ON THE USE OF PHONETIC INFORMATION FOR MAPPING FROM ARTICULATORY MOVEMENTS TO VOCAL TRACT SPECTRUM

Kenichi Nakamura †, *Tomoki Toda*‡, *Yoshihiko Nankaku*†, *Keiichi Tokuda*†

† Nagoya Institute of Technology Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555 Japan
‡Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0101 Japan
{ k-n, nankaku, tokuda } @ics.nitech.ac.jp
tomoki@is.naist.jp

ABSTRACT

This paper describes a method for determining the vocal tract spectrum from articulatory movements using an hidden Markov models (HMMs). In the proposed system, articulatory parameters are generated from a TTS system and converted to acoustic features to be synthesized. Comparing with conventional GMM-based systems, the proposed system has two additional properties: 1) phonetic information given input texts is available for the conversion, 2) the use of HMMs allows us to utilize the temporal structure of speech. In this paper, we investigate the optimal structure of HMMs for the conversion. Experimental results show that using phonetic and temporal information can improve the mapping accuracy in a spectral distortion measure.

1. INTRODUCTION

Many attempts to synthesize speech based on speech production mechanisms which are ignored in concatenative synthesis have been studied for several decades. In these approaches, the speech signal is generated from articulatory parameters [1] by a mathematical production model in which speech is characterized by the properties of the vocal apparatus instead of the speech acoustics. Slowly varying articulatory parameters are better candidates of features for speech modeling. Furthermore, the speech signal can be modified in an understandable way by manipulating articulatory parameters rather than acoustic parameters such as vocal tract spectrum.

Figure 1 shows the proposed TTS system with articulatory parameter conversion. First, F_0 and articulatory parameters are generated from the HMM-based TTS [2] which could be easily constructed by using articulatory parameters as training data. Then, articulatory parameters are modified to realize various speaking styles and they are converted to spectrum features, e.g., mel-cepstrum. Finally speech waveforms are synthesized from converted parameters and F_0 by using a speech synthesis filter.

In the conversion system, the mapping between acoustic and articulatory features is statistically determined using a parallel acoustic-articulatory speech database [3]. As a way to implement the transformation function for converting articulatory parameters to speech, the GMM-based system [4], and the HMM-based system [5] have been proposed. However, a detailed comparison between these systems has not been performed. Comparing with the GMM-based system, the proposed system has two additional properties: 1) phonetic information given input texts is available for the conversion process, 2) the use of HMMs allows us to utilizes the temporal structure of speech.

This work was partly supported by MEXT e-Society project.

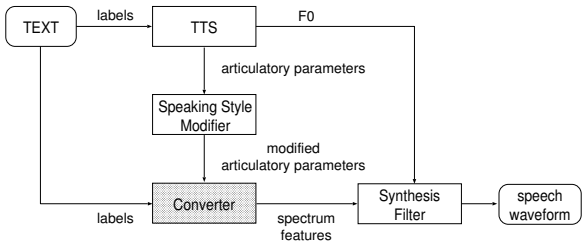

Fig. 1. TTS with articulatory parameter conversion

In this paper, we present a method for determining speech acoustics from articulatory movements using multi-mixture HMMs. We investigate the optimal structure of multi-mixture HMMs with context clustering [6]. To vary the degree of using the phonetic information and temporal structure of HMMs, we change the size of decision trees in context clustering, the number of HMM states, and mixture components while keeping the total number of model parameters fixed.

The rest of this paper is organized as follows. In the following section, we introduce the MOCHA database. In Section 3, the HMM-based speech conversion system is described. The maximum likelihood spectral estimation using dynamic features is applied to the HMM-based mapping in Section 4, and speech synthesis with the estimated spectral sequence is described in Section 5. Finally, we summarize this paper in Section 6.

2. ACOUSTIC-ARTICULATORY SPEECH DATABASE：MOCHA

The Multichannel articulatory database (MOCHA) [3] has been released from the University of Edinburgh. It consists of speech and some articulatory movements simultaneously recorded at Queen Margaret University College.

We use electromagnetic articulograph (EMA) data, one of representations of articulatory data provided in the MOCHA, as an articulatory representation. The movements of seven

articulators (top lip, bottom lip, bottom incisor, tongue tip, tongue body, tongue dorsum, and velum) and two reference points (the bridge of nose and the upper incisor) are sampled in the midsagittal plane at 500 Hz.

3. CONVERSION SYSTEM OVERVIEW

In the HMM-based conversion system, we construct feature vectors using parameters obtained from the parallel acoustic and articulatory speech database. The feature vector consists of mel-cepstral coefficients as spectral parameters, EMA data as articulatory parameters, and their delta and delta-delta parameters. The joint probability densities of articulatory and acoustic spectral parameters are modeled by the HMM using these feature vectors.

In the training stage, first, monophone HMMs are estimated by the isolated training and the following embedded training. After converting to context dependent HMMs, they are re-estimated by the embedded training. To avoid inaccurate estimates caused by a limited amount of data, we apply the tree-based context clustering technique [6].

In the conversion stage, first, the text to be synthesized is converted to a context dependent label sequence. Then, the sentence HMM is constructed by concatenating context dependent HMMs according to the label sequence. Articulatory parameters are converted to spectrum features based on the maximum likelihood estimation. Finally, a speech waveform is synthesized from the generated parameters by using a speech synthesis filter.

4. TRAINING JOINT PROBABILITY DISTRIBUTION WITH HMM

To convert the articulatory parameters to acoustic ones, the joint probability densities over two features are trained using the HMM. Each articulatory location is shown by x- and y-coordinates, therefore articulatory movements are represented as 14 dimensional vector sequence. Moreover, the proposed system can represent probability densities more precisely using multi-mixture compared with the conventional HMM-based system using single mixture [5]. Let $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ be articulatory and acoustic feature vectors, respectively. Let the vector $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ be a joint feature of these two features, and its vector sequence $\boldsymbol{Z} = \left[\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, \cdots, \boldsymbol{Z}_T^\top\right]^\top$ is modeled by the HMM λ. The output probability of $\boldsymbol{Z}$ given the HMM can be written as follows

$$p(\boldsymbol{Z} \mid \lambda) = \sum_{\text{all } \boldsymbol{q}} \sum_{\text{all } \boldsymbol{m}} \left[p(\boldsymbol{q} \mid \lambda) p(\boldsymbol{m} \mid \boldsymbol{q}, \lambda) \prod_{t=1}^{T} p(\boldsymbol{Z}_t \mid m_t, q_t, \lambda) \right] \quad (1)$$

where $\boldsymbol{q} = (q_1, q_2, \cdots, q_T)$ is a state sequence of the HMM, $\boldsymbol{m} = (m_1, m_2, \cdots, m_T)$ denotes a component number sequence of mixture distributions. The probabilities $p(\boldsymbol{q} \mid \lambda)$ and $p(\boldsymbol{m} \mid \boldsymbol{q}, \lambda)$ denote a state transition probability and mixture weights of output probability, respectively. In this paper, the mixture component is assumed to be a Gaussian distribution:

$$p(\boldsymbol{Z}_t \mid m_t = i, q_t = j, \lambda) = \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_{i,j}^{(Z)}, \boldsymbol{\Sigma}_{i,j}^{(Z)}\right), \quad (2)$$

$$\boldsymbol{\mu}_{i,j}^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_{i,j}^{(X)} \\ \boldsymbol{\mu}_{i,j}^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_{i,j}^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_{i,j}^{(XX)} & \boldsymbol{\Sigma}_{i,j}^{(XY)} \\ \boldsymbol{\Sigma}_{i,j}^{(YX)} & \boldsymbol{\Sigma}_{i,j}^{(YY)} \end{bmatrix}, \quad (3)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote a mean vector and a covariance matrix, respectively. In the above-mentioned condition, the parameters of the HMM λ is estimated via the EM algorithm.

4.1. Maximum likelihood spectral estimation

In the maximum likelihood (ML) spectral estimation, given the articulatory features $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ as an input, the optimal spectral features $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ is obtained by maximizing the following conditional probability,

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \lambda) = \sum_{\text{all } \boldsymbol{q}} \sum_{\text{all } \boldsymbol{m}} \Big[p(\boldsymbol{q} \mid \boldsymbol{X}, \lambda) \times p(\boldsymbol{m} \mid \boldsymbol{q}, \boldsymbol{X}, \lambda) \prod_{t=1}^{T} p(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, m_t, q_t, \lambda) \Big], \quad (4)$$

where the output probability distribution is written as follows:

$$p(\boldsymbol{Y}_t \mid \boldsymbol{X}_t, q_t = j, m_t = i, \lambda) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{i,j}(t), \boldsymbol{D}_{i,j}) \quad (5)$$

and

$$\boldsymbol{E}_{i,j}(t) = \boldsymbol{\mu}_{i,j}^{(Y)} + \boldsymbol{\Sigma}_{i,j}^{(YX)} \boldsymbol{\Sigma}_{i,j}^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_{i,j}^{(X)}\right), \quad (6)$$

$$\boldsymbol{D}_{i,j} = \boldsymbol{\Sigma}_{i,j}^{(YY)} - \boldsymbol{\Sigma}_{i,j}^{(YX)} \boldsymbol{\Sigma}_{i,j}^{(XX)^{-1}} \boldsymbol{\Sigma}_{i,j}^{(XY)}. \quad (7)$$

Furthermore, the posterior state transition probability $p(\boldsymbol{q} \mid \boldsymbol{X}, \lambda)$ and the mixture weight $p(\boldsymbol{m} \mid \boldsymbol{q}, \boldsymbol{X}, \lambda)$ is also calculated using articulatory parameters $\boldsymbol{X}$ and model parameters λ. Since equation (4) includes hidden variables, the optimal sequence of $\boldsymbol{Y}$ is estimated via the EM algorithm. The EM algorithm is an iterative method for approximating the maximum likelihood estimation. It maximizes the expectation of the complete data log-likelihood so called Q-function (auxiliary function):

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{\text{all } \boldsymbol{q}} \sum_{\text{all } \boldsymbol{m}} \left[p(\boldsymbol{m}, \boldsymbol{q} \mid \boldsymbol{Y}, \boldsymbol{X}, \lambda) \log p(\hat{\boldsymbol{Y}}, \boldsymbol{m}, \boldsymbol{q} \mid \boldsymbol{X}, \lambda) \right] \quad (8)$$

Taking the derivative of the Q-function, the spectral sequence $\hat{\boldsymbol{Y}}$ which maximizes the Q-function is given by

$$\hat{\boldsymbol{Y}} = \left(\overline{\boldsymbol{D}^{-1}}\right)^{-1} \overline{\boldsymbol{D}^{-1}\boldsymbol{E}}, \quad (9)$$

where

$$\overline{\boldsymbol{D}^{-1}} = \operatorname{diag}\left[\overline{\boldsymbol{D}_1^{-1}}, \overline{\boldsymbol{D}_2^{-1}}, \cdots, \overline{\boldsymbol{D}_T^{-1}}\right], \quad (10)$$

$$\overline{\boldsymbol{D}_t^{-1}} = \sum_{j=1}^{N} \sum_{i=1}^{M_j} \gamma_{i,j}(t) \boldsymbol{D}_{i,j}^{-1}, \quad (11)$$

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

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_t} = \sum_{j=1}^{N} \sum_{i=1}^{M_j} \gamma_{i,j}(t) \boldsymbol{D}_{i,j}^{-1} \boldsymbol{E}_{i,j}(t), \quad (13)$$

$$\gamma_{i,j}(t) = p(q_t = j \mid \boldsymbol{X}, \boldsymbol{Y}, \lambda) \times p(m_t = i \mid q_t = j, \boldsymbol{X}, \boldsymbol{Y}, \lambda). \quad (14)$$

The occupancy probability $\gamma_{i,j}(t)$ can be calculated by the forward-backward algorithm. Using the updated probabilities $\gamma_{i,j}(t)$, a new vector sequence $\hat{\boldsymbol{Y}}$ is calculated by equations (9), and then $\hat{\boldsymbol{Y}}$ is substituted for $\boldsymbol{Y}$. This procedure is iteratively performed until a certain convergence condition is satisfied.

4.2. Maximum likelihood spectral estimation using dynamic features

In this paper, we appropriately estimate the spectral feature sequence using dynamic features as described [4, 5]. Let $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top, \Delta^2\boldsymbol{x}_t^\top\right]^\top$ and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top, \Delta^2\boldsymbol{y}_t^\top\right]^\top$ be an articulatory feature and an acoustic feature, respectively. Where $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ denote static features, and the notations, Δ , Δ^2 represent first and second order dynamic features, respectively, calculated from the neighboring frames of time t. The relation between the static spectral sequence $\boldsymbol{y} = \left[{\boldsymbol{y}_1}^\top, {\boldsymbol{y}_2}^\top, \cdots, {\boldsymbol{y}_T}^\top\right]^\top$ and the static-dynamic features $\boldsymbol{Y}$ can be written as the following linear transformation:

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (15)$$

where $\boldsymbol{W}$ is a matrix which concatinates dynamic features to the static feature sequence $\boldsymbol{y}$. Under this relation, the static feature vector sequence $\hat{\boldsymbol{y}}$ which maximizes equation (8) is given by

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{-1}\boldsymbol{E}}. \quad (16)$$

Similarly to equation (9), the update is iterated until a certain convergence condition is satisfied.

5. EXPERIMENT

5.1. Experimental conditions

We investigated the effectiveness of using phonetic and temporal information by varying the importance of these properties. The acoustic-articulatory data described in Section 2 was used. Experimental conditions are shown in Table 1.

To investigate the mapping accuracy of the HMMs, we fixed the total number of parameters of HMMs, then assigned them variously. Where the importance of temporal information is represented by the state number of HMMs, and that of phonetic information is represented by the size of decision-tree in context clustering. In context clustering, a large single tree including all triphone HMMs was constructed for each temporal HMM state, which allows parameter sharing among different phone HMMs. Furthermore, to assign the optimal number of mixtures for each state (cluster), we apply the following procedure:

1. Construct a root node for all states of all HMMs.
2. Apply the questions which divide all temporal HMM states.
3. Perform the context clustering until the predetermined number of clusters are generated.
4. Back off the tree in the reverse order of divisions until the designed size of tree.
5. In the new leaf node obtained in 4, nodes of the sub tree are used as the mixture components, and their weights are determined by the occupancy count of the training data.

Table 1. Experimental Condition

database	MOCHA DATABASE
training sentences	414
evaluation sentences	46
acoustic data	
sampling frequency	16kHz
shift length	5ms
frame length	25ms
window function	Blackman window
analysis	24-order mel-cepstrum
articulatory data	
location (2-dimensional coordinate)	tongue tip, tongue body tongue dorsum top lip, bottom lip bottom incisor, velum
sampling frequency	500Hz → 200Hz
coordinate normalization	mean=0, variance=1
statistical model	left-to-right HMM
number of HMM states	1, 3, 5, 7, 9
total Gaussian distributions	64,128,256,512,768,1024

The variance parameters of HMMs were trained as diagonal covariances, and after the context clustering they were estimated by the embedded training as full covariance matrices.

In Section 4.2, we presented the process which iteratively estimates a spectral feature sequence and posterior probability distributions of the state transition and the mixture components. However, in this experiments we use the state alignment generated from the natural articulatory-acoustic data, hence only the posterior probabilities of mixture components were re-estimated, iteratively. In the experiment, F_0 sequences which automatically extracted from natural speech are used for synthesizing speech to focus on the spectral conversion.

The mel-cepstral distortion between the target and the estimated mel-cepstrum given by the following equation was used as the evaluation measure:

$$\text{MelCD} = \frac{10}{\log 10} \sqrt{2 \sum_{i=1}^{24} \left(mc_i^{(t)} - mc_i^{(e)}\right)^2} \quad (17)$$

where $mc_i^{(t)}$ and $mc_i^{(e)}$ denote the i-th coefficient of the target and the estimated mel-cepstrum, respectively.

5.2. Experimental results

To investigate only the effectiveness of phonetic information, we apply the context clustering to the GMM-based mapping. Figure 2 shows the MelCD of the GMM-based mapping with context clustering, which is equivalent to the multi-mixture

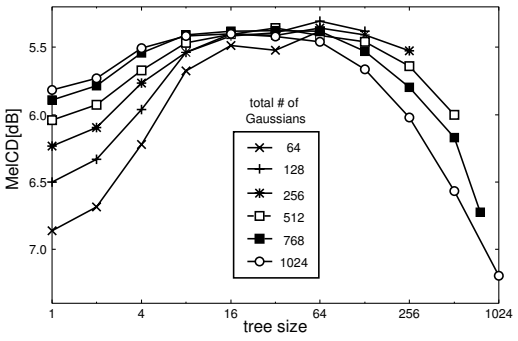

Fig. 2. MelCD of the GMM-based mapping with context clustering (♯ of HMM states = 1)

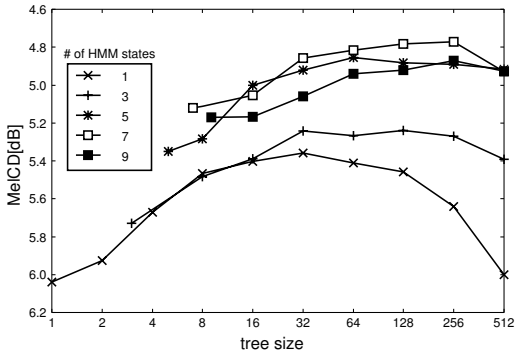

Fig. 3. MelCD of the HMM-based mapping (total ♯ of Gaussians = 512)

HMM-based mapping with the number of states is one, hence temporal information could not be modeled. As the decision-tree becomes large, phonetic information becomes positively used, and the left end of the graph indicates the mapping without using phonetic information, which is equivalent to the conventional GMM-based method.

It is observed that the mapping accuracy without phonetic information decreases in proportion to the decrease of the number of mixtures. However, the mapping using phonetic information (the tree size is around 64) achieved high accuracy even with fewer parameters. Furthermore, the result of 64 Gaussians with phonetic information is superior than the conventional GMM mapping of 1024 Gaussians. These results show that phonetic information is useful for converting articulatory features to acoustic ones.

We investigate the effectiveness of introducing temporal information to the GMM-based mapping. Figure 3 shows the MelCD of the multi-mixture HMM-based mapping, where the total number of Gaussian distributions is fixed to 512. It can be seen that the mapping accuracy can be improved by using temporal information. However the use of too many HMM states degrades the performance, which may be due to inadequate state alignments. The tree sizes, which achieved the highest accuracy in each number of HMM state tend to increase with the increase of HMM states. It is supposed that, independently of the number of HMM states, a similar number of clusters is required for each HMM state to represent its context dependency. This result suggests that the simultaneous use of phonetic and temporal information is effective for the conversion system.

6. CONCLUSION

In this paper, we examined an effectiveness of using phonetic and temporal information for converting articulatory movements to vocal tract spectrum. In the objective evaluation, it was confirmed that mapping accuracy is improved by using both phonetic and temporal information. Future works include investigating more effective contexts for articulatory-acoustic conversion. Constructing a TTS system for synthesizing articulatory features and listening tests are also future works.

Acknowledgments: Authors are grateful to Dr. Heiga Zen for helpful discussions.

7. REFERENCES

[1] M. M. Sondhi, "Articulatory modeling: a possible role in concatenative text-to-speech synthesis," *IEEE 2002 Workshop on Speech Synthesis*, Sept. 2002.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *Eurospeech*, vol. 5, pp. 2347–2350, Sept. 1999.

[3] A. Wrench, "http://www.cstr.ed.ac.uk/artic/mocha.html," Queen Margaret University College, 1999.

[4] T. Toda, A. W. Black, and K. Tokuda, "Mapping from ariticulatory movements to vocal tract spectrum with gaussian mixture model for ariticulatory speech synthesis," *5th ISCA Speech Synthesis Workshop-Pittsburgh*, pp. 31–36, June 2004.

[5] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoutstics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, pp. 175–184, Mar. 2004.

[6] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 1995.