymbol{W}_s^{(i)} \boldsymbol{\xi}_m = \boldsymbol{A}_s^{(i)} \bar{\boldsymbol{x}}_m + \boldsymbol{b}_s^{(i)}, \quad (2)$$

where $\boldsymbol{\xi}_m = [1, \bar{\boldsymbol{x}}_m^\top]^\top$, and $\boldsymbol{W}_s^{(i)} = [\boldsymbol{b}_s^{(i)} \boldsymbol{A}_s^{(i)}]$ is the speaker-adaptive matrix for the mean vector. The superscript $(\cdot)^\top$ is the matrix transpose. We assume that $\boldsymbol{W}_s^{(i)}$ is shared by S states $\{s_1, \cdots, s_S\}$.

3.4. EIGEN ARTICULATORY HMM

We discuss the SAT paradigm [1] for the normalization of multi-speaker articulatory data. Using SAT, the eigen articulatory model, consisting of speaker-independent articulatory features, is trained so that the resultant model of the MLLR-based speaker adaptation maximizes the likelihood for respective training speakers. In the training procedure of the eigen articulatory model, the maximum likelihood estimation of the mean vectors $\bar{\boldsymbol{x}}_m$ and the covariance matrices $\bar{\boldsymbol{\sigma}}_m$ of the Gaussian pdfs in state m of speaker i for the training data are given by

$$\bar{\boldsymbol{x}}_m = \Big(\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_m^{(i)}(t) \boldsymbol{A}_s^{(i)\top} \boldsymbol{\sigma}_m^{-1} \boldsymbol{A}_s^{(i)} \Big)^{-1} \times \Big(\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_m^{(i)}(t) \boldsymbol{A}_s^{(i)\top} \boldsymbol{\sigma}_m^{-1} (\boldsymbol{x}_t^{(i)} - \boldsymbol{b}_s^{(i)}) \Big) \quad (3)$$

$$\boldsymbol{\sigma}_m = \frac{\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_m^{(i)}(t) (\boldsymbol{x}_t^{(i)} - \bar{\boldsymbol{x}}_m^{(i)}) (\boldsymbol{x}_t^{(i)} - \bar{\boldsymbol{x}}_m^{(i)})^\top}{\sum_{i=1}^{I} \sum_{t=1}^{T_i} \gamma_m^{(i)}(t)}, \quad (4)$$

Table 1. Phoneme labels.

a i u e o N Q k s t n h m y r w g z d b p f G
ky sh ch ny hy my ry gy jy by py ts

Table 2. Phoneme types used in the evaluation.

Vowel	a i u e o
Labial	m b p f my by py
Alveolar	s t n z d sh ch ny ry jy ts
Velar	k g ky gy
Semivowel	w r y

where $\gamma_m^{(i)}(t)$ is the probability that the observation vector $\boldsymbol{x}_t^{(i)}$ is generated in state m at time t.

4. ARTICULATORY RECONSTRUCTION

Fig. 3 shows the procedure for reconstructing articulatory parameters and estimating the speech spectrum. The speaker-adapted articulatory HMM is obtained from the eigen (or average) articulatory HMM and speaker-adaptive matrix $\boldsymbol{W}^{(i)}$. Then, from this HMM, articulatory parameter sequences of speaker i are reconstructed using the algorithm for parameter generation from HMMs with dynamic features [7], where phoneme durations are obtained from the results of Viterbi alignment for the measured articulatory parameters. Finally, for the reconstructed articulatory parameters, the speech spectrum was estimated by using an articulatory-acoustic pair codebook search method [8].

5. EXPERIMENTS

5.1. EXPERIMENTAL CONDITIONS

As training data, 375 sentences were used. The types of HMM were 3-state left to right biphone models (considering subsequent phoneme) with no skips. The $\boldsymbol{\sigma}_m$ had a diagonal covariance. Table 1 shows phoneme symbols used in the experiments. In the table, /N/ is a syllabic nasal, /G/ is a nasalized sound of /g/, and /Q/ is a glottal stop. In addition to these 35 phoneme symbols, we included silence for special symbols that represent the onset and release of utterances. The HMM was trained using decision-tree-based state clustering to define 739 states.

5.2. RESULTS

We evaluate the proposed method in terms of the RMS error between the reconstructed and measured articulatory parameters. Fig. 4 shows the RMS error of articulatory parameters reconstructed from average articulatory HMM and eigen articulatory HMM for each phoneme type. In this figure, 'Total (To)' is the total average RMS error and 'Vowel (Vo)', 'Semivowel (Sv)', and consonants are the average

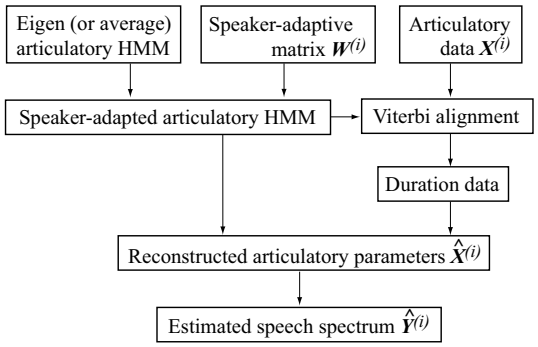

Fig. 3. Procedure for reconstructing articulatory parameters and estimating speech spectrum.

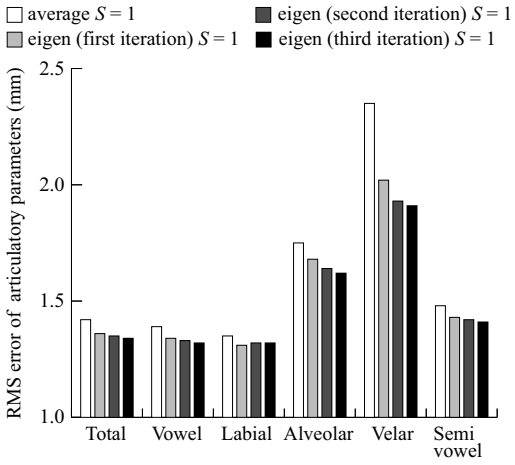

Fig. 4. RMS error for each phoneme type.

RMS errors for the phoneme types in Table 2 at articulation. The error was obtained for all articulatory positions for 'Vowel' and 'Semivowel', and for primary articulatory positions for consonants; the lip positions (UL and LL) for 'Labial (La)', the tongue tip position (T1) for 'Alveolar (Al)', and the tongue back (T3) for 'Velar (Ve)'.

For this experiment, 1 cluster was used. The RMS error of the articulatory parameters obtained from eigen articulatory HMM was smaller than those from average one for every phoneme types. In particular, the decrement of RMS error for velar consonants was larger than for the other phoneme types between average and eigen articulation. This suggests that the inter-subject variance of articulatory movements on velar consonants is larger than that for the other phoneme types. This is because the speaking tactics for velar consonants are different among the speakers due to the effect of a coordination between soft palate and tongue back. Another possible reason is that movements of the back of the tongue for velar consonants is independence of jaw movements. With all the RMS error for labial consonants slightly increased after the first iteration, the error saturated at the second iteration.

Table 3. Average RMS error for the number of speaker-adaptive matrices at second iteration.

phone. type	To	Vo	La	Al	Ve	Sv
$S = 1$ (mm)	1.35	1.33	1.32	1.64	1.93	1.42
$S = 6$ (mm)	1.31	1.29	1.30	1.60	1.80	1.38

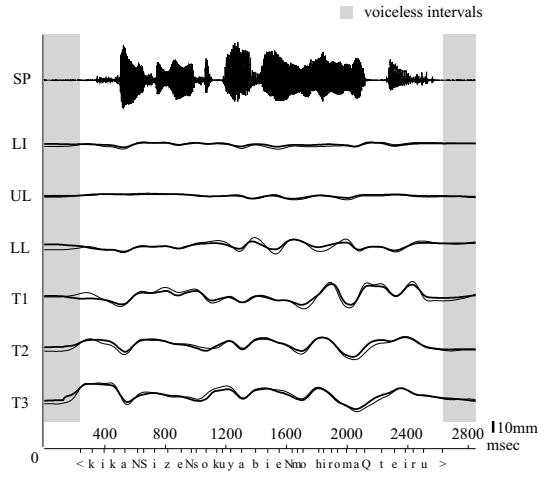

Fig. 5. Measured (thin lines) and reconstructed (thick lines) articulatory parameters of vertical positions of speaker B.

Table 3 shows the RMS error of articulatory parameters obtained from eigen articulatory HMM at second iteration for the number of speaker-adaptive matrices. 6 clusters were used. The RMS error decreased for every phoneme types as the number of speaker-adaptive matrices increased. However, the decrement of RMS error was small, except for velar consonants. Therefore, one speaker-adaptive matrix can well approximate speaker-dependent features of articulatory parameters. Fig. 5 shows an example of the reconstructed and measured articulatory parameters of vertical positions.

Fig. 6 shows a speech spectrum estimated from the reconstructed articulatory parameters of speaker A and B, respectively. For both speakers, the articulatory-acoustic pair codebook of speaker A was used. In other words, Fig. 6(b) shows speech spectrum sequences that speaker A produced according to the articulatory parameters of speaker B. This indicates that we were able to produce the speech spectrum with speaker-dependent features of articulation by controlling in the articulatory domain. Moreover, in the informal listening test, we were able to discriminate the differences among speakers.

6. CONCLUSIONS

We presented a multi-speaker articulatory reconstruction method based on the eigen articulatory HMM and speaker-adaptive matrix. These models were obtained by palate position normalization among speakers and the SAT paradigm. The average RMS error of the reconstructed articulatory parameters from eigen articulation was 1.35 mm. This result shows that this method is efficient for controlling speaker-dependent features of articulation.

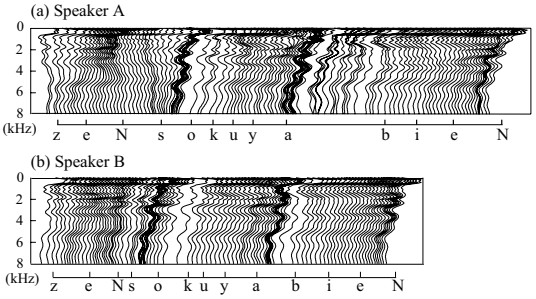

Fig. 6. Speech spectrum estimated from the reconstructed articulatory parameters of speaker A and B.

7. ACKNOWLEDGMENTS

The authors thank Dr. T. Moriya, Dr. M. Kashino, Dr. Y. Shiraki for many useful and helpful discussions, and H. Nakano of Kyushu University for help in programming and useful discussions.

8. REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[2] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, 2003.

[3] M. Hashi, J.R. Westbury, and K. Honda, "Vowel posture normalization," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2426–2437, 1998.

[4] T. Kaburagi and M. Honda, "Determination of sagittal tongue shape from the positions of points on the tongue surface," *J. Acoust. Soc. Am.*, vol. 96, no. 3, pp. 1356–1366, 1994.

[5] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Processing.*, vol. 12, no. 2, pp. 175–185, 2004.

[6] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[7] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. EUROSPEECH*, 1995, pp. 757–760.

[8] T. Kaburagi and M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proc. ICSLP*, 1998, pp. 433–436.