Acoustic-to-articulatory inversion using a speaker-normalized HMM-based speech production model

Sadao Hiroya

NTT Communication Science Laboratories, NTT Corporation, Japan
Department of Cognitive and Neural Systems, Boston University, United States
E-mail: hiroya@idea.brl.ntt.co.jp

Abstract

Acoustic-to-articulatory inverse mapping is a difficult problem because of its non-linear and one-to-many characteristics. We have previously developed a speech inversion method using a hidden Markov model (HMM)-based speech production model which takes into account the phoneme-specific dynamic constraints of articulatory parameters. We found that the constraint significantly decreases the estimation error of articulatory parameters. However, the model was trained for each speaker and articulatory parameters were estimated in a speaker-dependent manner. In this study, we present a speaker-normalized HMM-based speech production model which is constructed from a multi-speaker articulatory-acoustic database, and estimate articulatory parameters from multi-speakers' speech signals using the model. Result shows that the estimation error of articulatory parameters for vowels is about 1.0 mm.

1 Introduction

The acoustic-to-articulatory inverse mapping is characterized by one-to-many mapping [2]. One of the first studies to uniquely determine articulatory parameters from a speech signal by using continuity constraints on articulatory trajectories was done by Schroeter and Sondhi [6]. Moreover, to take into account the more reliable dynamic constraint based on articulatory measurements, a segmental acoustic and articulatory pair codebook [16], hypercube codebook [14], an extended Kalman filter [9], self-organizing hidden Markov model (SOHMMs) [15] and our own hidden Markov model (HMM)-based speech production model [10] have been proposed for speech inversion. However, the model was trained for each speaker and articulatory parameters were estimated in a speaker-dependent manner. The estimation method of articulatory movements from an arbitrary speaker's speech signal has practical applications in speech training and foreign language learning. Dusan and Deng performed speech inversion for unknown speakers by compensating for their vocal-tract length [8]. They suggested that the main source of inter-speaker variability in speech is geometrical differences in vocal-tract length. However, vocal-tract shape measurements showed that the need of non-uniform scaling along the vocal-tract length to normalize the variability in speech [3]. Therefore, the variability in speech cannot be sufficiently normalized only by adjusting vocal-tract length. To overcome the problem, we have proposed a speaker-adaptation method that statistically adapts the speech spectrum of a reference model to that of the unknown speaker by taking into account the dynamic constraints of articulatory parameters [11]. In this method, we assumed that the unknown speaker's dynamical constraints on articulatory parameters are the same as the reference speaker's. However, the estimation accuracy for the speaker-adapted model was half that of the speaker-dependent model. This was because there are differences in the dynamic constraints of articulatory parameters among the unknown and reference speakers. Figure 1 shows articulatory trajectories of the vowel interval [æ] of the word /pap/ for nine male speakers; the trajectories vary among the speakers (for details see also Sec. 2). The difference among the speakers was relatively large for

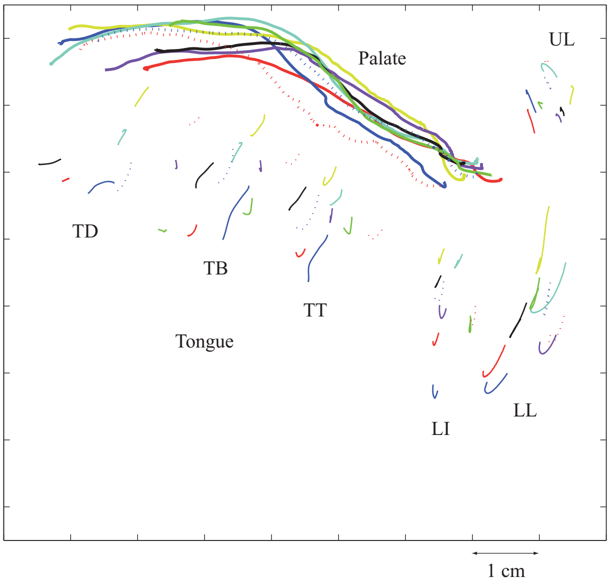

Figure 1: *Articulatory trajectories of the vowel [æ] and palate shapes for nine male speakers.*

TD. Therefore, to improve estimation accuracy, a speaker-independent phoneme-specific articulatory dynamical model is required.

2 Data collection

Articulatory parameters and speech signal data were obtained from simultaneous recordings using the MIT EMMA system [5] and from audio signals of continuous speech utterances (Fig. 2). In articulatory-acoustic recordings, nine male and eight female English native speakers read about 330 English sentences with clear and normal speaking: 'Say CVC (e.g. pip) for us' and 'Say CVC CVC (e.g. keep PECK) for us,' where C = [p, t, k] and V = [æ, ɑ, e, ɪ, i, u]. The articulatory data and the palate positions were collected at a sampling rate of 500 Hz and down-sampled to 250 Hz. The articulatory parameters were represented by the vertical and horizontal positions of six coils, which were placed on the lower incisor (LI), the upper and lower lips (LL, UL) and the tongue (TT, TB, TD: three positions). The speech signal was recorded at a sampling rate of 16 kHz. The first three formants (F1-F3), their bandwidths (B1-B3) and fundamental frequency (F0) for the vowel intervals of CVCs were the acoustic parameters, each with 250 Hz sampling rate. Vowel intervals were manually marked by experts.

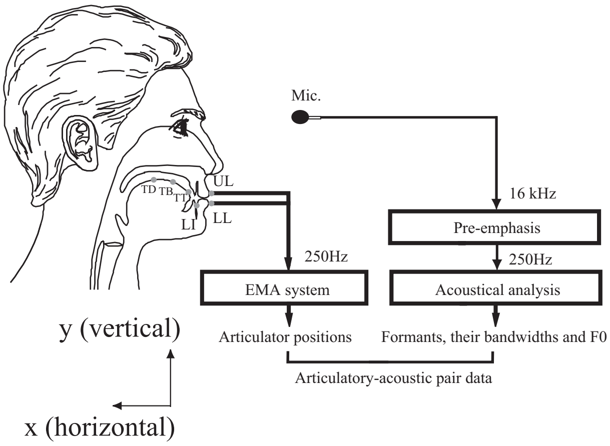

Figure 2: *Simultaneous recordings to obtain articulatory-acoustic data.*

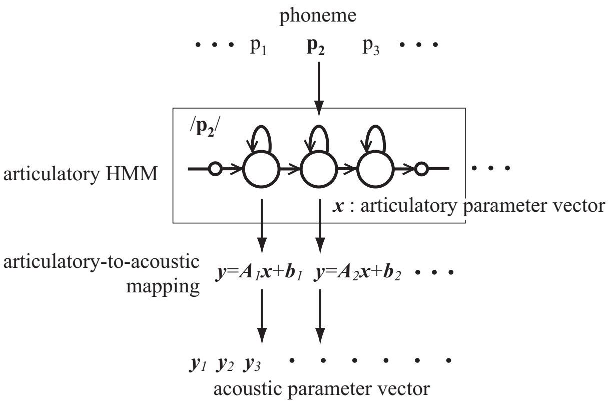

Figure 3: *HMM-based speech production model.*

3 HMM-based speech production model

The HMM-based speech production model consists of HMMs that represent the articulatory parameters for each phoneme, called the articulatory HMM, and an articulatory-to-acoustic mapping that transforms the articulatory parameters into acoustic parameters for each HMM state (Fig. 3). In the model, the linear function $\boldsymbol{y}_t = \boldsymbol{A}_j\boldsymbol{x}_t + \boldsymbol{b}_j$ was assigned to each HMM state j to approximate the articulatory-to-acoustic mapping $\boldsymbol{y}_t = f(\boldsymbol{x}_t)$ in a piecewise linear form. We denote the articulatory parameter vector sequence as $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_L^\top]^\top$ and the acoustic parameter vector sequence as $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_L^\top]^\top$, where we assume that articulatory parameter vector $\boldsymbol{x}_t$ and acoustic parameter vector $\boldsymbol{y}_t$ consist of static parameters and their velocity and acceleration. The superscript $(\cdot)^\top$ is the matrix transpose, and L the length of the observation sequence. The output probability of an acoustic parameter vector sequence in the HMM-based speech production model is

$$P(\boldsymbol{y}|\lambda) = \sum_{\boldsymbol{q}} \int P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{q}, \lambda) P(\boldsymbol{x}|\boldsymbol{q}, \lambda) P(\boldsymbol{q}|\lambda) d\boldsymbol{x}.$$

Here, $\boldsymbol{q} = (q_1, \cdots, q_L)$ is the HMM state sequence, and λ represents the phoneme-specific models. $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{q}, \lambda)$ is the occurrence probability of an acoustic parameter vector sequence for a given articulatory parameter vector sequence, and $P(\boldsymbol{x}|\boldsymbol{q}, \lambda)$ is the output probability in the articulatory HMM.

Using the model, we presented a method of maximum a-posteriori (MAP) estimation of articulatory parameters using dynamic features for given acoustic parameters: The estimated articulatory parameters are

$$\begin{aligned}\hat{\boldsymbol{x}}_p = (&\boldsymbol{R}^\top \boldsymbol{\sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{R} + \boldsymbol{R}^\top \boldsymbol{A}^\top \boldsymbol{\sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{A} \boldsymbol{R})^{-1} \\ &\times (\boldsymbol{R}^\top \boldsymbol{\sigma}_{\boldsymbol{x}}^{-1} \bar{\boldsymbol{x}} + \boldsymbol{R}^\top \boldsymbol{A}^\top \boldsymbol{\sigma}_{\boldsymbol{w}}^{-1} (\boldsymbol{y} - \boldsymbol{b})),\end{aligned}$$

where $\boldsymbol{R}$ is the transformation from the static articulatory parameter vector sequence $\boldsymbol{x}_p = [\boldsymbol{x}_{p1}^\top, \cdots, \boldsymbol{x}_{pt}^\top, \cdots, \boldsymbol{x}_{pL}^\top]^\top$ to the articulatory parameter vector sequence $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_L^\top]^\top$. $\bar{\boldsymbol{x}}$ and $\boldsymbol{\sigma}_{\boldsymbol{x}}$ are the mean and covariance of the articulatory parameter vector, and $\boldsymbol{\sigma}_{\boldsymbol{w}}$ is the covariance of the error in the linear approximation of the articulatory-to-acoustic mapping. Our previous study showed that the RMS error between the measured and estimated articulatory parameters for sentence utterances was 1.5 mm for three-state HMMs and showed statistically significant difference from that for one-state HMMs ($p < 0.01$). This indicates that the dynamical constraints on the basis of articulatory movements efficiently reduce redundancy in acoustic-to-articulatory inverse mapping.

4 Proposed model

Figure 4 shows the method for constructing the speaker-independent phoneme-specific HMM-based speech production model. First, by using the multi-speaker articulatory and acoustic data, we normalize articulatory and acoustic parameters using speaker-adaptive training (SAT) [12], respectively. Then, by using these normalized articulatory and acoustic data, we create a gender-specific speaker-normalized HMM-based speech production model.

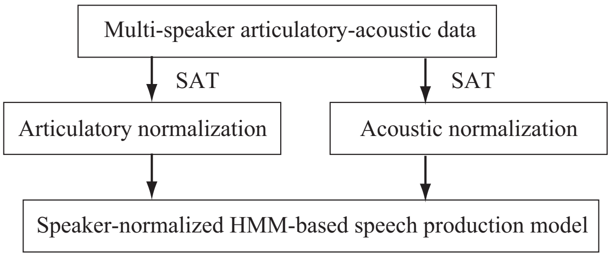
Multi-speaker articulatory-acoustic data
SAT
SAT
Articulatory normalization
Acoustic normalization
Speaker-normalized HMM-based speech production model

Figure 4: *Algorithm for speaker-normalized models.*

The SAT paradigm makes it possible to statistically obtain the speaker-independent articulatory HMMs and the speaker-adaptive matrix for individual speakers from the multi-speaker data. Our articulatory trajectory formation study has demonstrated that the estimation error of articulatory parameters based on this model is 1.29 mm for sentence utterances, not significantly different from that of the speaker-dependent model ($p = 0.02$) [12]. In particular, the RMS error for the tongue back with SAT was smaller than that without SAT. Figures 5 and 6 shows speaker-independent articulatory trajectories for six vowels for male and female speakers, respectively. These show that there are differences in articulatory dynamics between genders. We also conducted a SAT procedure for acoustic parameters (Fig. 7). It appeared that the differences in formant dynamics between genders are smaller than articulatory differences. This indicates that the speaking tactics in vowels differ between genders due to the effect of palate shape and vocal-tract length. This is related to the study of Simpson [1]: speakers' vowel articulations adapt to the form to their respective palates between genders during the acquisition of speech, but a more detailed analysis is required.

5 Experimental conditions

The types of HMMs were left to right ones with no skips. In the experiments, three-state monophone HMMs were used. Each state was composed of single Gaussian distribution. One speaker-adaptive matrix for each speaker for both articulatory and acoustic parameter vectors were used. As training data, 5035 and 4365 vowels for male and female subjects were used, respectively. We estimated six articu-

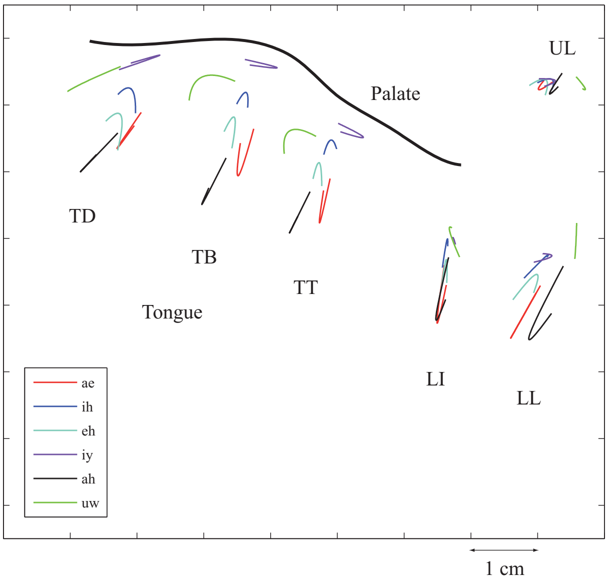

Figure 5: *Speaker-normalized articulatory parameters for each vowel for males.*

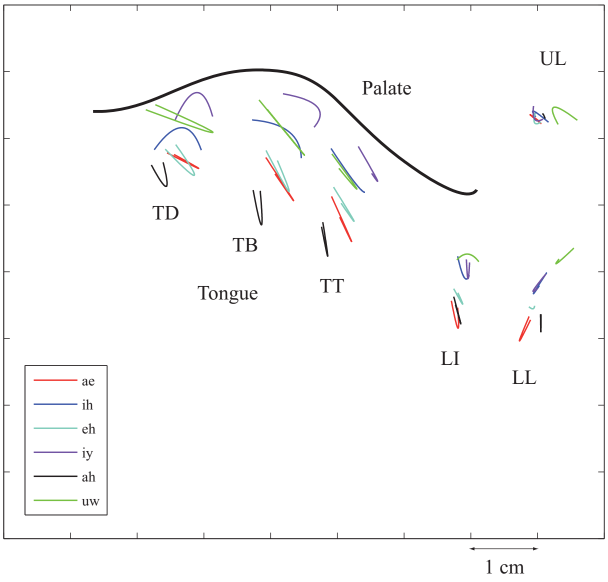

Figure 6: *Speaker-normalized articulatory parameters for each vowel for females.*

latory positions from the first three formants, their bandwidths and F0 using the proposed model. we applied a MAP estimate of articulatory parameters with dynamic features using a segment length 164 msec [10]. We evaluated the proposed method in terms of the RMS error between the measured and estimated articulatory parameters from acoustic parameters. The RMS error was evaluated in the vowel intervals and averaged for every speaker. All test vowels were included in the training ones.

6 Results

Figure 8 shows the measured and estimated articulatory movements for vowel [æ]. Figure 9 shows the RMS error of the articulatory parameters. The average error for all articulators was 1.03 mm. The maximum average error was 5.32 mm for males and 3.31 mm for females, respectively. Figure 10 shows the RMS error of the articulatory parameters for each vowel. The error for high vowels is smaller than that of low vowels. Figure 11 shows the RMS error of the articulatory parameters for each speaker. The error was less than 1.5 mm for every speaker. These results indicate that the speaker-independent phoneme-specific articulatory constraints efficiently decrease the estimation error for speech inversion.

7 Discussions

We have previously found that the estimation error of articulatory parameters from both formants and F0 based on the codebook search method [6] was smaller than that from formants only [13]. This was evidence that doesn't fit the source-filter theory for speech production [4]. Based on the finding, this study added a F0 to acoustic parameters but we should quantitatively evaluate the F0 effect for speech inversion using our dynamical models.

We also have found that context-dependent phonetic HMMs and phonemic information in an utterance significantly decreased the error ($p < 0.01$) [10]. However, vowel dynamical models in this study didn't take into account the preceding and subsequent consonants in CVCs (Figs. 5, 6 and 7). These vowel dynamics may vary depending on the consonants by coarticulation effect, so it is expected that the use of the context-dependent models will decrease the error of articulatory parameters.

Previously, a geometrical-based normalization method for articulatory parameters has been proposed [7] but Simpson pointed out that the method cannot reduce cross-speaker variance along the long axis of the vocal tract [1]. Thus, we expect that our normalized speech production model is useful for investigating the phoneme-specific invariant features

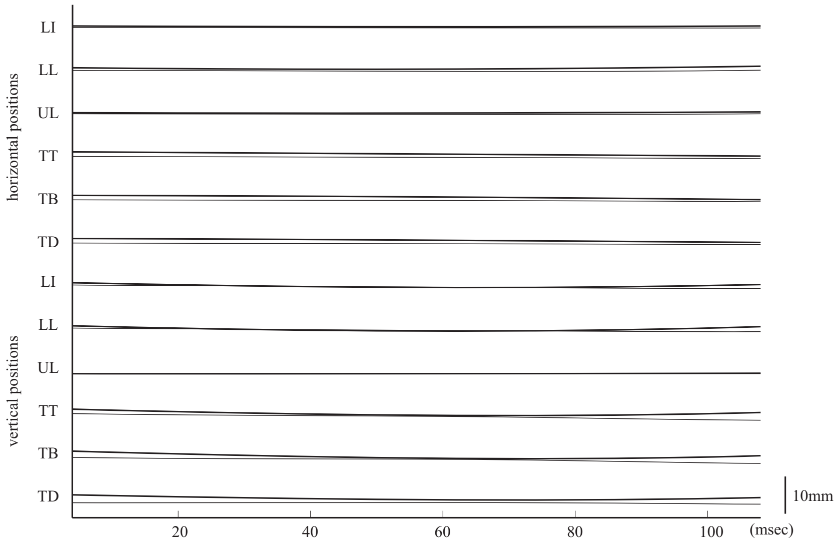

Figure 8: *Measured (thin lines) and estimated (thick lines) articulatory parameters for horizontal and vertical positions of vowel [æ] for speaker M01.*

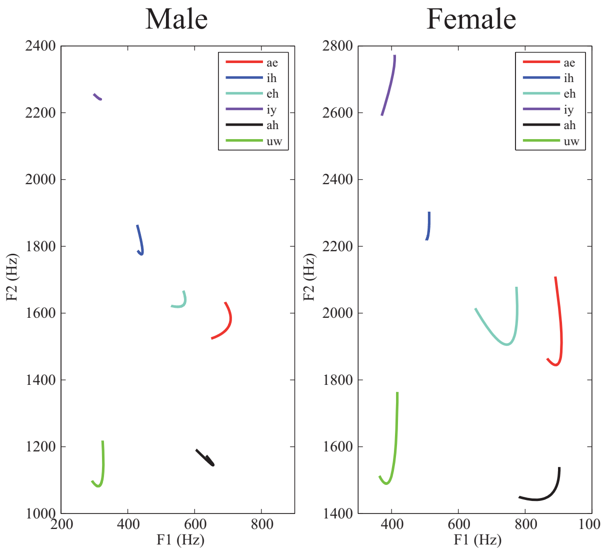

Figure 7: *Speaker-normalized formant trajectories for each vowel for each gender.*

included in speech signals in terms of articulatory parameters and the differences between genders.

8 Conclusions

This study demonstrated the small articulatory parameter estimation error obtained using a speaker-normalized articulatory dynamical model. We plan to develop an application for speech training.

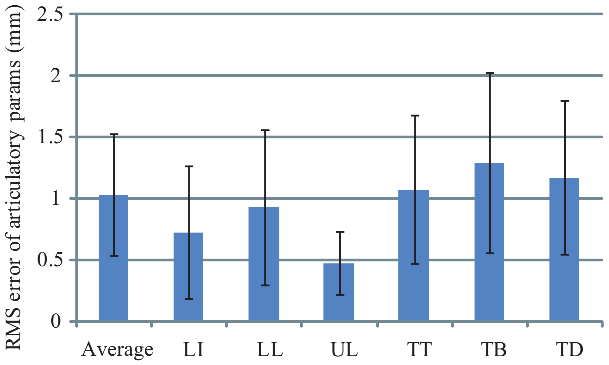

Figure 9: *RMS error of articulatory parameters for each articulator. Error bars indicate the standard deviation of the mean.*

Acknowledgement

I thank Prof. Joseph Perkell of MIT for permitting me to use the articulatory data, and Prof. Frank Guenther of Boston University, Dr. Miwako Hisagi and Mark Tiede of MIT for valuable discussions. This work was supported by NIH grants R01-DC03007 and R01-DC01925.

References

[1] A. P. Simpson. Dynamic consequences of differences in male and female vocal tract dimensions. *J. Acoust. Soc. Am.*, 109(5):2153–2164, 2001.

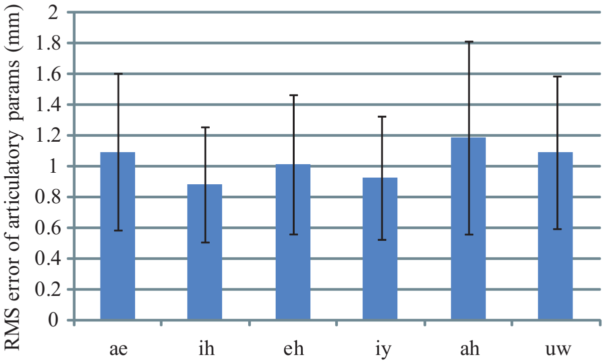

Figure 10: *RMS error of articulatory parameters for each vowel. Error bars indicate the standard deviation of the mean.*

[2] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computersorting technique. *J. Acoust. Soc. Am.*, 63(5):1535–1555, 1978.

[3] C. S. Yang and H. Kasuya. Speaker individualities of vocal tract shapes of Japanese vowels measured by magnetic resonance images. In *ICSLP*, pages 949–952, 1996.

[4] G. Fant. *Acoustic theory of speech production*. Monton & Co.'s-Gravenhage, 1960.

[5] J. Perkell, M. Cohen, M. Svirsky, M. Mathies, I. Garabieta, and M. Jackson. Electromagnetic midsagittal articulometer system for transducing speech articulatory movements. *J. Acoust. Soc. Am.*, 92(6):3078–3096, 1992.

[6] J. Schroeter and M. M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. SAP*, 2(1):133–150, 1994.

[7] M. Hashi, J. R. Westbury, and K. Honda. Vowel posture normalization. *J. Acoust. Soc. Am.*, 104(4):2426–2437, 1998.

[8] S. Dusan and L. Deng. Vocal-tract length normalization for acoustic-to-articulatory mapping using neural networks. *J. Acoust. Soc. Am.*, 106(4):2181, 1999.

[9] S. Dusan and L. Deng. Acoustic-to-articulatory inversion using dynamical and phonological constraints. In *5th Seminar on Speech Production*, pages 237–240, 2000.

[10] S. Hiroya and M. Honda. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Trans. SAP*, 12(2):175–185, 2004.

[11] S. Hiroya and M. Honda. Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model. *IEICE Trans. Inf. & Syst.*, E87-D(5):1071–1078, 2004.

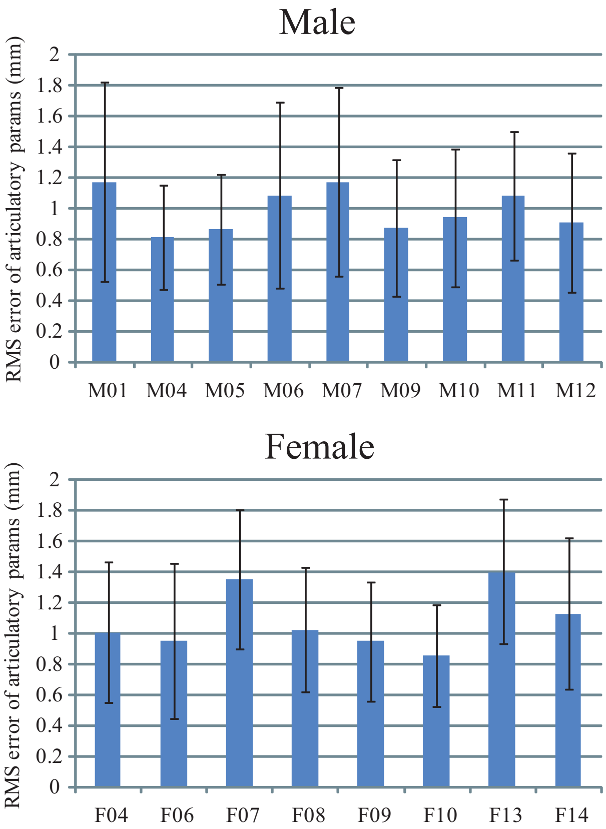

Figure 11: *RMS error of articulatory parameters for each subject. Error bars indicate the standard deviation of the mean.*

[12] S. Hiroya and T. Mochida. Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs. *Speech Comm.*, 48(12):1677–1690, 2006.

[13] S. Hiroya, T. Mochida, and M. Kashino. Reducing redundancy in acoustic-to-articulatory inversion by fundamental frequency. In *From sound to sense: 50+ years of discoveries in speech communication*, page 19, 2004.

[14] S. Ouni and Y. Laprie. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *J. Acoust. Soc. Am.*, 118(1):444–460, 2005.

[15] S. Roweis. Constraints hidden Markov models. In *NIPS*, pages 782–788, 1999.

[16] S. Suzuki, T. Okadome, and M. Honda. Determination of articulatory movement from speech acoustics using acoustic-articulatory codebook. *IEICE Trans. Fundamentals*, J85-A(8):840–846, 2002.