HMM-Based Robust Voice Conversion Using Adaptive F0 Quantization

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 an HMM-based voice conversion (VC) technique with quantized F0 symbol context using adaptive F0 quantization. In the HMM-based VC, an input utterance of a source speaker is decoded into phonetic and prosodic symbol sequences, and the converted speech is generated using the decoded information from the pre-trained target speaker’s phonetically and prosodically context-dependent HMM. In our previous work, we generated the F0 symbol by quantizing the average log F0 value of each phone using global mean and variance parameters calculated from the training data. In this study, these statistical parameters are obtained sentence-by-sentence, and this adaptive approach enables the more robust F0 conversion than the conventional one. Objective and subjective experimental results for English and Japanese speech show that the proposed adaptive quantization technique gives better F0 conversion performance than the conventional one. Moreover, the HMM-based VC is significantly robust for the variation of the source speaker’s individuality compared to the GMM-based one.

1. Introduction
Voice conversion (VC) is a technique for changing the acoustic characteristics of input speech with the linguistic information unchanged. One of the major purposes is to convert the speaker individuality of a source speaker’s speech to a target speaker’s one, and in this context, a wide variety of techniques have been proposed [1]. In these techniques, GMM-based statistical spectral conversion [2, 3] is one of the typical and efficient approaches. The advantage of the GMM-based technique is that it enables the continuous mapping of acoustic features between speakers based on soft clustered conversion functions. Moreover, the quality can be improved by incorporating the dynamic features and global variance (GV) [4]. However, there is also a limitation that only the static relationship between pre-aligned frames of source and target speakers’ parallel speech data is modeled, and it is difficult to convert the dynamic characteristics of speaker individuality appearing in the segmental level features such as phonemes.

On the other hand, segment-based feature mapping technique (e.g., [5]) converts the dynamic characteristics as well as the static ones, which is important to improve the conversion performance of the individuality. Although the conventional segment-based technique works well in the case where a large amount of training data of both source and target speakers is available [5], the performance is significantly degraded when the available training data is limited. Recently, alternative segment-based techniques have been proposed for nonparallel VC [6, 7]. However, these techniques focused on the conversion without parallel training data of source and target speakers, and have not outperform the GMM-based technique. A possible reason is that they used phone or diphone for representing the acoustic unit, and these contextually poor representation would degrade the conversion performance.

To overcome the problem, we have proposed segment-based VC using a context-dependent HMM-based framework in our previous work [8]. The basic idea has came from the HMM-based phonetic vocoder [9], which was proposed for very low bit-rate speech coding. To model and convert not only the spectral feature but also the fundamental frequency (F0), we used multi-space distribution HMM (MSD-HMM) [10] with quantized F0 context [11]. In this technique, an input speech utterance is decoded into the phoneme, duration, and F0 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.

In this paper, we propose an HMM-based VC technique using adaptive F0 quantization to improve the F0 conversion performance. In the previous work [8], the F0 values of each speaker were assumed to follow a single Gaussian distribution in the quantization of F0, and the global mean and variance parameters were used for decoding the F0 contour of every sentence. However, this assumption does not always work well because the mean and variance parameters of the input F0 value of a source speaker often differ from those of the training data depending on various factors. As a result, this mismatch sometimes degrades the F0 conversion performance, and is an obstacle of robust VC. In the proposed technique, the F0 values of each sentence are assumed to follow the Gaussian distribution. The mean and variance parameters used in the F0 quantization are calculated for the input sentence itself. We expect that this adaptive approach decreases the mismatch between the distribution of observed F0 values and statistical parameters, which would improve the conversion performance of F0. We assess the performance of the proposed VC for English and Japanese speech, and also examine the robustness for the variation of the source speaker in the spectral and F0 conversion.

2. HMM-based VC using quantized F0 context
In our HMM-based VC technique [8], the input speech of the source speaker is decoded into phonetic and prosodic symbol sequences, and the converted speech is generated from the pretrained target speaker’s acoustic HMMs with the decoded information using the HMM-based speech synthesis framework [12]. In the HMM-based speech synthesis, the multi-space distribution HMM (MSD-HMM) [10] is used to appropriately model the F0 distribution both for voiced and unvoiced regions.

2.1. Context generation using adaptive F0 quantization
To model the F0 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 F0 symbols [11]. We assume that the log F0 values of source and target speakers’ speech follow a single Gaussian distribution within each utterance.

In the F0 quantization, we first standardize the log F0 distribution into $N(0, 1)$ for each utterance using mean and variance parameters of log F0 calculated from the utterance itself. Then, the F0 symbol $s_p$ for each phone unit $p$ is obtained by quantizing the mean value of log F0, $f_p$, of each phone into a discrete value as follows.

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

where $Q[.]$ denotes an operation of scalar quantization, and $M$ is the number of the quantization levels. We set the points that equally divide the range $[-2, 2]$ into $M$ levels as the quantization boundaries. In this study, we call this technique adaptive F0 quantization. An example of four-level F0 quantization is illustrated in Fig. 1. The detail of the proposed VC system is described in the next section.

2.2. Overview of the proposed VC system

A block diagram of the proposed technique is shown in Fig. 2. In the decoding part, sequences of phoneme, mel-cepstrum, and F0 are extracted from the input speech of a source speaker. Then, phoneme durations are obtained by phoneme alignment between the phoneme and the mel-cepstrum sequences, and used in the F0 quantization. For this purpose, we train the source speaker’s triphone HMMs in advance with a sufficient amount of training data. An F0 symbol sequence is also obtained using the standardization and quantization described in Sect. 2.1.

In the synthesis part, a label sequence for speech synthesis is generated using the phonemes and F0 symbols obtained in the decoding part. We adopt a context-dependent label set in which the preceding and succeeding phonemes and F0 symbols are taken into account. Then, the converted speech parameters are generated from the target speaker’s context-dependent MSD-HMMs trained using a sufficient amount of speech data. In this study, the phoneme duration is not transmitted to the synthesis part, and the duration is determined from the duration distribution of the target speaker’s model. As a result, there is no need to convert the duration features. In the training of the target speaker’s MSD-HMM, the F0 standardization and quantization are applied to the labeling of F0 symbols. Finally, the converted speech is synthesized using a mel log spectrum approximation (MLSA) filter [13] as the synthesis filter.

In most conventional frame-based mapping approaches to VC, parallel speech data, in which the source and target speakers utter the same sentences, is required for obtaining the conversion parameters. In contrast, the proposed technique does not have such a restriction because the models of the source and target speakers can be trained separately. Although VC techniques with nonparallel training data have been already proposed (e.g., [6, 7]), it should be noted again that they do not take both the phonetic and prosodic contexts into account in the training and conversion process.

3. Speech database and experimental setup

For English speech data, we used the CMU ARCTIC database for speech synthesis where the speech were automatically labeled with phonetic and prosodic contexts using Festvox voice building tools [14] with CMU SphinxTrain [15]. The details of the annotation are found in [16]. From the database, three male and one female speakers, AWB, BDL, RMS, and SLT, were chosen as the source speakers, and one female speaker, CLB, was chosen as the target speaker. We used the first 50 sentences as the test data, and the next 200 sentences as the training data. For Japanese speech data, we used the ATR Japanese speech database set B which includes ten professional narrators’ speech data. Each speaker uttered 503 phonetically balanced Japanese sentences. 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 test sentences were 53 sentences (subset J), and the training data was chosen from the rest of 450 sentences (subset A to D).

Speech signals were sampled at a rate of 16kHz, and the STRAIGHT analysis [17] 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 a 52-dimensional feature vector by adding the log F0 and its delta to the feature vector used in the decoding part. For the F0 extraction, we used the ESPS get_f0 [18] tool for English speech, and an instantaneous-frequency-based technique [19] for Japanese speech.

For the acoustic model, we used hidden semi-Markov model (HSSM) [20] that has explicit duration distributions. In the decoding part, we simply converted HSSM to HMM in advance by calculating the transition probabilities from the mean parameters of duration distributions of HSSM. We used 5-state left-to-right triphone model for the decoding part, and 5-state left-to-right model with triphone and quantized F0 context for the synthesis part. As the quantized F0 context, we used preceding, current, and succeeding F0 symbols. The output distribution in each state of the MSD-HSMM was modeled by a single Gaussian density function, and the covariance matrix was assumed to be diagonal. In the context clustering for parameter tying, decision trees were automatically constructed based on a minimum description length (MDL) criterion [21]. We conducted the experiments under condition where the phonetic transcription was given as same in [7]. To mitigate the effect of over-smoothing in the speech synthesis, we used a parameter generation algorithm considering GV [22] in the subjective evaluation test.

4. Experimental results for English speech

4.1. Comparison of F0 quantization methods

In our technique, we must specify the number of quantization level to create the quantized F0 contexts for training and synthesis. In our previous study [8], we have already explored the appropriate number of quantization levels for Japanese speech using global F0 quantization. Here we again evaluated the conversion performance for English speech using both global and adaptive F0 quantization. As the objective measures of spectral and F0 distortion, we used mel-cepstral distance and root mean square (RMS) error of F0 between original and converted speech, which are commonly used in the objective evaluation of VC performance [4, 23].

We evaluated four conversion combinations of source and target speakers, that is, female speaker CLB as the target speaker, and speakers AWB, BDL, RMS, and SLT as the source speakers. Figure 3 shows the average F0 distortion of the four combinations with different number of quantization levels: 1, 2, 4, 8, and 16. In the figure, the values of standard deviation are also shown as error bars. GLOBAL and ADAPTIVE indicate our techniques using global and adaptive F0 quantization, respectively. For comparison, we also evaluated the widely-used simple F0 conversion with a single linear transformation (LINEAR) given by

\[ \tilde{f}_y = \frac{f_x - \mu_x}{\sigma_x} \cdot \sigma_y + \mu_y, \]  

where \( f_x \) and \( f_y \) are the log F0 values of the source speaker and converted speech, \( \mu_x \) and \( \sigma_x \) are the mean and standard deviation calculated from the source speaker's training data, and \( \mu_y \) and \( \sigma_y \) are those from the target speaker's one, respectively.

From the results, we found that the F0 distortion decreased when the number of quantization levels increased in our techniques. The conversion with both global and adaptive quantization methods outperformed the conventional simple linear conversion when the number of quantization levels was two or more. It was also found that the adaptive quantization gave better performance than that of the global one in average and variance of distortion among the source speakers. This indicates that the proposed technique is more robust for the source speaker's variation in the F0 conversion than the conventional one.

Next, we evaluated the spectral conversion performance. Table 1 shows the average spectral distortion of the four combinations with different number of quantization levels: 1, 2, 4, 8, and 16. In the table, the values of standard deviation are also listed in the parentheses. From the results, we can clearly see the difference between the results of F0 and spectral distortion. The F0 distortion decreased as the number of quantization levels increased, whereas there was little change in the spectral distortion for different number of quantization levels. Although the quantized F0 context was not effective for the spectral feature, the effect was not the negative one, and we have decided to use the same number of quantization levels as F0 in the spectral modeling. In consideration of these results, we used the adaptive F0 quantization and fixed the number of quantization levels to eight in the following experiments for English speech.

4.2. Comparison with GMM-based spectral conversion

Next, we compared the spectral conversion performance of the proposed technique using adaptive F0 quantization with the conventional GMM-based one. For the GMM-based conversion, we used the technique based on a maximum likelihood (ML) criterion with dynamic features [4]. We changed the number of mixtures from 32 to 2048. In our technique, we control the total number of HMM states by changing the size of the decision tree constructed in the context clustering for parameter tying. Specifically, we changed the weighting factor \( \omega \) of the penalty term of MDL \( D(\lambda) \) given by

\[ D(\lambda) = -L(\lambda) + w \cdot \frac{1}{2} K \log N, \]
where $L(\lambda)$ is log-likelihood of model $\lambda$, and $K$ and $N$ are the numbers of model parameters and observations, respectively. When $w$ is smaller than unity, the size of the decision tree becomes larger than that in the MDL criterion. When $w$ is zero, no parameters are tied. Results are shown in Figs. 4 (a) and (b). Figures show the average spectral distortion of the four conversion combinations. The standard deviations are also shown as error bars. In the GMM-based technique, there are large standard deviations for every mixtures, and this indicates that the performance of the GMM-based conversion highly depends on the individuality of the source speaker. By contrast, our technique gave significantly smaller deviations, and was found to be robust for the source speaker’s variation. Moreover, when we chose the optimal number of mixtures and HMM states, our technique outperformed the GMM-based one.

To examine the influence of the source speaker’s individuality in more detail, we show the spectral distortion in the conversions from respective speakers given the optimum model size in Fig. 5. In the HMM-based conversion, we used adaptive F0 quantization. The average value of optimum numbers of HSMM states of four target speakers was 2522. We also show the results when using the MDL criterion, i.e., $w = 1.0$ in Fig. 4 (b), in the clustering. When the MDL criterion was used, the average number of HSMM states of four speakers was 534. From the figure, we can see that the GMM-based technique with optimum model size gave better performance than that of the HMM-based one only in the female speaker SLT. This fact means that the GMM-based conversion works well when the relationship of source and target speakers’ spectral features can be well modeled by the GMM. However, except for the speaker SLT, the HMM-based technique was comparable to or outperformed the GMM-based one when the optimum model size was used in each technique. Moreover, the GMM-based conversion caused significantly larger variation of spectral distortions among the source speakers than the HMM-based one, and this would lead to an obstacle of VC from arbitrary speakers.

4.3. Results of subjective evaluation

Next, we evaluated the performance of the proposed technique with adaptive F0 quantization (ADAPTIVE) by XAB tests for English speech. We compared the proposed technique to our conventional technique with global F0 quantization (GLOBAL) and GMM-based technique (GMM). In the GMM-based technique, we converted speaking rate by warping frames of spectral and F0 features. For the warping, we applied linear transformation using the ratio of total utterance length of training data of source and target speakers. To correctly obtain the utterance length excluding silence, we used the time information from labels for the HMM training. It is noted that we used the MDL criterion for the context clustering of the HMM-based techniques and did not optimize the model size, whereas we used optimum number of mixtures, 512, in the GMM-based conversion.
7. Japanese participants listened to pairs of synthetic speech samples of ADAPTIVE and GLOBAL or GMM, and chose the sample whose speaker individuality was more similar to the reference sample X. The reference sample was vocoded speech of the target speaker. For each participant, we randomly chose eight sentences as the test samples from 50 sentences. The results are shown in Figs. 6 (a) and (b) with confidence intervals of 95%. When the quantization methods are compared, the proposed adaptive F0 quantization gave better performance than the conventional global one. From the results of Fig. 6 (b), we found that the HMM-based VC outperformed the GMM-based one except in the case where the source speaker was SLT, and this is consistent with the objective evaluation results in Sect. 4.2.

5. Experimental results for Japanese speech

5.1. Conversion performance with different amount of training data

So far, we fixed the number of training sentences to 200 for English speech. Here, we assessed the conversion performance of the proposed technique with adaptive F0 quantization when changing the amount of training data for Japanese speech. For comparison, we also evaluated the GMM-based technique with linear F0 transformation and our technique with global F0 quantization. In the HMM-based conversion, we set the number of quantization levels to eight from the experimental results for English speech and our previous study for Japanese speech [8].

Figures 7 (a) and (b) show the spectral distortion by mel-cepstral distance when changing the number of training sentences to 100, 200, 300, and 450. In the GMM-based conversion, we optimized the number of mixtures for each amount of training data. The optimum number of mixtures are shown in the parentheses of the figures. On the other hand, we did not optimized the number of HMM states in the HMM-based conversion, and used the MDL criterion to automatically determine the model size. From the results, we found that the performance of the proposed technique with adaptive F0 quantization was almost comparable to that of the our conventional HMM-based technique with global F0 quantization even when the amount of training data varies.

Figures 8 (a) and (b) show the F0 distortion by RMS error of F0 when changing the number of training sentences to 100, 200, 300, and 450. We can see that the HMM-based conversion with adaptive F0 quantization gave the better performance in all of the training data sets.

5.2. Results of subjective evaluation

We evaluated the performance of the proposed technique with adaptive F0 quantization by subjective evaluation tests for Japanese speech. In the HMM-based VC, we used the MDL criterion for the context clustering as in the case of English speech of Sect 4.3. First, we conducted opinion tests on speaker similarity of converted speech. In the HMM-based conversion, we set the number of mixtures to 100, 200, 300, and 450. We can see that the HMM-based conversion with adaptive F0 quantization gave the better performance in all of the training data sets.
**8. References**


---

**7. Acknowledgements**

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