A Technique for Controlling Voice Quality of Synthetic Speech Using Multiple Regression HSMM

Makoto Tachibana, Takashi Nose, Junichi Yamagishi, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
{makoto.tachibana, takashi.nose, junichi.yamagishi, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper describes a technique for controlling voice quality of synthetic speech using multiple regression hidden semi-Markov model (HSMM). In the technique, we assume that the mean vectors of output and state duration distribution of HSMM are modeled by multiple regression with a parameter vector called voice quality control vector. We first choose three features for controlling voice qualities, that is, "smooth voice – nonsmooth voice," "warm – cold," "high-pitched – low-pitched," and then we attempt to control voice quality of synthetic speech for these features. From the results of several subjective tests, we show that the proposed technique can change these features of voice quality intuitively.

Index Terms: HMM-based speech synthesis, voice quality control, multiple regression HMM, HSMM.

1. Introduction

To obtain easily desired voice quality of synthetic speech, it is more desirable that we can control voice quality intuitively based on linguistic expressions of voice quality, such as "brightness," "softness," and "warmth," rather than operate the physical features, such as fundamental frequency (F0), formant, and spectrum. In an HMM-based speech synthesis system, voice characteristics of synthesized speech can be easily changed by transforming HMM parameters [1],[2], and an eigenvoice technique [3] has been proposed for controlling voice quality. In the eigenvoice technique, a number of speaker dependent HMM sets are represented by a few parameters obtained by using principal component analysis (PCA), and one can define an arbitrary voice quality by setting a few parameters, i.e., weights for the eigenvoices. However, the eigenvoice technique has a problem that, in general, each axis of the eigenspace does not represent specific physical meaning, and therefore, it is not easy to control desired voice quality intuitively.

On the other hand, we have proposed an alternative technique for controlling HMM parameters based on multiple regression HMM (MRHMM) [4], more specifically, we can control HMM parameters according to a specified axis which represents a certain feature. In [4], we showed that we can change emotional expressions and/or speaking styles of the synthesized speech intuitively by specifying the intensity of a desired style represented in a low-dimensional space. Furthermore, we have proposed multiple regression hidden semi-Markov model (MRHSMM) [5], which enables us to control not only output distributions for spectral and F0 but also state duration distributions for phone duration [6].

In this paper, we apply the MRHSMM-based technique to controlling voice quality of synthetic speech. "Voice quality" is a term that has wide meaning [7],[8], and we use the term voice quality as idiosyncratic features, in other words, physiological differences between speakers. We first choose some features for controlling voice quality in accordance with the result of subjective evaluation of speech database, and define a low-dimensional control vector called *voice quality control vector* whose component represents the degree or intensity of a certain feature of voice quality. Then we generate speech by using MRHSMM-based speech synthesis system. From the results of subjective tests, we show that we can control the impression of voice qualities of the synthesized speech to some extent.

2. Multiple Regression HSMM-Based Speech Synthesis

2.1. Multiple Regression HSMM

Here we briefly review the multiple regression HSMM [5]. HSMM [9] is an extension of HMM and has output and state duration probability distributions at each state. We assume that each speech synthesis unit is modeled by an N-state HSMM λ. We also assume that the i-th state output $b_i(\boldsymbol{o})$ and duration distributions $p_i(d)$ are Gaussian distributions characterized by mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$, and mean m_i and variance σ_i^2, respectively,

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2). \tag{2}$$

In MRHSMM, we further assume that the mean parameters of the output and duration distributions at each state are modeled using multiple regression as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} \tag{3}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{4}$$

where

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{5}$$

and $\boldsymbol{v}$ is the control vector, which is a low-dimensional vector whose component v_k represents the degree or intensity of a certain feature in voice quality. In addition, $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are $M \times (L+1)$- and $1 \times (L+1)$-dimensional multiple regression matrices, and M is the dimensionality of $\boldsymbol{\mu}_i$. Then the probability distribution functions at state i are given by

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \boldsymbol{\Sigma}_i) \tag{6}$$

$$p_i(d) = \mathcal{N}(d; \boldsymbol{H}_{p_i}\boldsymbol{\xi}, \sigma_i^2). \tag{7}$$

When training data $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ and corresponding control vectors $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ are given, we can estimate the parameters of MRHSMM, $\boldsymbol{H}_{b_i}, \boldsymbol{\Sigma}_i, \boldsymbol{H}_{p_i}$, and σ_i^2, in ML sense by using EM algorithm. The derivation of re-estimation formulas can be found in [5].

2.2. System Overview

A block diagram of a multiple regression HSMM-based speech synthesis system is shown in Fig. 1. The system consists of two stages, the training stage and the synthesis stage.

In the training stage, we first train context-dependent HSMMs without context clustering for respective speakers independently. Then we apply a shared decision tree context clustering (STC) technique [10] to these models to construct a common tree structure for all speakers. Finally, we obtain a single model with the common tree structure for all speakers by incorporating a voice quality control vector into the re-estimation procedure based on the EM algorithm.

In the synthesis stage, we transform an arbitrarily given text into a sequence of context-dependent phoneme labels, and specify the value of the voice quality control vector corresponding to a desired voice quality. In accordance with the label sequence and voice quality control vector, mean vectors of PDFs of the model are determined by (3) and (4). Finally, a speech parameter sequence is generated from pdf sequences.

3. Experiments

3.1. Choice of Voice Quality Control Vector

As for the training data of MRHSMM, it is desirable to use a database which has variations in voice quality to some extent. On the other hand, since we train the context dependent HSMMs for each speaker in the training stage, it is also desirable that the amount of training data should be large, for example, a few hundred utterances for each speaker. As a consequence, we used the ATR Japanese speech database (Set B) uttered by 6 male (mho, mht, mmy, msh, mtk, and myi) as the training data, and we attempted to control the feature of the voice quality among those speakers.

In order to choose a voice quality control vector, we first collected expression words which express the features of voice quality subjectively by means of questionnaires. The test samples were the recorded speech uttered by six male speakers included in the database. The ATR database sentences consist of ten subsets — subsets A, B, ..., and J, and ten test sentences were chosen for each subject at random from the subset J of 53 sentences. Three subjects were presented with a sample chosen from six speakers in random order, and described their impressions freely about the features of voice quality of the test sample. The number of descriptions was from three to five in each sample. In addition, as the examples of the description, some pairs of the expression words which were introduced in [11], [12] were shown.

We obtained a total of 367 descriptions. From the result of the questionnaires, the voice characteristics of these speakers were not much different, and we could not find completely independent control features. However, we considered frequencies in the use of the description among speakers and similarity in expression, etc., and chose three pairs of features for controlling voice quality, that is v_0: "smooth – nonsmooth," v_1: "warm – cold," v_2: "high-pitched – low-pitched." In the following, we use them as a voice quality control vector $\boldsymbol{v} = [v_0, v_1, v_2]^\top$.

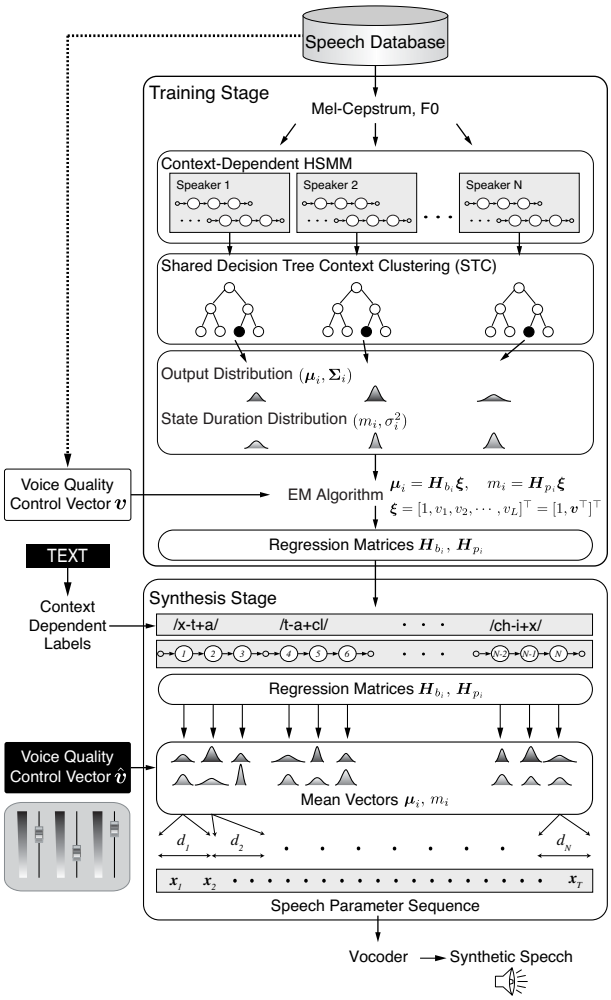

Figure 1: *A block diagram of a multiple regression HSMM-based speech synthesis system.*

3.2. Setting of Voice Quality Control Vector in MRHSMM Training

3.2.1. v_0: "smooth – nonsmooth " and v_1: "warm – cold"

To set the components of the voice quality control vector v_0: "smooth – nonsmooth" and v_1: "warm – cold" in the training of MRHSMM, we used the results of the subjective evaluation tests for the degree or intensity of subjective impression of voice quality. The test samples were the same recorded speech as used in Sect. 3.1. Ten test sentences were chosen for each subject at random. The subjects were presented with a sample, and asked to evaluate the degree of impression of the voice quality on a seven-point scale, that is, 3 for very smooth/warm, -3 for very nonsmooth/cold.

Figure 2 shows the result of the average score for every speaker. We can see that most scores are distributed between -1 and 1, but there is a certain degree of variation of vocal quality in the database. In the training of MRHSMM, we set the average score as the voice quality control vector for each speaker, and did

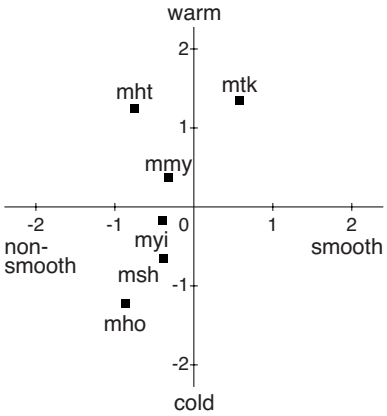

Figure 2: *Subjective impression of voice quality for recorded speech.*

not consider variation within a speaker.

3.2.2. v_2: "low – high"

Since the value of F0 affects the impression of "low-pitched – high-pitched," we calculated the average value of F0 of training data for each speaker. We confirmed that the speakers whose voice quality was often described as being high-pitched in the questionnaires in Sect. 3.1 had high values of average F0 and the speakers whose voice quality were described as being low-pitched had low values of average F0. Then we used the F0 value as the value of v_2: "low – high." Specifically, the value of v_2 in the training of MRHSMM was set to the normalized value for each training sentence given by

$$v_2^{(s,n)} = \frac{2F^{(s,n)} - (F_{\max} + F_{\min})}{F_{\max} - F_{\min}} \tag{8}$$

where $v_2^{(s,n)}$ is the value for the n-th training data of speaker s, $F^{(s,n)}$ is a average value of logarithm of F0 (logF0) of the n-th training data, and $F_{\max}$ and $F_{\min}$ are the highest and lowest values of the average values of whole training data of each speaker, respectively.

Figure 3 shows the distributions of v_0, v_1, and v_2. Note that, in actual MRHSMM training, v_2 was changed in accordance with the value given by (8) for each training sentence. However, for simplicity, we show the average value of whole training data of each speaker in this figure.

3.3. Experimental Conditions

We used 5-state left-to-right MRHSMMs and trained the model using 450 sentences for each speaker. Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. The feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logF0 and their delta and delta-delta coefficients.

3.4. Voice Quality Control with Changing Single Control Parameter

We synthesized speech by changing the values of components of the voice quality control vector separately. We used five kinds

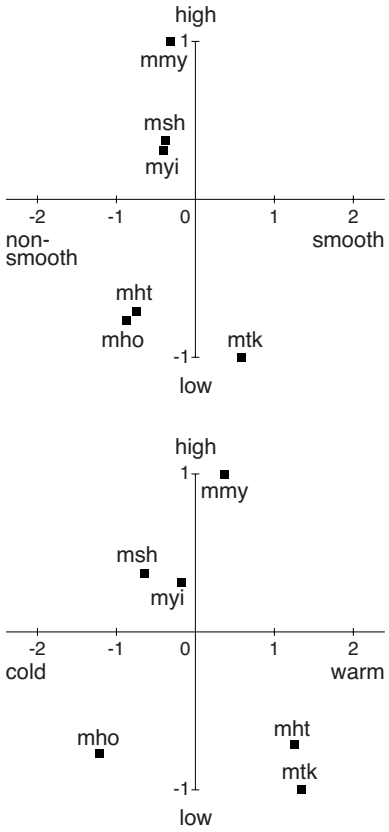

Figure 3: *Distributions of v_0, v_1, and v_2.*

of value in each component, specifically, -1, -0.5, 0, 0.5, and 1. The subjects were presented with a sample chosen from these five cases in random order, and asked to evaluate the degree of impression of voice quality on a seven-point scale, that is, 3 for very smooth/warm/high, -3 for very nonsmooth/cold/low. Ten test sentences were chosen at random for each subject from 53 test sentences that were not contained in the training data. The subjects were the same for evaluation described in Sect. 3.2.1.

Figure 4 shows the results of the average score. We can see that we can control the degree of impression of the voice quality by changing the value of corresponding voice quality control vector.

3.5. Voice Quality Control with Changing Multiple Control Parameters

We next synthesized speech by changing the values of the control vector as shown in Fig. 5. Note that the value of v_2: "low – high" was fixed to zero in this experiment. The evaluation method was almost the same as Sect. 3.2.1 except that each test sample was synthesized speech.

Figure 6 shows the result of the average score. From this figure, we can see that generally the relative distribution of the score of each generated speech sample is maintained and thus we can change these features of voice quality to some extent. We can also see that the degree of impression of "cold – warm" becomes colder as the test samples are close to nonsmooth. This is due to the fact that the components of voice quality vector v_0 and v_1 correlated each other. In fact, Pearson's product-moment correlation coefficient between them is 0.55. This means that the v_0-axis is

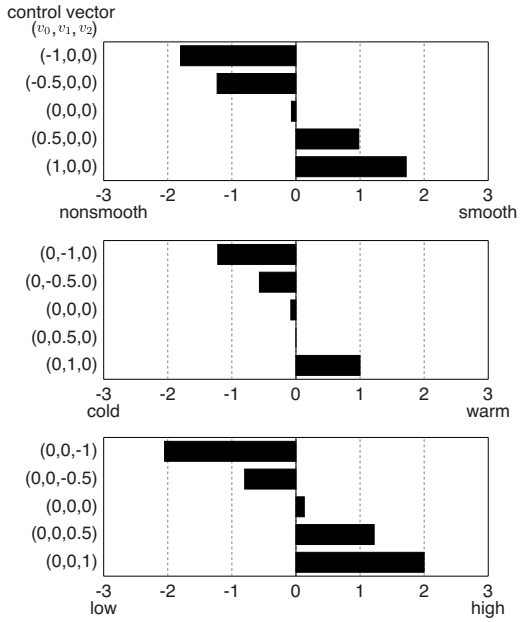

Figure 4: *Subjective evaluation score for variations of single control parameter.*

not orthogonal to the v_1-axis, and we cannot quite control these features of voice quality independently.

4. Conclusions

We have described a technique for controlling voice quality of synthetic speech using multiple regression hidden semi-Markov model (HSMM). We first chose three features for controlling voice qualities experimentally, that is "smooth voice – nonsmooth voice," "cold – warm," and "low-pitched – high-pitched," and then we attempted to control voice qualities of synthetic speech for these features. From the results of several subjective evaluation tests, we have shown that the proposed technique can control voice quality intuitively. The proposed technique is not limited to the Japanese language and can be applied to other languages. Future work will focus on detailed subjective evaluation and investigation of other features of voice quality by using more diverse speakers.

5. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *J. Acoust. Soc. Jpn. (E)*, vol. 21, pp. 199–206, Apr. 2000.

[2] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *Proc. EUROSPEECH 2001*, Sept. 2001, vol. 1, pp. 345–348.

[3] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP 2002*, Sept. 2002, pp. 1269–1272.

[4] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. INTERSPEECH 2004-ICSLP*, Oct. 2004, pp. 1437–1440.

[5] N. Niwase, J. Yamagishi, and T. Kobayashi, "Human walking motion synthesis with desired pace and stride length based on HSMM," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2492–2499, Nov. 2005.

[6] T. Nose, J. Yamagishi, and T. Kobayashi, "Style control of synthetic speech using multiple regression HSMM," in *Proc. INTERSPEECH 2006-ICSLP*, Sept. 2006. (to appear)

[7] J. Laver, *The Phonetic Description of Voice Quality*, Cambridge University Press, Oct. 1980.

[8] H. Kasuya and C.-S. Yang, "Voice quality associated with voice source," *J. Acoust. Soc. Jpn. (Japanese Edition)*, vol. 51, no. 11, pp. 869–875, 1995.

[9] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. INTERSPEECH 2004-ICSLP*, Oct. 2004, vol. 2, pp. 1393–1396.

[10] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 3, pp. 534–542, Mar. 2003.

[11] I. Hieda, "Subjective indices for evaluation of synthesized voice," *Japan Ergonomics Society (Japabese Edition)*, vol. 24, no. 6, pp. 387–393, 1988.

[12] M. Tsuzaki and H. Kawai, "Estimation of a perceptual space for the variation of speech quality in a long term corpus: A method and a pilot result," in *Proc. 2003 Autumn Meeting of ASJ (Japanese Edition)*, Sept. 2003, vol. 1, pp. 237–238, 1-8-28.

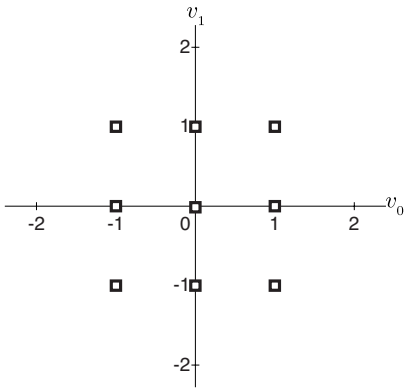

Figure 5: *Desired control vectors.*

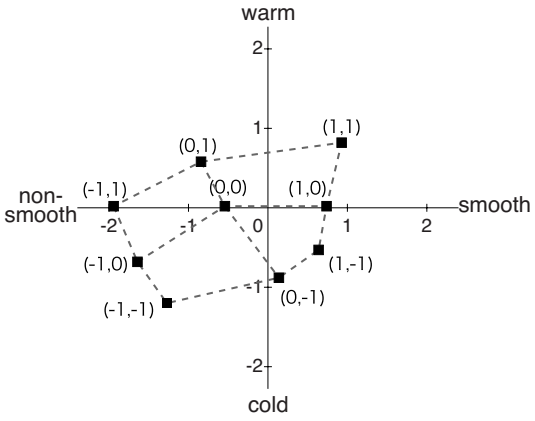

Figure 6: *Subjective evaluation score for variations of multiple control parameters.*