HMM-Based Voice Conversion Using Quantized F0 Context

SUMMARY We propose a segment-based voice conversion technique using hidden Markov model (HMM)-based speech synthesis with nonparallel training data. In the proposed technique, the phoneme information with durations and a quantized F0 contour are extracted from the input speech of a source speaker, and are transmitted to a synthesis part. In the synthesis part, the quantized F0 symbols are used as prosodic context. A phonetically and prosodically context-dependent label sequence is generated from the transmitted phoneme and the F0 symbols. Then, converted speech is generated from the label sequence with durations using the target speaker’s pre-trained context-dependent HMMs. In the model training, the models of the source and target speakers can be trained separately, hence there is no need to prepare parallel speech data of the source and target speakers. Objective and subjective experimental results show that the segment-based voice conversion with phonetic and prosodic contexts works effectively even if the parallel speech data is not available.

key words: voice conversion, HMM-based speech synthesis, F0 quantization, prosodic context, nonparallel data

1. Introduction

Recent developments in statistical parametric speech processing have provided us many useful and beneficial applications in speech recognition and speech synthesis. Voice conversion is one of such attractive applications which can change nonlinguistic or paralinguistic information, e.g., speaker individuality or emotional expressions appearing in speech. The demands for voice conversion applications are increasing in many fields, such as in entertainment [1], foreign language education [2], and software for the physically challenged [3].

In this context, a variety of techniques have been proposed [4]. The techniques that have been widely studied so far are based on statistical mapping of spectral features at the frame level using a probabilistic model, i.e., Gaussian mixture model (GMM) [5], [6]. Although a source speaker’s spectral features can be easily converted so as to be closer to those of a target speaker by using the GMM-based framework, there are some problems, such as the requirement of parallel data, over-smoothing effect, and insufficient prosody conversion. Recently, several approaches have been proposed to overcome these problems. In [7], spectral mapping with nonparallel training data was proposed by introducing hidden Markov model (HMM)-based modeling and adaptation with phonetic information. The over-smoothing effect is alleviated by introducing global variance (GV) parameters into the estimation of a parameter trajectory [8]. For the prosody conversion, nonlinear modification of fundamental frequency (F0) has been proposed based on multi-space distribution GMM (MSD-GMM) [9].

However, in the above techniques, it is not easy to appropriately convert segmental or supra-segmental speaker characteristics because no phonetic or prosodic context is taken into account in the model training and conversion process. As a result, the conversion performance is not always satisfactory, and highly depends on the combination of the source and target speakers. To alleviate this problem, a segment-based voice conversion using unit selection is one of the effective approaches where the dynamic characteristics of speaker individuality is converted as well as the static ones [10]. In the technique, the phone units are used as the segments, and a mapping table is generated between triphones of the source and target speakers. Although this unit-selection-based approach significantly improved the conventional VQ-based frame-by-frame mapping one [11], it was also pointed out that a large amount of speech data of the source and target speakers must be prepared to achieve high conversion performance [10].

In this paper, we propose a novel voice conversion technique for converting segmental and supra-segmental features by introducing an HMM-based speech synthesis with phonetic and prosodic contexts [12]. The basic idea of our technique comes from the HMM-based phonetic vocoder [13], which was proposed for very low bit-rate speech coding. In this phonetic vocoder, a phoneme sequence with durations is extracted from spectral features of the input speech using a phoneme decoder, and then a spectral feature trajectory is generated using the HMM-based speech synthesis framework. By replacing the acoustic model for the synthesis with that of another speaker, we can convert the spectral feature at the segmental level. In the proposed technique, we first extract phonetic and prosodic information from the input source speaker’s speech, and then the converted speech is generated from the target speaker’s model with the extracted information using the HMM-based speech synthesis framework. To model not only the spectral features but also the prosodic features of the synthetic speech units, we simultaneously model the spectral, F0, and duration features using multi-space distri-
bution HMM (MSD-HMM) [14]. For modeling the F0 features using the MSD-HMM, we propose an automatic context generation method. This method is based on the standardization and quantization of F0 values at the phoneme level, and no hand-labeled speech data with prosodic information is required.

In the proposed technique, there are two major contributions associated with the voice conversion performance and flexibility. The first is the improvement of the spectrum and F0 conversion performance by using a phonetically and prosodically context-dependent acoustic model. In the spectral conversion, the source speaker’s individuality is almost lost when input speech is coded into a phoneme sequence. As a result, we can alleviate the degradation of conversion performance caused by the variability of the source speaker’s individuality. In the F0 conversion, the conventional technique based on the global linear transformation cannot convert the dynamic characteristics appearing within a phoneme or between phonemes because there is no explicit modeling of prosodic features at the segmental level. In our technique, we convert such phoneme-level characteristics by introducing context-dependent modeling with quantized F0 contexts and the MSD-HMM. The second contribution is that our technique does not require any parallel speech data of the source and target speakers. This will be important for making the voice conversion more flexible because it is not realistic to prepare the parallel speech data in some applications, e.g., voice conversion for spontaneous or cross-language speech.

2. Voice Conversion Using Context-Dependent HMM with Prosodic Context

2.1 Automatic Generation of Prosodic Context Using Quantized F0 Symbols

When speech features including both spectral and F0 features are modeled using the MSD-HMM, the prosodic context labels are required for the training data as well as the phonetic contexts for modeling the F0 features appropriately. Although a standard way of prosodic labeling is to incorporate the accent information, it is not always possible to automatically give “full context” labels including accent information with high accuracy. To overcome this problem, we employ the F0 quantization for the prosodic labeling, and uses coarsely quantized F0 symbols as the prosodic context.

We assume that the log F0 values of source and target speakers’ speech follow normal distributions. In the quantization, we first standardize the log F0 distribution into $N(0, 1)$. The global means and variances of log F0 are calculated in advance for the source and target speakers’ training data, and are used for the standardization. By using this standardization before quantization, the quantized F0 symbols of the source and target speakers are expected to roughly correspond to each other, and we can use the same label between the source and target speakers without modification when synthesizing the converted speech.

In our technique, each phone unit of speech is labeled with quantized F0 symbols. Therefore, in the model training, we conduct phoneme segmentation in advance to obtain the phone boundary information, and calculate the mean value $f_p$ for the standardized log F0 for each phone unit $p$. Then, the F0 symbol $s_p$ is obtained by quantizing $f_p$ into a discrete value as follows:

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

where $Q[\cdot]$ denotes an operation of scalar quantization, and $M$ is the number of the quantization levels. We set the points, which equally divide the region $[-2, 2]$ into $M$ levels, as the quantization boundaries. An example of eight-level F0 quantization is shown in Fig. 1. In the voice conversion, the information of phone boundary is obtained using forced alignment between the extracted phoneme sequence and the spectral features of the input speech, and the above quantization process is applied to the log F0 sequence to create the context-dependent labels for speech synthesis.

2.2 Overview of Proposed Technique

A block diagram of the proposed technique is shown in Fig. 2. In the decoding part, sequences of phoneme, melcepstrum, 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 melcepstrum sequences. 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, durations, and F0 symbols obtained in the decoding part. We adopt a context-dependent label 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 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 [15] as the synthesis filter.

In most conventional frame-based mapping approaches to voice conversion, 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 voice conversion techniques with nonparallel training data have been already proposed (e.g., [7], [16]), it should be noted that they do not take the phonetic and prosodic contexts into account in the training and conversion process.

3. Experiments

3.1 Experimental Conditions

We used speech data of four speakers contained in the ATR Japanese speech database set B. These speakers were two males (MHT and MYI) and two females (FKN and FYM). Each speaker uttered 503 phonetically balanced sentences consisting of ten subsets — A, B, ···, and J. The subset J contains 53 sentences, and each of the others contains 50 sentences. Speech signals were sampled at a rate of 16 kHz, 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 52 dimensional feature vector by adding the log F0 and its delta to the feature vector used in the decoding part. We used F0 data contained in the database.

For the acoustic models, we used 5-state left-to-right triphone HMMs for the decoding part, and 5-state left-to-right MSD-HMMs with triphone and quantized F0 contexts 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 HMM/MSD-HMM was modeled by a single Gaussian density function, and the covariance matrices of these models 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 [18]. To mitigate the effect of over-smoothing in the speech synthesis, we used a parameter generation algorithm considering GV [19] in the subjective evaluation tests. We chose 53 sentences from the subset J as the test data.

In the following objective and subjective evaluation tests, voice conversion was conducted for all twelve combinations among the four speakers. We also conducted the experiments under conditions where the phoneme sequence for the input speech of the source speaker was known except for the subjective evaluation in Sect. 3.5.

3.2 Effect of Prosodic Context Using F0 Quantization

As described in Sect. 2.1, the proposed technique uses coarsely quantized F0 symbols for the prosodic context. The trained model with single-level quantization is equal to that without F0 context. From the preliminary experimental results, we found that the two-level F0 quantization was insufficient to transmit the intonation and accent information. Here, we experimentally explore the sufficient number of quantization levels using objective measure. Specifically, we evaluated the spectral and F0 conversion performance of the proposed technique when changing the number of the quantization levels from 1 to 16. The training data of the source and target speakers consisted of 450 parallel sentences from subsets A to I. In this study, we used mel-cepstral distance for the objective measure of the spectral similarity as used in [8], [9]. To calculate the distortion frame by frame, time-alignment was conducted in advance using dynamic time warping (DTW) with the spectral features. Figure 3 shows the average mel-cepstral distances with different number of quantization levels, 1, 2, 4, 8, and 16, for male to male, male to female, female to male, and female to female conversion of the source and target speakers. From the figure, it is found that the spectral distortion slightly decreased compared to the case when using only the triphone context. However, the distortion stays almost the same value when the number of quantization levels is larger than four.

Next, we objectively evaluated the similarity of the F0 contours between the target speaker’s original and converted speech. As the objective measure, we used the root mean square (RMS) error of log F0 as used in [9], [11]. The RMS error was obtained using only the frames where both the source and target speaker’s F0 values exist. We also calculated the F0 distortion when using the global linear transfor-
Fig. 3 Average mel-cepstral distances between input and converted speech samples with different number of quantization levels.

Fig. 4 Average RMS errors of F0 between input and converted speech samples with different number of quantization levels.

\begin{equation}
\hat{f}_y = \frac{f_x - \mu_x}{\sigma_x} \cdot \sigma_y + \mu_y,
\end{equation}

where $f_x$ and $\hat{f}_y$ 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.

Figure 4 shows the average RMS errors of log F0 with different number of quantization levels, 1, 2, 4, 8, and 16, for male to male, male to female, female to male, and female to female conversion of the source and target speakers. When the number of quantization levels is one, i.e., there is no prosodic context, the F0 distortion of the proposed technique is much larger than that of the linear transformation. However, the distortion decreases with increasing the number of quantization levels, and the proposed technique outperforms the linear transformation when the number of quantization levels is four or more. The F0 distortion does not change significantly between the 8- and 16-level quantization. Taking these results into account, we fixed the number of quantization levels to eight in the following experiments. In Fig. 5, we show an example of F0 contours of target speaker’s original and converted speech.

3.3 Performance Evaluation of Voice Conversion with Nonparallel Training Data

In the proposed technique, the acoustic models of source and target speakers can be trained separately, and thus parallel training data is not indispensable for the model training. In this experiment, we compared the conversion performance of the proposed technique when using parallel and nonparallel training data through objective evaluation tests. We used 200 parallel and nonparallel sentences chosen from subsets A to I for the model training of the source and target speakers. Because the conversion performance varies depending on the choice of training data sets of the speakers, we prepared nine different combinations of training data sets by choosing the subsets as listed in Table 1. It should be noted

Table 1 Training data sets for performance comparison using parallel and nonparallel training data.

<table>
<thead>
<tr>
<th>Training set</th>
<th>Parallel</th>
<th>Source speaker</th>
<th>Target speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td>A,B,C,D</td>
<td>F,G,H,I</td>
</tr>
<tr>
<td>b</td>
<td></td>
<td>B,C,D,E</td>
<td>G,H,I,A</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>C,D,E,F</td>
<td>H,I,A,B</td>
</tr>
<tr>
<td>d</td>
<td></td>
<td>D,E,F,G</td>
<td>I,A,B,C</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td>E,F,G,H</td>
<td>A,B,C,D</td>
</tr>
<tr>
<td>f</td>
<td></td>
<td>F,G,H,I</td>
<td>B,C,D,E</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>G,H,I,A</td>
<td>C,D,E,F</td>
</tr>
<tr>
<td>h</td>
<td></td>
<td>H,I,A,B</td>
<td>D,E,F,G</td>
</tr>
<tr>
<td>i</td>
<td></td>
<td>I,A,B,C</td>
<td>E,F,G,H</td>
</tr>
</tbody>
</table>

(b) single- and eight-level F0 quantization

Fig. 5 Example of F0 contours of target speaker’s original and converted speech.
that the same training data sets of the target speaker were used for both parallel and nonparallel data. This means that the target speaker’s acoustic models were identical in both cases.

The average mel-cepstral distances and F0 RMS errors for all twelve conversion combinations of source and target speakers when using parallel and nonparallel data are shown in Figs. 6 and 7, respectively. From the results of the mel-cepstral distortion, we found that the difference in spectral distortions between parallel and nonparallel data is small in every training data set. This is because the target speaker’s models were identical for parallel and nonparallel cases in this experiment, and only the phoneme alignment information and prosodic context determined by the source speaker’s model affected the conversion performance.

In contrast, there were differences in F0 distortions between parallel and nonparallel data depending on the training sets. However, the overall performance was almost comparable in parallel and nonparallel data. In addition, the distortion variation among the nine different training sets with the nonparallel data was smaller than that with the parallel one. A possible reason is that the dependency of the training data was alleviated by using different sentences for the model training of the source and target speakers. The above results indicate that the proposed technique is still effective when the parallel training data is not available.

3.4 Performance Comparison between GMM-Based and Proposed Techniques

We compared the conversion performance with the GMM-based mapping technique [8]. In the GMM-based technique, the dynamic feature is taken into account, and a smooth speech parameter trajectory is generated using maximum likelihood (ML) criterion as well as the proposed technique.

In this experiment, we used parallel training data of the source and target speakers. The training data set was 450 sentences from subsets A to I. In the GMM-based technique, we first aligned the frames of the source and target speaker’s training data of the mel-cepstrum using DTW, and generated joint feature vectors. Then the GMM was trained using the joint vectors. In the model training of the GMM, we excluded the frames corresponding to silence according to the label information used for the model training of the proposed technique. In addition, we assumed that the covariance and cross-covariance parameters of the source and target speakers were diagonal. In the parameter conversion process, we converted the spectral features, excluding the silence frames, and used the source speaker’s parameters without modification in the silence frames. Then, the converted mel-cepstrum was modified so that the power did not change before and after the conversion.

3.4.1 Objective Evaluation Tests

Figure 8 shows the average mel-cepstral distances between the target speaker’s original and converted speech of all twelve combinations of speakers. The mel-cepstral distance was calculated using the frames, excluding the silence ones according to the label information for the test data. We changed the training sentences from 100 to 450 as listed in Table 2. In the figure, the numbers in the parentheses are
the average values of the total numbers of distributions in the HMMs of the target speakers and the GMMs.

In the GMM-based technique, we calculated the average mel-cepstral distances in advance by changing the number of mixtures from 32 to 1024, and chose the optimal one for each training set. In contrast, the number of HMM states of the proposed technique was automatically determined in the tree-based context clustering using the MDL criterion. In the objective evaluation, we did not consider GV when generating the parameter trajectory in both the GMM-based and our techniques. From the results, we found that the spectral distortion decreased with the proposed technique, which indicates that the segment-based conversion with phonetic and prosodic contexts works effectively. Moreover, as listed in Table 3, the standard deviations of the mel-cepstral distances of all twelve conversion combinations are much smaller with the proposed technique than with the GMM-based one. This implies that our technique is robust for the variations of source and target speaker combinations compared to the conventional GMM-based one.

Figure 9 shows the average F0 distortions of all twelve combinations of speakers. The numbers in the parentheses are the average values of the total numbers of HMM states for the target speakers. In this figure, we also show the results for the global linear transformation discussed in Sect. 3.2 given by Eq. (2). We can see that the distortion was mitigated more compared to the global linear transformation.

3.4.2 Subjective Evaluation Tests

We also subjectively compared the proposed and GMM-based techniques. First, we evaluated the naturalness of the converted speech samples using a mean opinion score (MOS) test. The training data and other conditions were the same as those for the objective evaluation tests. It should be noted that we used the parameter generation algorithm considering GV for both the GMM-based and proposed techniques. For each participant, five test samples were chosen randomly from 53 test sentences. Fifteen participants listened to the test samples of all twelve combinations of the source and target speakers, and rated the naturalness of the test samples using a 5-point scale: “5” for excellent, “4” for good, “3” for fair, “2” for poor, and “1” for bad. Figure 10 shows the average of the MOS of 1) four combinations of same gender conversion, 2) eight combinations of cross gender conversion, and 3) all twelve combinations of the source and target speakers. Confidence intervals of 95% are also shown in this figure. From the figure, we can see that the proposed technique significantly improved the naturalness of converted speech compared to the GMM-based one.

Next, we evaluated the speaker individuality of the converted speech samples using an XAB test. The training data and other conditions were the same as those for the MOS test. The reference sample $X$ is the target speaker’s vocoded speech, and $A$ and $B$ are the converted speech samples of the GMM-based and proposed techniques, respectively. For each participant, five test samples were chosen randomly from 53 test sentences. Fifteen participants listened to the test samples of $A$ and $B$ in random order, and asked which sample sounded more similar to the individuality of the reference sample $X$. Table 4 shows the average scores of the XAB test. We confirmed that the difference between the results of the GMM-based and our techniques is statistically

<p>| Table 2  |
| Training data sets. |</p>
<table>
<thead>
<tr>
<th>Number of sentences</th>
<th>Subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>A, B</td>
</tr>
<tr>
<td>200</td>
<td>A, B, C, D</td>
</tr>
<tr>
<td>300</td>
<td>A, B, C, D, E, F</td>
</tr>
<tr>
<td>450</td>
<td>A, B, C, D, E, F, G, H, I</td>
</tr>
</tbody>
</table>

<p>| Table 3  |
| Standard deviations [dB] of mel-cepstral distances of all twelve conversion combinations. |</p>
<table>
<thead>
<tr>
<th>Number of sentences</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>450</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-based</td>
<td>0.20</td>
<td>0.19</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.12</td>
<td>0.09</td>
<td>0.11</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Fig. 9  Comparison of RMS errors of F0 between linear transformation and proposed technique.

Fig. 10  Results of MOS test on speech naturalness for GMM-based and proposed techniques.
Table 4  Result of XAB test on speaker individuality of GMM-based and proposed techniques.

<table>
<thead>
<tr>
<th></th>
<th>GMM-based</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43%</td>
<td>57%</td>
</tr>
</tbody>
</table>

Table 5  Phoneme accuracy for input speech samples of source speakers.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>MHT</th>
<th>MYI</th>
<th>FKN</th>
<th>FYM</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>91.7</td>
<td>86.7</td>
<td>86.6</td>
<td>90.0</td>
<td>88.7</td>
</tr>
</tbody>
</table>

significant at the 1% level.

3.5 Performance Evaluation Using Recognized Phoneme Sequence

In the previous experiments, we assumed that the phoneme sequence was known for the input speech. Although this condition is acceptable in some applications such as post-processing of text-to-speech systems and narrative generation tools, there are cases where we cannot use phonetic transcriptions. Hence, in this experiment, we evaluated how the conversion performance is affected when the phonetic information is extracted automatically using a standard phoneme recognition method. In addition, we also evaluated our technique with and without phoneme duration transmission.

We trained the source and target speakers’ models using 450 parallel sentences. For the phoneme recognition, we used five-state left-to-right triphone HMMs with diagonal covariance matrices. The phoneme recognition was performed based on the Viterbi algorithm, and phonetic networks generated from Japanese phonetic concatenation rules were used in the recognition. The average phoneme error rates for the 53 test sentences of four source speakers are listed in Table 5. The phoneme accuracy was calculated by

\[
\text{Accuracy} \, (\%) = \frac{N - D - S - I}{N} \times 100 \quad (3)
\]

where \(N\), \(D\), \(S\), and \(I\) represent the numbers of labels in the reference transcriptions, deletions, substitutions, and insertions, respectively. We evaluated the conversion performance on speaker individuality using a comparison category rating (CCR) test. The scale for the CCR test was “5” for very similar and “1” for very dissimilar to reference speech. We used the target speaker’s vocoded speech of the test sentences as reference samples. For each participant, five test samples were chosen randomly from 53 test sentences. Each participant listened to the test samples, and rated the speaker individuality of the test samples compared to the reference samples.

Figure 11 shows the average scores of the CCR test for 1) four combinations of same gender conversion, 2) eight combinations of cross gender conversion, and 3) all twelve combinations of the source and target speakers. Confidence intervals of 95% are also shown in the figure. From these results, we found that the performance degrades when using the automatically recognized label sequence without duration transmission. This is because incorrect phonemes were generated with the target speaker’s duration parameters, and the recognition errors were clearly perceived by the participants. In contrast, when the duration information of the input speech was transmitted, the conversion performance was substantially improved, and became closer to that with the correct phonemes. When the correct phoneme sequence was given for the input speech, the performance was slightly improved using the target speaker’s duration parameters.

4. Conclusion

We have proposed a novel segment-based voice conversion technique using HMM-based speech synthesis with phonetic and prosodic contexts. From the objective experimental results, we found that spectral and F0 distortions decreased using the proposed technique compared to those of the conventional GMM-based one. Moreover, we have shown that the proposed technique is robust for the variation of the source and target speakers’ pair in spectral conversion. We have also shown that the perceptual naturalness and similarity of the converted speech improved based on the subjective experimental results. In addition, the objective experimental results have shown that our technique works even if the parallel speech data of the source and target speakers are not available. This would make the implementation of the voice conversion applications using spontaneous or cross-language speech data much easier. In the future work, we will focus on reducing the training data by using speaker-independent and speaker-adaptive approaches.

Acknowledgment

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063.
References


