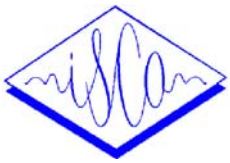

5th ISCA Speech Synthesis Workshop
Pittsburgh, PA, USA
June 14-16, 2004

MAPPING FROM ARTICULATORY MOVEMENTS TO VOCAL TRACT SPECTRUM WITH GAUSSIAN MIXTURE MODEL FOR ARTICULATORY SPEECH SYNTHESIS

Tomoki Toda[†‡], *Alan W Black*[†], *and Keiichi Tokuda*[‡]

[†]Language Technologies Institute, Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA 15213 USA
[‡]Graduate School of Engineering, Nagoya Institute of Technology Gokiso-cho, Showa-ku, Nagoya-shi, Aichi, 466-8555 Japan

ABSTRACT

This paper describes a method for determining the vocal tract spectrum from articulatory movements using a Gaussian Mixture Model (GMM) to synthesize speech with articulatory information. The GMM on joint probability density of articulatory parameters and acoustic spectral parameters is trained using a parallel acoustic-articulatory speech database. We evaluate the performance of the GMM-based mapping by a spectral distortion measure. Experimental results demonstrate that the distortion can be reduced by using not only the articulatory parameters of the vocal tract but also power and voicing information as input features. Moreover, in order to determine the best mapping, we apply maximum likelihood estimation (MLE) to the GMM-based mapping method. Experimental results show that MLE using both static and dynamic features can improve the mapping accuracy compared with the conventional GMM-based mapping.

1. INTRODUCTION

Current methods of unit selection [1] and concatenative synthesis [2] have dramatically improved the naturalness of synthetic speech. These approaches make it possible to use Text-to-Speech (TTS) more widely [3]; however, they are not appropriate for flexibly synthesizing various voices. In principle, only speech segments included in the corpus can be used for synthesizing speech. Therefore, in order to synthesize various types of speech (e.g. speech of multiple speakers, emotional speech and other speaking styles), representative speech samples need to be recorded in advance. Thus, large-sized speech corpora are needed to synthesize speech with sufficient naturalness.

Many attempts at synthesizing speech based on speech production mechanisms ignored in concatenative synthesis have been studied for several decades. The speech signal is generated from articulatory parameters by a mathematical production model. Speech is characterized not by the properties of the speech acoustics but by the properties of the vocal apparatus in this framework. Slowly varying articulatory parameters are better candidates for speech coding [4]. Moreover, the speech signal can be modified in an understandable way by manipulating articulatory parameters rather than acoustic parameters such as the vocal tract spectrum. Articulatory modeling having these advantages is also applied to TTS [5]. However, synthetic speech quality is generally degraded since the speech production mechanism is too complex to be mathematically modeled without some approximations.

Some corpus-based approaches for estimating the vocal tract spectrum from articulatory parameters have recently been studied. Instead of mathematically representing the production mechanism, correspondence of a configuration of articulatory parameters to the vocal tract spectrum is statistically extracted from a parallel acoustic-articulatory speech database. One of such methods is to use an acoustic-articulatory codebook. Shiga and King [6] proposed a method for statistically estimating the vocal tract spectrum from harmonic spectra in multiple frames having similar articulatory configurations. Kaburagi and Honda [7] reported that the estimation accuracy of the spectrum can be improved by using phonetic information as well as articulatory information. Hiroya and Honda [8][9][10] extended the codebook approach to statistically modeling a space of articulatory and acoustic parameters with HMMs, which is called a speech production model. In this model, the acoustic-articulatory correspondence is represented as a linear mapping in each state of the diphone HMMs.

We also focus on speech synthesis from articulatory parameters to discover a more flexible framework than that of concatenative speech synthesis. An approach in which phonetic information is not always needed has many advantages, e.g. to enable speech modification that is independent of languages. Therefore, we address the problem of spectral determination from articulatory parameters without phonetic information. Articulatory parameters are converted into the acoustic spectrum using a mapping algorithm with a Gaussian Mixture Model (GMM) proposed by Stylianou [11], which is often used for voice conversion. We investigate the effectiveness of using some features on source information and taking into account the correlation between frames for the spectral determination. The MOCHA database [12] is used as acoustic-articulatory data in this paper.

The paper is organized as follows. In **Section 2**, we introduce the MOCHA database. In **Section 3**, the GMM-based mapping method and evaluations of its performance using some input features are described. In **Section 4**, we apply maximum likelihood spectral estimation using dynamic features to the GMM-based mapping. In **Section 5**, speech synthesis with the estimated spectral sequence is described. Finally, we summarize this paper in **Section 6**.

2. ACOUSTIC-ARTICULATORY SPEECH DATABASE: MOCHA

The Multichannel Articulatory database (MOCHA) [12] is available from the Centre for Speech Technology Research, University of Edinburgh. The MOCHA database consists of speech and some articulatory movements simultaneously recorded at Queen Margaret University College.

Acoustic-articulatory data of two speakers is used. One is female (fsew0), and the other is male (msak0). The 460 British

TIMIT sentences are uttered by each speaker. Speech data is recorded at 16 kHz sampling.

We use electromagnetic articulograph (EMA) data, one of representations of articulatory data provided in MOCHA, as an articulatory parameter. The movement of seven articulators (top lip, bottom lip, bottom incisor, tongue tip, tongue body, tongue dorsum, and velum) and two reference points (the bridge of the nose and the upper incisor) are sampled in the midsagittal plane at 500 Hz. Each articulatory location is shown by x- and y-coordinates. We performed a normalization process described in [13] for reducing the effect of noise resulting from measurement error. The 14-dimensional articulatory feature vector converted to Z-score is used in this paper.

3. GMM-BASED MAPPING FROM ARTICULATORY MOVEMENTS TO VOCAL TRACT SPECTRUM

3.1. GMM-based mapping

In the GMM-based mapping algorithm [11], a mapping function from an articulatory feature vector $\boldsymbol{x}_t$ to a spectral feature vector $\boldsymbol{y}_t$ in frame t is defined as

$$\hat{\boldsymbol{y}}_t = \sum_{i=1}^{M} p(m_i|\boldsymbol{x}_t, \Theta) \boldsymbol{E}(\boldsymbol{y}_t|\boldsymbol{x}_t, m_i, \Theta), \quad (1)$$

$$\boldsymbol{E}(\boldsymbol{y}_t|\boldsymbol{x}_t, m_i, \Theta) = \boldsymbol{\mu}_i^{(y)} + \boldsymbol{\Sigma}_i^{(yx)} {\boldsymbol{\Sigma}_i^{(xx)}}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_i^{(x)}), \quad (2)$$

$$p(m_i|\boldsymbol{x}_t, \Theta) = \frac{w_i N(\boldsymbol{x}_t; \boldsymbol{\mu}_i^{(x)}, \boldsymbol{\Sigma}_i^{(xx)})}{\sum_{j=1}^{M} w_j N(\boldsymbol{x}_t; \boldsymbol{\mu}_j^{(x)}, \boldsymbol{\Sigma}_j^{(xx)})}, \quad (3)$$

where $\hat{\boldsymbol{y}}_t$ denotes an estimated spectral feature vector. w_i denotes a weight of the i-th mixture, and M denotes the total number of mixtures. $\boldsymbol{\mu}_i^{(x)}$ and $\boldsymbol{\mu}_i^{(y)}$ denote the mean vector of the i-th mixture for $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. $\boldsymbol{\Sigma}_i^{(xx)}$ and $\boldsymbol{\Sigma}_i^{(yx)}$ denote the covariance matrix of the i-th mixture for $\boldsymbol{x}$ and the cross-covariance matrix of the i-th mixture for $\boldsymbol{x}$ and $\boldsymbol{y}$. $N(\boldsymbol{x}_t; \boldsymbol{\mu}_i^{(x)}, \boldsymbol{\Sigma}_i^{(xx)})$ denotes the normal distribution with $\boldsymbol{\mu}_i^{(x)}$ and $\boldsymbol{\Sigma}_i^{(xx)}$. Θ denotes a set of model parameters, i.e. weights, mean vectors and covariance matrices.

Kain [14] proposed joint density estimation as a robust method for estimating model parameters compared with a least squares estimation especially in cases of small amounts of training data. In this method, the following GMM on a joint vector $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ is trained with the EM algorithm.

$$p(\boldsymbol{z}|\Theta) = \sum_{i=1}^{M} w_i N(\boldsymbol{z}; \boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma}_i^{(z)}), \quad (4)$$

$$\boldsymbol{\Sigma}_i^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(xx)} & \boldsymbol{\Sigma}_i^{(xy)} \\ \boldsymbol{\Sigma}_i^{(yx)} & \boldsymbol{\Sigma}_i^{(yy)} \end{bmatrix}, \quad \boldsymbol{\mu}_i^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(x)} \\ \boldsymbol{\mu}_i^{(y)} \end{bmatrix}. \quad (5)$$

It is noted that the covariance matrices are full because we perform a mapping between the different features, i.e. from the articulatory parameter to the spectral parameter.

3.2. Evaluation of mapping accuracy

In order to measure the accuracy of the GMM-based articulatory-acoustic mapping, we perform experimental evaluations using a spectral distortion

3.2.1. Features

We use mel-cepstrum as a feature for representing the vocal tract spectrum. A spectrum is calculated with STRAIGHT analysis [15], a high-quality method based on pitch-adaptive analysis with an interpolation in the time-frequency region, and is then converted to 25-dimensional minimum phase mel-cepstrum. The 1-st through 24-th mel-cepstral coefficients are used as the spectral features. The shift length is 5 ms.

The 14-dimensional EMA data (downsampled to match the 5 ms shift rate) is used as an articulatory feature. Since the EMA data captures only vocal tract characteristics, it cannot capture source characteristics that also affect the acoustic spectrum. Therefore, we use some features on source information as well as the EMA data. As a feature for voicing information, we use an unvoiced/voiced decision binary feature, "UV", or a log-scaled F_0, "F_0", which also includes U/V information. F_0 is automatically extracted with fixed-point analysis [16] in STRAIGHT. As a power information feature, we use the 0-th mel-cepstral coefficient showing a power of a log-scaled amplitude spectrum, "CPow", or a log-scaled power of a linear amplitude spectrum, "Pow".

3.2.2. Experimental conditions

The acoustic-articulatory data of the two speakers described in **Section 2** was used. For each speaker, we used 414 sentences for training parameters of the mapping function and 46 sentences not included in the training data for evaluation. Silence frames were removed using phonetic segmentation information included in MOCHA. The number of frames in the training data was 223,480 for the female speaker (fsew0) and 188,988 for the male speaker (msak0). The mel-cepstral distortion between the target and the estimated mel-cepstra given by the following equation was used as the evaluation measure,

$$\text{Mel-CD} = 10/\ln 10 \sqrt{2 \sum_{d=1}^{24} \left(mc_d^{(t)} - mc_d^{(e)} \right)^2}, \quad (6)$$

where $mc_d^{(t)}$ and $mc_d^{(e)}$ denote the d-th coefficient of the target and the estimated mel-cepstra, respectively. Evaluations were performed for some combinations of input features. In each case, the number of mixtures was varied from 1 to 256.

3.2.3. Experimental results

Mel-cepstral distortion as a function of the number of mixtures is shown in **Fig. 1**. The distortion decreases as the number of mixtures increases although using too many mixtures causes degradation of the performance due to over-training.

Average and standard deviation of mel-cepstral distortions when using the optimum number of mixtures in each combination of input features are shown in **Table 1**. Results for the combination of features, "EMA+UV+Pow", which is not shown in **Fig. 1**, are also shown in this table. As for voicing information, the log-scaled F_0 can slightly reduce the distortion compared with the simple U/V parameters. For power information, the log-scaled power of the linear spectrum is more effective than the power of the log-scaled spectrum. It can be seen that the mapping accuracy can be dramatically improved by using not only the articulatory EMA data but also the log-scaled power of linear spectrum, and further improvement can be achieved by also using the log-scaled F_0.

The mel-cepstral distortion calculated in each phonetic category for the female speaker is shown in **Fig. 2**. Using power information causes a decrease of distortion in all phonetic categories.

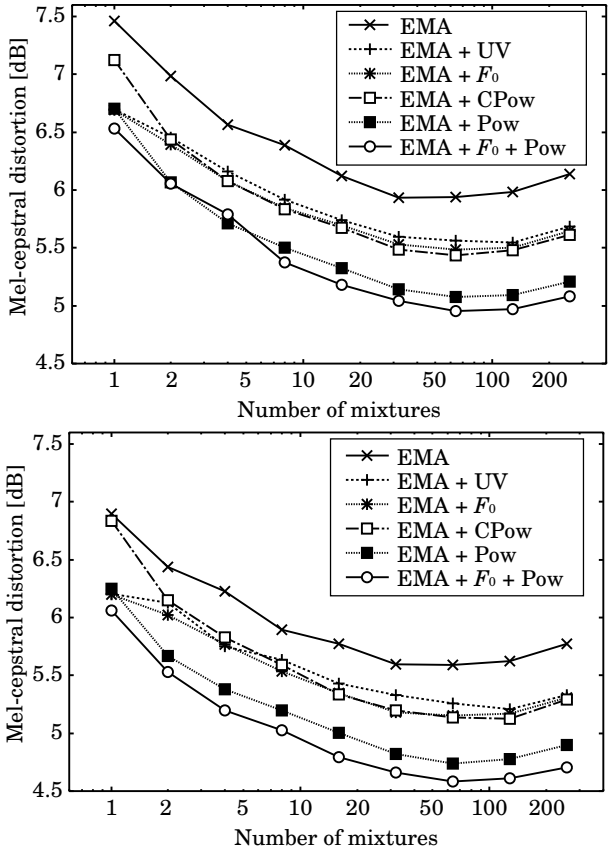

Fig. 1. Mel-cepstral distortion as a function of the number of mixtures. The upper figure shows the result for the female speaker, and the lower figure shows the result for the male speaker.

The mapping accuracy in some phonetic categories consisting of voiced and unvoiced consonants can be improved by further using voicing information. It is observed that the mapping to spectrum for fricatives and affricates is more difficult than for other phonemes. We can also see a tendency that the mapping accuracy of longer vowels is better than that of shorter vowels. These results are also observed for the male speaker.

The mel-cepstral distortion for the female speaker is obviously worse than for the male speaker. In general, it is harder to estimate the acoustic spectrum of speech uttered by a high fundamental frequency compared with a low fundamental frequency because spectral information in frequency bands between F_0 harmonics disappears. Therefore, variance of the acoustic spectrum regarded as noise increases as F_0 rises. **Fig. 3** shows frequency distributions of the mel-cepstral distortion in voiced and unvoiced frames. The average F_0 is 206 Hz in the female voice and 115 Hz in the male voice, respectively. In the voiced frames, the distribution for the male speaker is obviously shifted to the lower distortion compared with that for the female speaker. Whereas, such a shift of the distribution is not observed in the unvoiced frames although the spread of the distribution for the male speaker is smaller than for the female speaker. These results show that the influence of F_0 on spectral estimation is one of factors causing the difference of the mapping accuracy between two speakers.

Table 1. Average and standard deviation of mel-cepstral distortion [dB] in each combination of input features. The number in each bracket shows the optimum number of mixtures

	Female	Male
EMA	5.93 ± 2.45 (32)	5.59 ± 2.23 (64)
EMA+UV	5.55 ± 2.17 (128)	5.21 ± 1.94 (128)
EMA+F_0	5.48 ± 2.18 (64)	5.15 ± 2.00 (64)
EMA+CPow	5.44 ± 2.43 (64)	5.12 ± 2.10 (128)
EMA+Pow	5.08 ± 1.94 (64)	4.74 ± 1.77 (64)
EMA+UV+Pow	5.01 ± 1.89 (64)	4.62 ± 1.62 (128)
EMA+F_0+Pow	4.96 ± 1.86 (64)	4.59 ± 1.61 (64)

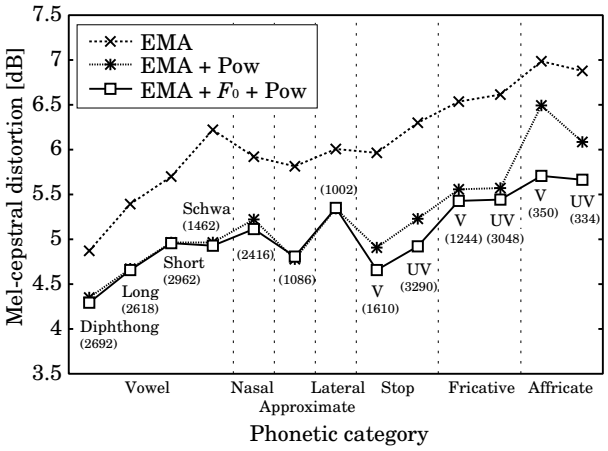

Fig. 2. Mel-cepstral distortion in each phonetic category for the female speaker. "V" denotes voiced phonemes, and "UV" denotes unvoiced phonemes. The number in each bracket denotes the number of frames for each phonetic category.

4. MAXIMUM LIKELIHOOD SPECTRAL ESTIMATION USING DYNAMIC FEATURES

In the mapping function in Eq. (1), the estimated spectrum is defined as the weighted sum of the product of each of the conditional mean vectors in individual mixtures and the conditional probabilities that the input feature vector belongs to each one of the mixtures. This function is not supported by a proper statistical model. To perform the mapping based on a statistical model, we apply maximum likelihood estimation (MLE) to the GMM-based mapping algorithm. MLE has also been applied to the HMM-based speech production model [8][9][10]. In this production model, a HMM state sequence is determined by the Viterbi algorithm. Meanwhile, we use the EM algorithm described in [18] for maximizing a likelihood in this paper.

4.1. Spectral estimation based on maximum likelihood criterion

The conditional probability of the target feature vector $\boldsymbol{y}_t$ for the given input feature vector $\boldsymbol{x}_t$ is written as

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\Theta}) = \sum_{i=1}^{M} p(m_i|\boldsymbol{x}_t, \boldsymbol{\Theta}) p(\boldsymbol{y}_t|\boldsymbol{x}_t, m_i, \boldsymbol{\Theta}), \quad (7)$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, m_i, \boldsymbol{\Theta}) = N(\boldsymbol{y}_t; \boldsymbol{E}_t(i), \boldsymbol{D}(i)), \quad (8)$$

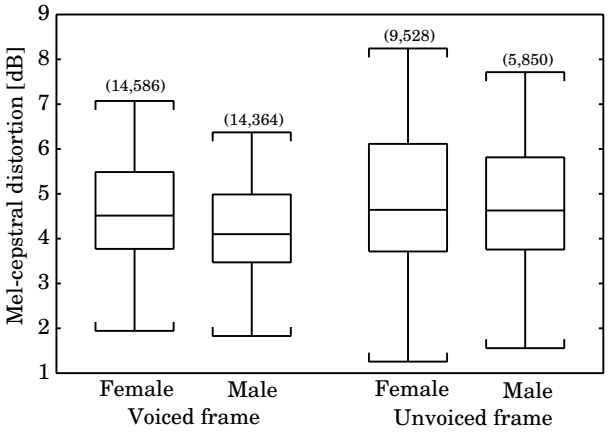

Fig. 3. Frequency distributions of mel-cepstral distortion in voiced and unvoiced frames. Input features are EMA+F_0+Pow. The number in each bracket denotes the number of frames for each phonetic category.

where $\boldsymbol{E}_t(i)$ is equal to $\boldsymbol{E}(\boldsymbol{y}_t|\boldsymbol{x}_t, m_i, \boldsymbol{\Theta})$ in Eq. (2) and $\boldsymbol{D}(i)$ is defined as

$$\boldsymbol{D}(i) = \boldsymbol{\Sigma}_i^{(yy)} - \boldsymbol{\Sigma}_i^{(yx)}\boldsymbol{\Sigma}_i^{(xx)^{-1}}\boldsymbol{\Sigma}_i^{(xy)}. \quad (9)$$

Let $\boldsymbol{X} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_T^\top\right]^\top$ be a time sequence of the input feature vector and $\boldsymbol{Y} = \left[\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$ be that of the target feature vector. In order to maximize a likelihood function $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})$, we maximize an auxiliary function of a current feature vector sequence $\boldsymbol{Y}$ and a new feature vector sequence $\hat{\boldsymbol{Y}}$ defined by

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{\text{all } \boldsymbol{m}} p(\boldsymbol{Y}, \boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\Theta}) \log p(\hat{\boldsymbol{Y}}, \boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\Theta}) \quad (10)$$

$$= p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta}) \left\{-\frac{1}{2}\hat{\boldsymbol{Y}}^\top \overline{\boldsymbol{D}^{-1}}\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}^\top \overline{\boldsymbol{D}^{-1}\boldsymbol{E}} + \overline{K}\right\}, \quad (11)$$

where

$$\overline{\boldsymbol{D}^{-1}} = \text{diag}\left[\overline{\boldsymbol{D}_1^{-1}}, \overline{\boldsymbol{D}_2^{-1}}, \cdots, \overline{\boldsymbol{D}_T^{-1}}\right], \quad (12)$$

$$\overline{\boldsymbol{D}_t^{-1}} = \sum_{i=1}^{M} \gamma_t(i)\boldsymbol{D}(i)^{-1}, \quad (13)$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}} = \left[\overline{\boldsymbol{D}_1^{-1}\boldsymbol{E}_1}^\top, \overline{\boldsymbol{D}_2^{-1}\boldsymbol{E}_2}^\top, \cdots, \overline{\boldsymbol{D}_T^{-1}\boldsymbol{E}_T}^\top\right]^\top, \quad (14)$$

$$\overline{\boldsymbol{D}_t^{-1}\boldsymbol{E}_t} = \sum_{i=1}^{M} \gamma_t(i)\boldsymbol{D}(i)^{-1}\boldsymbol{E}_t(i), \quad (15)$$

$$\gamma_t(i) = p(m_i|\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\Theta}),$$

$$= \frac{p(m_i|\boldsymbol{x_t}, \boldsymbol{\Theta})N(\boldsymbol{y}_t; \boldsymbol{E}_t(i), \boldsymbol{D}(i))}{\sum_{j=1}^{M} p(m_j|\boldsymbol{x_t}, \boldsymbol{\Theta})N(\boldsymbol{y}_t; \boldsymbol{E}_t(j), \boldsymbol{D}(j))}, \quad (16)$$

where the constant $\overline{K}$ is independent of $\hat{\boldsymbol{Y}}$. $\hat{\boldsymbol{Y}}$ that maximizes $Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$ is given by

$$\hat{\boldsymbol{Y}} = \left(\overline{\boldsymbol{D}^{-1}}\right)^{-1} \overline{\boldsymbol{D}^{-1}\boldsymbol{E}}. \quad (17)$$

The target vector sequence given by the conventional mapping function is used as the initial vector sequence $\boldsymbol{Y}$. $\hat{\boldsymbol{Y}}$ is calculated by the above equations, and then $\hat{\boldsymbol{Y}}$ is substituted for $\boldsymbol{Y}$. This procedure is iteratively performed until a convergence condition is satisfied.

There is little difference between the acoustic spectrum estimated with diagonal covariance matrices and that with full covariance matrices in our preliminary experiments. Therefore, we use only diagonal elements of $\boldsymbol{D}(i)$.

4.2. Introducing dynamic features

Some discontinuities are often shown in a sequence of the estimated target feature because the correlation between frames is not considered in the conventional GMM-based mapping [17]. To address this problem, we introduce dynamic features for modeling and estimating processes. Specifically, we use the parameter generation algorithm described in [18]. This method has also been applied to the HMM-based speech production model [9][10].

Not only static but also dynamic features are used as the input and target feature vectors, which are given by

$$\boldsymbol{x}'_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top, \quad \boldsymbol{y}'_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top. \quad (18)$$

To estimate a model parameter set $\boldsymbol{\Theta}'$ for these feature vectors, the GMM on the joint vector $\boldsymbol{z}' = [\boldsymbol{x}'^\top, \boldsymbol{y}'^\top]^\top$ is trained with EM algorithm.

We can represent the relationship between a sequence of the static feature and a sequence of the static and dynamic features as a linear conversion,

$$\boldsymbol{Y}' = \boldsymbol{W}\boldsymbol{Y} = \left[\boldsymbol{y}'^\top_1, \boldsymbol{y}'^\top_2, \cdots, \boldsymbol{y}'^\top_T\right]^\top, \quad (19)$$

where $\boldsymbol{W}$ is a transformation matrix described in [18]. The target static feature sequence is estimated by maximizing an auxiliary function defined as

$$Q(\boldsymbol{Y}', \hat{\boldsymbol{Y}}') = \sum_{\text{all } \boldsymbol{m}} p(\boldsymbol{Y}', \boldsymbol{m}|\boldsymbol{X}', \boldsymbol{\Theta}') \log p(\hat{\boldsymbol{Y}}', \boldsymbol{m}|\boldsymbol{X}', \boldsymbol{\Theta}')$$

$$= p(\boldsymbol{Y}'|\boldsymbol{X}', \boldsymbol{\Theta}') \left\{-\frac{1}{2}\hat{\boldsymbol{Y}}^\top \boldsymbol{W}^\top \overline{\boldsymbol{D}'^{-1}}\boldsymbol{W}\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}^\top \boldsymbol{W}^\top \overline{\boldsymbol{D}'^{-1}\boldsymbol{E}'} + \overline{K'}\right\}, \quad (20)$$

where $\overline{\boldsymbol{D}'^{-1}}$ and $\overline{\boldsymbol{D}'^{-1}\boldsymbol{E}'}$ are comprised of static and dynamic parts and are calculated in a similar way as was described in the previous sub-section using the following $\gamma'_t(i)$,

$$\gamma'_t(i) = \frac{p(m_i|\boldsymbol{x}'_t, \boldsymbol{\Theta}')N(\boldsymbol{y}'_t; \boldsymbol{E}'_t(i), \boldsymbol{D}'(i))}{\sum_{j=1}^{M} p(m_j|\boldsymbol{x}'_t, \boldsymbol{\Theta}')N(\boldsymbol{y}'_t; \boldsymbol{E}'_t(j), \boldsymbol{D}'(j))}. \quad (21)$$

The sequence of the estimated target static feature, $\hat{\boldsymbol{Y}}$, that maximizes $Q(\boldsymbol{Y}', \hat{\boldsymbol{Y}}')$ is given by

$$\hat{\boldsymbol{Y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}'^{-1}}\boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}'^{-1}\boldsymbol{E}'}. \quad (22)$$

The iteration process as mentioned above is performed until a convergence condition is satisfied. In this case, we also use only diagonal elements of $\boldsymbol{D}'(i)$.

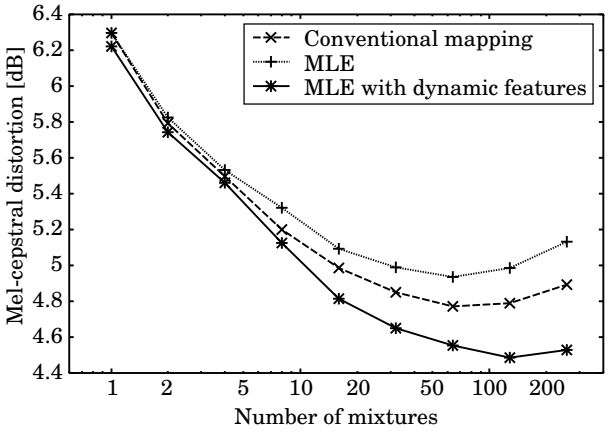

Fig. 4. Mel-cepstral distortion as a function of the number of mixtures in each of the estimation methods. The mel-cepstral distortion shows the average of the distortions for the female speaker and for the male speaker.

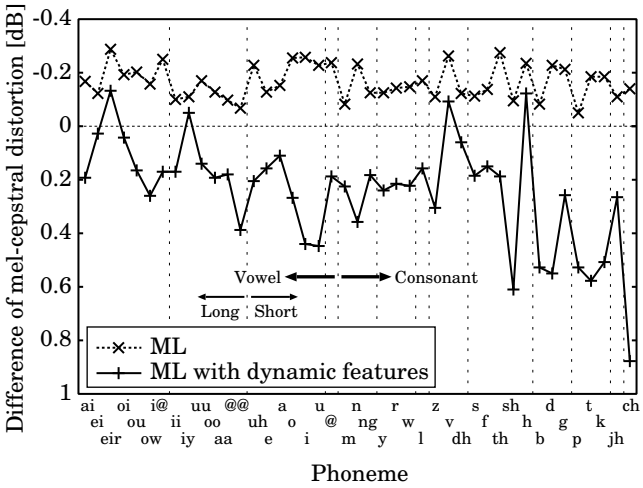

Fig. 5. Difference between mel-cepstral distortions in MLE and the conventional mapping. Minus difference means decrease of mel-cepstral distortion compared with that of the conventional mapping. The difference shows the average of the differences for the female speaker and for the male speaker.

4.3. Experimental evaluation of spectral estimation method

We investigate the effectiveness of applying MLE and introducing dynamic features to the GMM-based mapping.

Experimental conditions were the same as those in the previous section. The EMA data, the log-scaled F_0 and the log-scaled power of a linear spectrum were used as the input features.

Figure 4 shows mel-cepstral distortion as a function of the number of mixtures in each of estimation methods. MLE with only static features is worse than the conventional mapping function. There is a tendency that the MLE causes more discontinuities than the conventional mapping. In the conventional mapping, the discontinuities are alleviated by an interpolation between conditional mean vectors using the conditional probability. This interpolation also causes the decrease of mel-cepstral distortion. The problem of the MLE can be solved by introducing dynamic features. Consequently, the smallest average and standard deviation of the mel-cepstral distortion (female: 4.69 $\pm$ 1.67 dB, 128 mix., male: 4.29 $\pm$ 1.41 dB, 128 mix.) is achieved by MLE with the dynamic features. Using dynamic features causes an increase in the optimum number of mixtures because a large number of mixtures is needed for modeling a joint space on both static and dynamic features.

The differences between mel-cepstral distortion in the ML-based mappings and that of the conventional mapping calculated for individual phonemes are shown in **Fig. 5**. MLE using only static features causes the deterioration of mel-cepstral distortion in all phonemes. By introducing dynamic features, however, the mel-cepstral distortions in many phonemes can be improved. Larger improvement is achieved in the shorter vowels compared with longer vowels. As described in the previous experimental result shown in **Fig. 2**, the mapping performance in the shorter vowels is worse than that in the longer vowels when the mapping is performed with the conventional function taking account into only static features. The mel-cepstra in the shorter vowels have more various transitions than in the longer vowels. Therefore, the dynamic features capturing the transition of mel-cepstra are especially useful for improving the mapping accuracy in the short vowel.

5. SYNTHESIZING SPEECH WITH ESTIMATED VOCAL TRACT SPECTRUM

Speech synthesis with the vocal tract spectral sequence estimated from articulatory configurations is performed as follows: 1) converting the estimated mel-cepstra into linear spectra, 2) controlling a power of each spectrum, and 3) synthesizing the speech signal from the spectra and F_0 using STRAIGHT synthesis [15], where source excitation is constructed by pulse and noise with phase manipulation.

5.1. Preliminary perceptual evaluation

We have performed preliminary evaluations of synthetic speech quality to demonstrate 1) the effectiveness of using features of source information and 2) the effectiveness of performing MLE with dynamic features. The number of listeners was three. Training data was the same as that used in the previous experiments. Ten sentences not included in the training data were used for evaluations. An F_0 and a power of a linear spectrum automatically extracted from natural speech were used for synthesizing speech.

Two preference tests were conducted. In one test, we compared synthetic voices using only the EMA ("EMA"), the EMA and the log-scaled power of a linear spectrum ("EMA+Pow"), and the EMA, the power and the log-scaled F_0 ("EMA+F_0+Pow"). The number of mixtures was set to 32 in "EMA" and 64 in both "EMA+Pow" and "EMA+F_0+Pow", respectively. In the other test, we compared synthetic voices with the conventional mapping function ("Conv"), MLE ("MLE"), and MLE using dynamic features ("MLE with dyn"). The number of mixtures was set to 64 in both "Conv" and "MLE" and 128 in "MLE with dyn". In each test, 120 stimulus pairs were evaluated by each listener.

Results are shown in **Table 2**. We can see the similar tendency as was shown in the previous objective evaluations: synthetic speech quality is improved by using features of source information and applying MLE with dynamic features.

Table 2. Preference score [%]. The upper table shows the result of comparison between input features. The lower table shows the result of comparison between mapping methods. Confidence intervals (95%) are also shown in the total results. The preference score shows the ratio of the number of samples selected as having better quality to the number of samples presented to listeners

Feature	Total	Female	Male
EMA	14.0≤ 18.8 ≤24.3	11.7	25.8
EMA+Pow	56.5≤ 62.9 ≤69.0	66.7	59.2
EMA+F_0+Pow	62.0≤ 68.3 ≤74.2	71.7	65.0
Method	**Total**	**Female**	**Male**
Conv	38.2≤ 44.6 ≤51.1	40.0	49.2
MLE	39.4≤ 45.8 ≤52.4	49.2	42.5
MLE with dyn	53.1≤ 59.6 ≤65.8	60.8	58.3

6. CONCLUSION

In order to realize speech synthesis using articulatory information, a mapping from articulatory parameters to the vocal tract spectrum was performed using a Gaussian Mixture Model (GMM). The MOCHA acoustic-articulatory speech database was used for training the GMM on a joint space of articulatory and spectral feature vectors. From results of experimental evaluations based on a spectral distortion measure, it was shown that using both articulatory features of the vocal tract and acoustic features capturing characteristics of the source make it possible to improve the mapping accuracy of the acoustic spectrum. In order to achieve a more appropriate mapping, we applied maximum likelihood estimation (MLE) using dynamic features to the GMM-based mapping. We also performed experimental evaluations based on the distortion measure to demonstrate the effectiveness of using MLE. As a result, it was shown that the distortion could be reduced by considering dynamic features. These improvements could be also seen in the results of preliminary perceptual tests.

We will perform further perceptual evaluations. Synthetic speech with the estimated spectral sequence has vocoder-like speech quality. It is worthwhile to introduce a residual excitation in the speech synthesis phase to improve the quality of synthetic speech. Moreover, we need to construct a framework for modifying speech by manipulating articulatory parameters to achieve high-quality and flexible speech synthesis.

Acknowledgment: This research was supported in part by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists. The authors are grateful to Dr. Hideki Kawahara of Wakayama University in Japan for permission to use the STRAIGHT analysis-synthesis method.

7. REFERENCES

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proc. ICASSP*, pp. 679–682, New York, U.S.A., Apr. 1988.

[2] W.N. Campbell and A.W. Black. Prosody and the selection of source units for concatenative synthesis. *Progress in Speech Synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, N.Y., Springer, pp. 279–292, 1997.

[3] A.K. Syrdal, C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay. Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. ICSLP*, Vol. 3, pp. 410–415, Beijing, China, Oct. 2000.

[4] J. Schroeter and M.M. Sondhi. Speech coding based on physiological models of speech production. *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi, Marcel Dekker New York, pp. 231–267, 1992.

[5] M.M. Sondhi. Articulatory modeling: a possible role in concatenative text-to-speech synthesis. *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, U.S.A., Sep. 2002.

[6] Y. Shiga and S. King. Estimating the spectral envelope of voiced speech using multi-frame analysis. *Proc. EUROSPEECH*, pp. 1737–1740, Geneva, Switzerland, Sep. 2003.

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

[8] S. Hiroya and M. Honda. Determination of articulatory movements from speech acoustics using an HMM-based speech production model. *Proc. ICASSP*, Orlando, U.S.A, pp. 437–440, May. 2002.

[9] S. Hiroya and M. Honda. Acoustic-to-articulatory inverse mapping using an HMM-based speech production model. *Proc. ICSLP 2002*, pp.2305–2308, Denver, U.S.A., Sep. 2002.

[10] S. Hiroya and M. Honda. Speech inversion for arbitrary speaker using a stochastic speech production model. *An Interdisciplinary Workshop on Speech Dynamics by Ear, Eye, Mouth and Machine*, pp. 9–14, Kyoto, Japan, June 2003.

[11] Y. Stylianou. *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification.* Ph.D. Thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[12] A. Wrench. The MOCHA-TIMIT articulatory database. *http://www.cstr.ed.ac.uk/artic/mocha.html*, Queen Margaret University College, 1999.

[13] K. Richmond. *Estimating articulatory parameters from the acoustic speech signal*. Ph.D. Thesis, The Centre for Speech Technology Research, University of Edinburgh, 2001.

[14] A. Kain. *High Resolution Voice Transformation.* Ph.D. Thesis, OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[15] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[16] H. Kawahara, H. Katayose, A.de Cheveigné, and R.D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity. *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sep. 1999.

[17] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu. Voice conversion with smoothed GMM and MAP adaptation. *Proc. EUROSPEECH*, pp. 2413–2416, Geneva, Switzerland, Sep. 2003.

[18] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.