



Speaker-independent HMM-based Voice Conversion Using Quantized Fundamental Frequency

Takashi Nose, Takao Kobayashi

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

takashi.nose@ip.titech.ac.jp, takao.kobayashi@ip.titech.ac.jp

Abstract

This paper proposes a segment-based voice conversion technique between arbitrary speakers with a small amount of training data. In the proposed technique, an input speech utterance of source speaker is decoded into phonetic and prosodic symbol sequences, and then the converted speech is generated from the pre-trained target speaker's HMM using the decoded information. To reduce the required amount of training data, we use speaker-independent model in the decoding of the input speech, and model adaptation for the training of the target speaker's model. Experimental results show that there is no need to prepare the source speaker's training data, and the proposed technique with only ten sentences of the target speaker's adaptation data outperforms the conventional GMM-based one using parallel data of 200 sentences.

Index Terms: voice conversion, segment-based mapping, HMM-based speech synthesis, speaker adaptation, average voice model

1. Introduction

Voice conversion (VC) is a technique for changing nonlinguistic information appearing in speech, and one of the common purposes is to convert the speaker individuality of input speech to that of another speaker. In this context, a variety of techniques have been proposed [1]. In these techniques, GMM-based statistical spectral mapping [2, 3] is one of the typical approaches. This technique enables the continuous mapping of acoustic features by soft clustering, and the quality can be improved by incorporating the dynamic features and global variance (GV) [4]. However, this technique only models the relationship between frames of source and target speakers, and cannot convert the dynamic characteristics of speaker individuality included in the segmental level such as phonemes. In contrast, segment-based feature mapping, shown in Fig. 1, converts the dynamic characteristics as well as the static ones, which is important to improve the individuality of the converted speech [5].

In our previous study [6], we have proposed segment-based VC using HMM-based speech recognition and synthesis. The basic idea comes from the HMM-based phonetic vocoder [7], which was proposed for very low bit-rate speech coding. To model and convert not only the spectral feature but also the fundamental frequency (F_0), we used multi-space distribution HMM (MSD-HMM) [8] with quantized F_0 context [9]. In this technique, an input speech utterance is decoded into the phoneme, duration, and F_0 symbol sequences which represent the linguistic and prosodic information. Then, the converted speech is generated from the pre-trained target speaker's MSD-HMM using decoded information. By means of decoding the speech, we can relax the dependency of the source speaker's characteristics in the conversion [10]. However, this technique requires a certain amount of training data of source and target

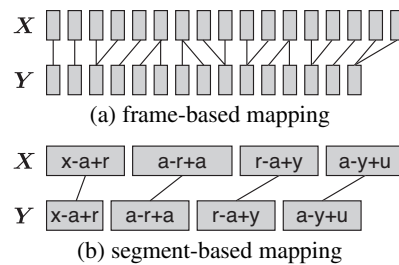


Figure 1: Two types of mapping approaches of feature vector sequences X and Y for VC.

speakers, typically more than ten minutes, and it is not easy to convert speech using only a few minutes data while maintaining its naturalness.

In this paper, we describe context-dependent HMM-based VC using speaker-independent model and speaker adaptation. We focus on the case where the phonetic transcription is given as same in [11]. In this case, the decoding process of the input speech becomes a simple segmentation problem. Therefore, we expect to use the speaker-independent model in the decoding, which means that there is no need to prepare any training data of the source speaker. In the synthesis part, the amount of target speaker's training data can be also reduced by using model adaptation technique which have been well discussed in the text-to-speech synthesis [12].

2. Context-dependent HMM-based VC with quantized F_0 context

In our HMM-based VC technique [6], the input speech of the source speaker is decoded into phonetic and prosodic symbol sequences, and the converted speech is generated from the pre-trained target speaker's acoustic HMMs with the decoded information using the HMM-based speech synthesis framework [13]. In the HMM-based speech synthesis, the multi-space distribution HMM (MSD-HMM) [8] is used to appropriately model the F_0 distribution both for voiced and unvoiced regions.

2.1. Quantized F_0 symbol for prosodic context of acoustic model

To model the F_0 sequence using context-dependent MSD-HMM, the labeling of prosodic information is necessary for every utterance. However, it is not always possible to automatically extract the prosodic information such as accent, tone, and intonation with high accuracy. To automatically generate the prosodic labels, we use the quantized F_0 symbols [9]. We assume that the log F_0 values of source and target speakers' speech follow a normal distribution within each utterance [10].

In the F_0 quantization, we first standardize the log F_0 dis-

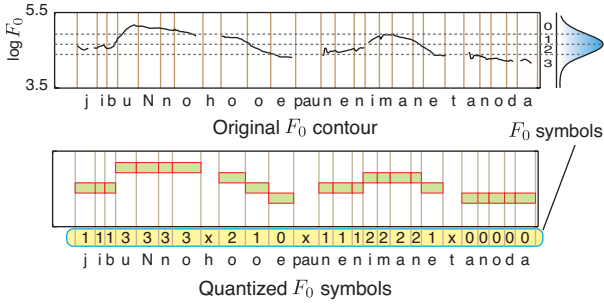


Figure 2: Example of F_0 quantization into four levels.

tribution into $\mathcal{N}(0, 1)$ for each utterance using global mean and variance of $\log F_0$ calculated from the utterance. Then, the F_0 symbol s_p for each phone unit p is obtained by quantizing the mean value of $\log F_0$, f_p , of each phone into a discrete value as follows.

$$s_p = Q[f_p], \quad s_p \in \{0, 1, \dots, M-1\}, \quad (1)$$

where $Q[\cdot]$ denotes an operation of scalar quantization, and M is the number of the quantization levels. We set the points that equally divide the region $[-2, 2]$ into M levels as the quantization boundaries. An example of four-level F_0 quantization is illustrated in Fig. 2. The detail of the proposed VC system is described in the next section.

3. HMM-based VC using SI model and model adaptation

3.1. Decoding of input speech using SI model

In the speaker-dependent (SD) case of our previous study [6], we decoded the input speech of a source speaker using the SD model trained with a sufficient amount of training data. On the other hand, when the phonetic transcription is given, only the phoneme segmentation of input speech is required for the F_0 symbolization. In this study, we therefore use the speaker-independent (SI) model instead of the SD model for the segmentation. It should be noted that SI model does not work well when the phonetic information is not given because the speaker individuality of SI model does not always match the SD model.

3.2. Synthesis of converted speech using model adaptation

As well as in the decoding part, we use the speaker-independent approach in the synthesis part. We use the technique of speaker adaptation from average voice model [14] which is effective when the amount of training data of the target speaker is very limited in the HMM-based speech synthesis. For simplicity, we call the average voice model SI model in the rest of this paper.

Figure 3 shows a block diagram of the proposed VC system. In the decoding part, sequences of phoneme, mel-cepstrum, and F_0 are extracted from the input speech of a source speaker. Then, phoneme durations are obtained by phoneme alignment with SI model and used in the F_0 quantization. In the synthesis part, a label sequence for speech synthesis is generated using the phonemes and F_0 symbols obtained in the decoding part. Then, the converted speech parameters are generated from the pre-trained speaker-adapted (SA) MSD-HMMs. In this study, the phoneme duration is not transmitted to the synthesis part, and the duration is determined from the duration distribution of the SA model. As a result, there is no need to convert the duration features in the proposed technique. Finally, the converted speech waveform is synthesized using a mel log spectrum approximation (MLSA) filter [15] as the synthesis filter.

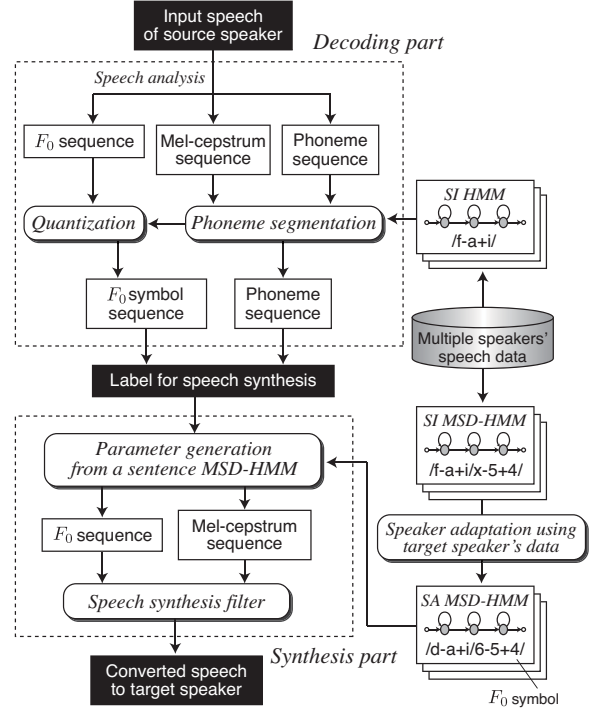


Figure 3: Overview of proposed VC system.

4. Experiment

4.1. Experimental conditions

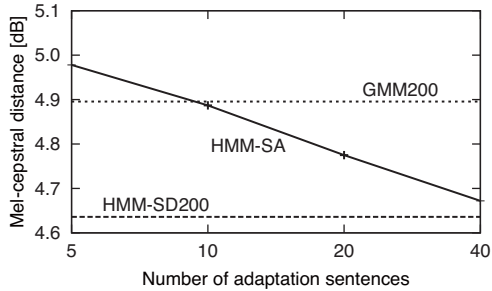
We used the ATR Japanese speech database set B which includes ten professional narrators' speech data. From the database, one male and one female speakers, MSH and FYM, were chosen as the source speakers, and one female speaker, FKN, was chosen as the target speaker. The training and test sentences were phonetically balanced 200 sentences (subset A to D) and 53 sentences (subset J), respectively. For the SI models used in decoding and synthesis parts, six speakers not included in the source and target speakers were chosen. The number of training sentences were 450 sentences per speaker, 2700 sentences in total.

Speech signals were sampled at a rate of 16kHz, and the STRAIGHT analysis [16] was used for spectral feature extraction with a 5-ms shift. In the decoding part, we used the feature vector consisting of 25 mel-cepstral coefficients including the zeroth coefficient and their delta coefficients. As a result, the total dimensionality of the feature vector was 50. In the synthesis part, we constructed 52 dimensional feature vector by adding the $\log F_0$ and its delta to the feature vector used in the decoding part. For the F_0 extraction, we used an instantaneous-frequency-based technique [17].

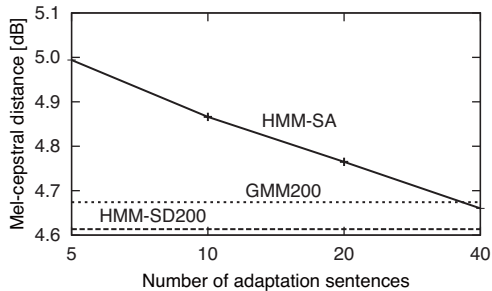
For the acoustic model, we used hidden semi-Markov model (HSMM) [18] that has an explicit duration distribution. In the decoding part, we simply converted HSMM to HMM in advance by calculating the transition probabilities from the mean parameters of duration distributions of HSMM using

$$a_{ii} = 1 - 1/d_i, \quad a_{i,i+1} = 1/d_i, \quad (2)$$

where $a_{i,i+1}$ is transition probability from state i to $i+1$, and d_i is mean parameter of i -th state duration. We used 5-state left-to-right triphone model for the decoding part, and 5-state left-to-right model with triphone and quantized F_0 context for the synthesis part. As the quantized F_0 context, we used preceding, current, and succeeding F_0 symbols. The number of quantization levels of F_0 was set to eight on a basis of the previous



(a) male-to-female conversion



(b) female-to-female conversion

Figure 4: Spectral distortion between original and converted speech with different number of adaptation sentences.

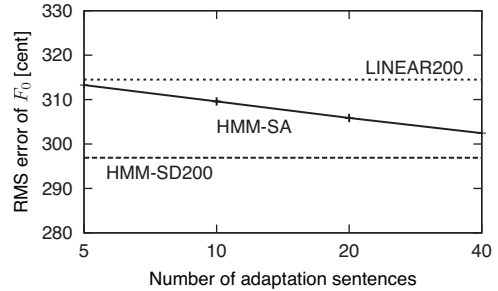
study [6]. The output distribution in each state of the MSD-HSMM was modeled by a single Gaussian density function, and the covariance matrices were assumed to be diagonal. In the context clustering for parameter tying, a decision tree is automatically constructed based on a minimum description length (MDL) criterion [19]. As the model adaptation algorithm, we used a combined method of constrained structural maximum a posteriori linear regression (CSMAPLR) and MAP adaptation based on HSMM [12]. To mitigate the effect of over-smoothing in the speech synthesis, we used a parameter generation algorithm considering GV [20] in the subjective evaluation test¹.

4.2. Performance evaluation of VC from arbitrary speakers

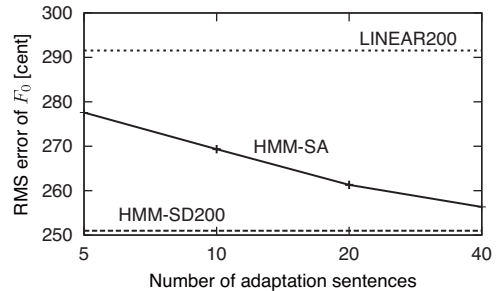
We first conducted objective evaluation of the conversion performance when a sufficient amount of training data is available for the target speaker, and focus on the effect of SI model of decoding part in this experiment. The target speaker's SD model was trained using 200 sentences. As the objective measure of spectral and F_0 similarity to the original features of the target speaker, we used mel-cepstral distance and root mean square (RMS) error of F_0 . To calculate the distortion frame by frame, time-alignment was conducted in advance by dynamic time warping (DTW) using the spectral features. We evaluated three types of techniques: conventional GMM-based spectral conversion using a maximum likelihood criterion with linear transformation of F_0 (GMM+L) [4], HMM-based conversion using the SD model in the decoding part (HMM-SD), and the proposed VC using the SI model in decoding (HMM-SI). In GMM+L, we used diagonal covariance and cross-covariance matrices, and set the number of mixtures to 512 based on preliminary experimental results. In GMM+L and HMM-SD, we used parallel training data of 200 sentences of source and target speakers as used in the training of the target speaker's model in HMM-SI.

Experimental results are shown in Table 1. We found that

¹Speech samples used in the evaluation tests are available at <http://www.kbys.ip.titech.ac.jp/demo/is2010/nose/>.



(a) male-to-female conversion



(b) female-to-female conversion

Figure 5: RMS error of F_0 between original and converted speech with different number of adaptation sentences.

Table 1: Average mel-cepstral distance [dB] and F_0 RMS error [cent] between original and converted speech.

source speaker		GMM+L	HMM-SD	HMM-SI
FYM	spec	4.67	4.61	4.61
	F_0	291.5	251.9	251.0
MSH	spec	4.90	4.63	4.63
	F_0	314.5	298.3	296.9

the decoding with SI model dose not degrade the conversion performance compared to the SD model, and HMM-based conversion shows better performance than GMM. We also found that the F_0 similarity of HMM-SI is slightly better than those of HMM-SD. A possible reason is that the training data for the source speaker's model was 200 sentences and relatively small, whereas the SI model was trained with a larger amount of training data of 2700 sentences, and this affected the accuracy of phoneme segmentation in F_0 symbolization process.

4.3. Performance evaluation of VC between arbitrary speakers

The experimental results of Sect. 4.2 indicate that the SI model can be used instead of the SD model in the decoding process of VC. Therefore, in this experiment, we used the SI model in the decoding, and focus on the conversion performance when using speaker adaptation for the model training of the target speaker.

4.3.1. Objective evaluation results

First, we objectively evaluated the conversion performance with different amount of target speaker's training data of 5, 10, 20, and 40 sentences. These sentences were chosen randomly from the 200 sentences used for the SD model training. To alleviate the dependency of the choice of the adaptation data, we repeated the same experiments ten times by changing the sentences of adaptation data. For comparison, we also evaluated the GMM+L and HMM-SI with 200 sentences as described in Sect 4.2.

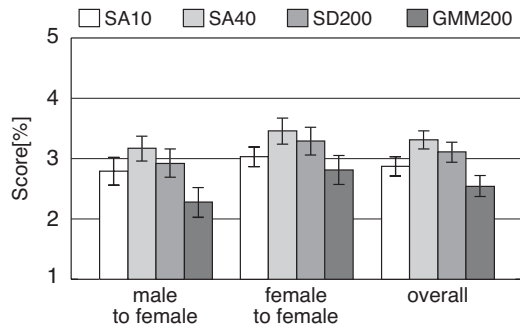


Figure 6: Results of similarity test on speaker individuality.

Figures 4 and 5 show the results. Both in the spectral and F_0 conversions, the distortions decreased as the number of adaptation sentences increases, and the performance became close to that of the speaker-dependent case (HMM-SD200). As for the spectral conversion, the proposed technique (HMM-SA) outperformed the GMM-based mapping (GMM200) when the number of adaptation sentences was 10 for male-to-female conversion and 40 for female-to-female conversion, respectively. It is also found that the results of the proposed technique are less sensitive to the variation of source speakers than the GMM-based mapping. As for the F_0 conversion, our technique outperformed the linear transformation even if only five adaptation sentences were available for the target speaker.

4.3.2. Subjective evaluation results

Next, we conducted a subjective evaluation test on speaker similarity of converted speech. We evaluated the proposed technique with 10 and 40 sentences (SA10 and SA40), speaker-dependent HMM-based one with 200 sentences (SD200), and GMM-based one with parallel 200 sentences (GMM200). In the GMM-based technique, speaking rate was converted by linear interpolation of frames for spectral and F_0 features using the ratio of total utterance lengths of source and target speakers' training data. To correctly obtain the utterance length without silence, we used the time information from label files for the training of the SD HMM. In the HMM-based technique, we did not use any training data of the source speaker. Six participants listened to pairs of test and reference samples and evaluated the speaker similarity in a five-point scale from "1: bad" to "5: excellent." For each participant, we randomly chose fifteen sentences as the test samples from 53 sentences. We used the target speaker's vocoded speech of the test sentences as reference samples. The scores of male-to-female, female-to-female, and overall are shown in Fig. 6 with confidence intervals of 95%.

From the results, the proposed technique with ten adaptation sentences of the target speaker gives better performance than GMM-based one with 200 parallel sentences. It is also seen that the scores of SA40 is slightly better than SD200. A possible reason is that the amount of training data of the target speaker was not enough for the speaker-dependent model, and phoneme duration of some samples was perceived as unnatural.

5. Conclusions

We have proposed an HMM-based voice conversion technique which requires a less amount of training data than our conventional speaker-dependent technique. Experimental results showed that the speaker-independent model is feasible for the decoding of the input speech into the prosodic symbol sequence under condition where the phonetic transcription is known. It was also found that the proposed technique gives better performance with a significantly smaller amount of non-parallel training data than GMM-based one which requires parallel data. In

future work, we will examine the performance of our technique using different types of speech such as emotional or singing voices.

6. Acknowledgements

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063 and 21800020.

7. References

- [1] Y. Stylianou, "Voice transformation: a survey," *Proc. ICASSP 2009*, pp. 3585–3588, 2009.
- [2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.
- [3] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP 1998*, vol. 1, 1998.
- [4] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.
- [5] M. Abe, "A segment-based approach to voice conversion," in *ICASSP 91*, 1991, pp. 765–768.
- [6] T. Nose, Y. Ota, and T. Kobayashi, "HMM-based voice conversion using quantized F_0 context," *IEICE Trans. Inf. & Syst.*, 2010, (to appear).
- [7] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura, "A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques," in *Proc. ICASSP 1998*, vol. 2, 1998, pp. 609–612.
- [8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.
- [9] T. Nose, K. Ooki, and T. Kobayashi, "HMM-based speech synthesis with unsupervised labeling of accentual context based on F_0 quantization and average voice model," in *Proc. ICASSP 2010*, Mar. 2010, pp. 4622–4625.
- [10] T. Nose and T. Kobayashi, "Robust voice conversion technique based on context-dependent HMM with F_0 quantization," in *Proc. 7th ISCA Speech Synthesis Workshop (SSW7)*, 2010, (to appear).
- [11] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1301–1312, 2006.
- [12] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Iso-gai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.
- [13] 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*, Sept. 1999, pp. 2347–2350.
- [14] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *Proc. EUROSPEECH*, Sept. 2001, pp. 345–348.
- [15] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *IECE Trans. A (Japanese Edition)*, vol. J66-A, no. 2, pp. 122–129, Feb. 1983.
- [16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "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.
- [17] T. Tanaka, T. Kobayashi, D. Arifianto, and T. Masuko, "Fundamental frequency estimation based on instantaneous frequency amplitude spectrum," in *Proc. ICASSP 2002*, vol. 1, 2002.
- [18] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.
- [19] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 2, pp. 79–86, Mar. 2000.
- [20] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.