TEXT-TO-SPEECH SYNTHESIS WITH ARBITRARY SPEAKER'S VOICE FROM AVERAGE VOICE

Masatsune Tamura†, Takashi Masuko†, Keiichi Tokuda††, Takao Kobayashi†

†Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan

††Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555 Japan

Email: {mtamura,masuko,tkobayas}@ip.titech.ac.jp, tokuda@ics.nitech.ac.jp

Abstract

This paper describes a technique for synthesizing speech with any desired voice. The technique is based on an HMM-based text-to-speech (TTS) system and MLLR adaptation algorithm. To generate speech of an arbitrarily given target speaker, speaker-independent speech units, i.e., average voice models, is adapted to the target speaker using MLLR framework. In addition to spectrum and pitch adaptation, we derive an algorithm for adaptation of state duration. We demonstrate that a few sentences uttered by a target speaker are sufficient to adapt not only voice characteristics but also prosodic features. Synthetic speech generated from adapted models using only four sentences is very close to that from speaker dependent models trained using a large amount of speech data.

1. INTRODUCTION

Realization of text-to-speech (TTS) synthesizer which can speak with any desired voice is one of the most important issues in the research area of human computer interaction systems. For this purpose, there have been reported various voice conversion techniques [1]-[5]. However, most of these techniques consider spectral conversion mainly and fewer attention is paid to prosodic feature conversion. For example, only average pitch is taken into account when pitch is modified to match the target speaker's prosodic feature [4] [5].

We have proposed an approach which enable the TTS system to change not only spectral voice characteristics [6] but also prosodic features [7]. In the approach, the HMM-based TTS system [8] is used and the voice characteristics of synthetic speech are changed by transforming HMM parameters of the speech units in the MLLR adaptation framework. Although we demonstrated that this approach can generate speech which resembles target speaker's voice in both spectral and prosodic features, adaptation of state duration was not implemented and remained as feature work.

In this paper, we present an adaptation technique of state duration for the HMM-based TTS system. In the HMM-based TTS system, spectrum, pitch, and state duration are modeled simultaneously in a unified framework of HMM [8]. Specifically, spectrum and pitch are modeled by continuous probability distribution HMMs and multi-space probability distribution (MSD) HMMs [9], respectively, and state duration is modeled by multi-dimensional Gaussian distributions. We derive an MLLR algorithm which can be applied to adaptation of state duration. As a result, all of adaptation procedures of spectral and prosodic features including pitch and duration is done in the MLLR framework using the proposed technique together with the previous work of [6] and [7]. To generate speech of an arbitrarily given target speaker, we make speaker-independent speech units, i.e., average voice models, in the training stage, then we adapt the average voice models to the target speaker using the proposed technique.

2. HMM-BASED SPEECH SYNTHESIS SYSTEM

2.1. System overview

A block-diagram of the HMM-based TTS system is shown in Fig.1. The system consists of three stages, the training stage, the adaptation stage, and the synthesis stage.

In the training stage, mel-cepstral coefficients and fundamental frequency are extracted at each analysis frame as the static features from multi-speaker speech database. Then, the dynamic features, i.e., delta and delta-delta parameters, are calculated from the static features. Spectral parameters and pitch observations are combined into one observation vector frame by frame, and speaker independent phoneme HMMs, which we refer to as the average voice HMMs, are trained using the observation vectors. To model variations of spectrum, pitch and duration, phonetic and linguistic contextual factors, such as phoneme identity factors and stress related factors, are taken into account [8]. Spectrum and pitch are modeled by multi-stream HMMs and output distributions for spectral and pitch parts are continuous probability distribution and multi-space probability distribution (MSD) [9], respectively. Then, a decision tree based context clustering technique [10] [11] is separately applied to the spectral and pitch parts of the context dependent phoneme HMMs. Finally, state durations are modeled by multi-dimensional Gaussian distributions, and the state clustering technique is also applied to the duration distributions [12].

In the adaptation stage, the average voice HMMs

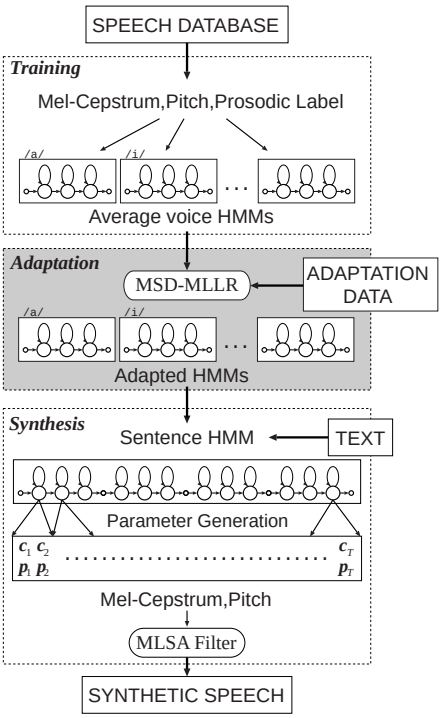


Figure 1: A block-diagram of HMM-based speech synthesis system.

are adapted to a target speaker using a small amount of speech data uttered by the target speaker. We use the maximum likelihood linear regression (MLLR) algorithm [13] and MSD-MLLR algorithm [7], which is an extension of MLLR to multi-space probability distribution HMM, for spectrum and pitch adaptation, respectively. In this paper, we also extend the MLLR algorithm to duration adaptation as shown in section 3. The extended MLLR technique is applied to the mean vectors of the distributions in each stream of the average voice HMMs. As a result, distributions for spectral, pitch, and duration parameters are adapted simultaneously.

In the synthesis stage, first, an arbitrarily given text to be synthesized is transformed into a context dependent phoneme label sequence. According to the label sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating adapted phoneme HMMs. From the sentence HMM, phoneme durations are determined based on state duration distributions [12]. Then, spectral and pitch parameter sequences are generated using the algorithm for speech parameter generation from HMMs with dynamic features [14]. Finally, by using MLSA filter [15], speech is synthesized from the generated mel-cepstral and pitch parameter sequences.

2.2. Duration modeling and synthesis

In the HMM-based speech synthesis system, state duration densities were modeled by single Gaussian distributions,

$$p_i(d_i) = \mathcal{N}(d_i|\mu_i, \sigma_i^2), \tag{1}$$

where d_i is the number of frames for state i. The mean μ_i and variance σ_i^2 of Gaussian distributions for state i are calculated on the trellis obtained in embedded training stage [12]. Then state duration distributions are converted to multi-dimensional phoneme duration distributions. The number of dimensions of duration distributions is equal to the number of states and each dimension represents the duration distributions for each state. Phoneme duration models are clustered using a decision-tree based context clustering technique based on MDL criterion [11]. In the synthesis stage, we first select the phoneme duration distributions by traversing the tree. Then state duration d_i is determined by rounding μ_i off to integer number.

3. SPEAKER ADAPTATION BASED ON MLLR

To generate speech with an arbitrarily given target speaker's voice, we adapt the speaker independent models, i.e., average voice models, to the target speaker. We use a transformation-based model adaptation approach. Standard MLLR algorithm [13] is used for spectral adaptation and MSD-MLLR [7] ,which is extended MLLR for MSD-HMMs, is used for pitch adaptation. Here we derive a similar algorithm for duration model adaptation.

3.1. Spectrum and pitch adaptation using MSD-MLLR

Spectral and pitch adaptation is performed using MLLR [13] and MSD-MLLR [7], respectively. In MSD-MLLR algorithm, the new mean vector $\hat{\boldsymbol{\mu}}_{ig}$ of gth space of state i is estimated by

$$\hat{\boldsymbol{\mu}}_{ig} = \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig} \tag{2}$$

where $\boldsymbol{\xi}_{ig} = [1, \boldsymbol{\mu}_{ig}^T]^T$, and $\boldsymbol{\mu}_{ig}, \boldsymbol{U}_{ig}$ is the mean vector and the covariance matrix of output probability $\mathcal{N}_{ig}(\boldsymbol{x})$. Regression matrix $\boldsymbol{W}_{ig}$ is calculated by maximizing the likelihood of given adaptation samples.

When the amount of adaptation data is limited, the transformation matrices $\boldsymbol{W}_{ig}$ should be tied across several Gaussian distributions. We construct a regression class tree to group the distributions. By doing this, we can estimate transformation matrices for distributions which there were no observations at all and estimate transformation matrices according to the amount of adaptation data.

Regression class tree is constructed by recursively applying a binary splitting algorithm based on LBG for all Gaussian distributions. In the binary tree, each leaf has a distribution, and all the distributions below the lowest node in which the amount of adaptation data is larger than the prescribed threshold are adapted using the same transformation matrix.

3.2. MLLR for duration distribution

Mean adaptation of MLLR is based on affine transformation. Let μ_i, σ_i^2 be the mean and the variance of Gaussian duration distribution $\mathcal{N}_i(d)$ for the state i, respectively. For given adaptation samples $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$, the new mean vector $\hat{\mu}_i$ is estimated by

$$\hat{\mu}_i = \boldsymbol{W}_i \, \xi_i \tag{3}$$

where $\xi_i = [1 \, \mu_i]^T$. To derive a maximum likelihood estimation of transformation matrix $\boldsymbol{W}_i$, we define the probability $\gamma_t(i)$ of being in state i at time t, given the model λ and the observation $\boldsymbol{O}$, as follows:

$$\begin{aligned} \gamma_t(i) &= P(q_t = i|\boldsymbol{O}, \lambda) \\ &= \frac{1}{P(\boldsymbol{O}|\lambda)} \sum_{\text{all}\boldsymbol{q}} P(\boldsymbol{O}, q_t = i|\lambda), \end{aligned} \tag{4}$$

where $\boldsymbol{q} = (q_1, q_2, \cdots, q_T)$ is the possible state sequences of length T. Using $\gamma_t(i)$, we also define a probability $\chi_{t_0,t_1}(i)$ of occupying state i from time t_0 to t_1 as

$$\chi_{t_0,t_1}(i) = (1 - \gamma_{t_0-1}(i)) \cdot \prod_{t=t_0}^{t_1} \gamma_t(i) \cdot (1 - \gamma_{t_1+1}(i)), \tag{5}$$

where $\gamma_{-1}(i) = \gamma_{T+1}(i) = 0$.

Then the regression matrix $\boldsymbol{W}_i$ is found by solving the following equation

$$\begin{aligned} &\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i) \frac{1}{\sigma^2(i)} (t_1 - t_0 + 1) \, \xi^T \\ &\qquad = \sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i) \frac{1}{\sigma^2(i)} \boldsymbol{W}_i \, \xi \, \xi^T. \end{aligned} \tag{6}$$

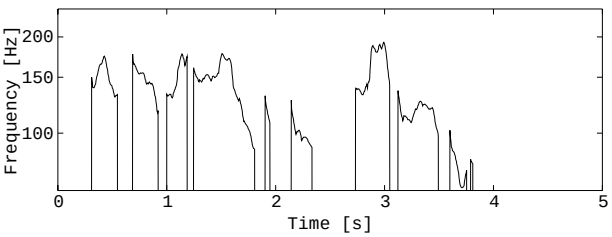
Regression matrix $\boldsymbol{W}_i$ should be tied across several Gaussians as well as spectral and pitch adaptation. The algorithm of tying and estimating the regression matrix is based on regression class tree described in section 3.1.

4. EXPERIMENTS

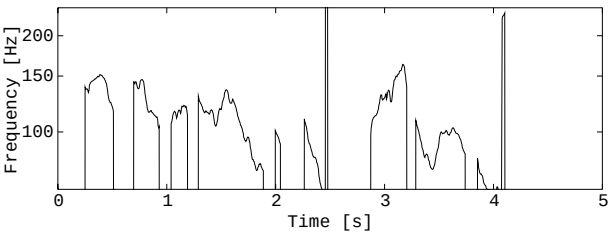
4.1. Experimental conditions

We used phonetically balanced sentences from ATR Japanese speech database for training and adaptation. Based on phoneme labels and linguistic information included in the database, we made context dependent phoneme labels. We used 42 phonemes including silence and pause. The details of contextual factors are shown in [8].

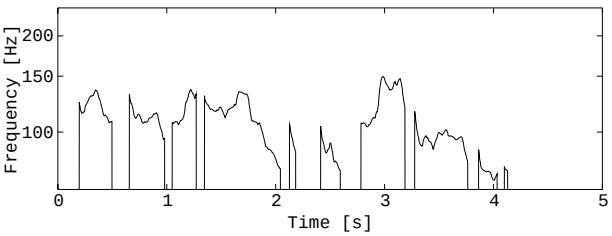
Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. Pitch values were obtained using ESPS `get_f0` program [16]. Delta and delta-delta pitch parameters were calculated only within voiced regions,



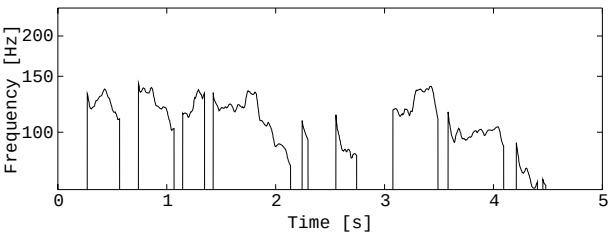
(1) average voice models



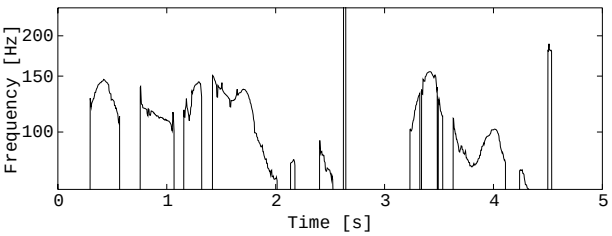
(2) speaker dependent models



(3) adapted models using 4 sentences.



(4) adapted models using 50 sentences.



(5) natural speech

Figure 2: comparison of generated pitch contours

and the frames where delta or delta-delta pitch parameters were not computable because of the boundaries of voiced and unvoiced regions were treated as unvoiced. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. We used 5-state left-to-right models in which the spectral part of each state is modeled by a single diagonal Gaussian output distribution.

The average voice HMMs (SI models) were trained using 5 male speakers' speech data, 400 sentences for each speaker. The states of the models were clustered using a decision tree based context clustering technique with MDL-criterion [11] . The total number of states in SI models is 3765 for spectral part and 12761 for pitch part, and the number of Gaussians for duration distribution is 6318. We chose a male speaker MHT from the database as the target speaker, who was not included in the training speakers of average voice HMMs. We also trained HMMs using 450 sentences uttered by MHT (SD models) to compare with the adapted models. The total number of states in SD models is 906 for spectral part and 1894 for pitch part, and the number of Gaussians for duration distribution is 1021. Thresholds of samples for traversing regression class tree were set to 1500 for spectral stream, 100 for pitch stream and 100 for duration distributions.

4.2. Prosodic parameter generation

Figure 2 shows an example of the generated pitch contours from a Japanese sentence /he-ya-i-ppa-i-ni-ta-ba-ko-no-no-o-mu-ga-ta-chi-ko-me-pau-yu-ru-ya-ka-ni-u-go-i-te-i-ru/ ("Smoke of tobocco fill the whole room, and is moving gently," in English) which is not included in the training sentences. In Fig. 2, x-axis and y-axis represents time and frequency. Fig. 2 (a), (b), (e), and (d) show the the pitch contours generated from average voice HMMs, speaker dependent HMMs, speaker adapted HMMs using 4 sentences, and speaker adapted HMMs using 50 sentences, respectively. Fig. 2 (e) show the pitch contour extracted from natural speech. There are some pitch halving and doubling errors in (e) and (b) because we did not collect the pitch extraction errors manually in the training stage.

From this figure, it can be seen that the pitch contour generated from SA HMMs becomes closer to that extracted from original speech. It has been observed by informal listening tests that the syntheitc voice resembles target speaker's voice in both spectral and prosodic features.

5. CONCLUSION

In this paper, we described a technique for adapting voice characteristics and prosodic features of HMM-based TTS system to an arbitrarily given target speaker. To convert the duration model parameters, we derived the MLLR algorithm for duration models. We have shown that synthetic speech generated from adapted models using average voice models becomes closer to the target speaker's voice. Our future work is subjective and objective evaluation of this technique.

6. References

[1] M. Hashimoto and N. Higuchi, "Training data selection for voice conversion using speaker selection and vector field smoothing," in *Proc. ICSLP-96*, Oct. 1996, pp. 1397–1400.

[2] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. ICASSP-97*, Apr. 1997, pp. 1611–1614.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[4] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP-98*, May 1998, pp. 285–288.

[5] T. Toda, J. Lu, H. Saruwatari, and K. Shikano, "Straight-based voice conversion algorithm based on gaussian mixture model," in *Proc. ICSLP-2000*, Oct. 2000, vol. 3, pp. 279–282.

[6] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, Nov. 1998, pp. 273–276.

[7] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP-2001*, May 2001.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH-99*, Sept. 1999, pp. 2374–2350.

[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP-99*, Mar. 1999, pp. 229–232.

[10] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Mar. 1994, pp. 307–312.

[11] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EUROSPEECH-97*, Sept. 1997, pp. 99–102.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP-98*, Dec. 1998, pp. 29–32.

[13] C.J.Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[14] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP-95*, May 1995, pp. 660–663.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, Mar. 1992, pp. 137–140.

[16] Entropic Research Laboratory Inc, *ESPS Programs Version 5.0*, 1993.