HMM-based singing voice synthesis system using pitch-shifted pseudo training data

Ayami Mase, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda

Department of Computer Science, Nagoya Institute of Technology, Japan
{ayami-m, uratec, nankaku}@sp.nitech.ac.jp, tokuda@nitech.ac.jp

Abstract

Over the last few years, a statistical parametric approach to singing voice synthesis based on hidden Markov models (HMMs) has been grown over. In this approach, spectrum, excitation, and duration of singing voices are simultaneously modeled by context-dependent HMMs, and waveforms are generated from HMMs themselves. However, pitches that rarely appear in training data cannot be properly generated because the system cannot model their fundamental frequency \(F_0\) contours. In this paper, we propose a technique for training HMMs using pitch-shifted pseudo data. Subjective listening test results show that the proposed technique improves the naturalness of the synthesized singing voices.

Index Terms: singing voice synthesis, HMM-based speech synthesis, pitch-shift

1. Introduction

In recent years, various applications of speech synthesis systems have been proposed and investigated. Singing voice synthesis, which enables computers to sing any song, is one of the hot topics especially in the fields of entertainment and amusement. Singing voice synthesis is popular especially in Japan because of a singing synthesizer named “Vocaloid” [1]. By using this system, songwriters need not employ a real singer or sing a song by themselves. This kind of systems are leading to model di

2. HMM-based Singing Voice Synthesis System

Figure 1 is an overview of the HMM-based singing voice synthesis system. The system consists of training and synthesis parts. Although it is quite similar to the HMM-based speech synthesis system [5], two specific techniques were introduced for singing voice synthesis: time-lag modeling [6] and vibrato modeling [3, 7].

The time-lag is an important factor for singing voice synthesis. In the case of singing voice synthesis, the rhythm and tempo of the music must not be ignored. Therefore, the start timings of the notes and phoneme durations in each note must be determined from the musical score data. However, if the musical score is strictly followed, the synthesized singing voice will be unnatural because of time-lags. To overcome this problem, the time-lags of individual notes are modeled by Gaussian distributions.

Vibrato is also an important singing technique which should be modeled, although it is not included in the musical score. The timing and power of vibrato vary from singer to singer. Therefore, vibrato modeling is required for naturalness of synthesized singing voice. To model vibrato automatically, we introduce a vibrato modeling technique for the HMM-based singing voice synthesis.

2.1. Training Part

In the training part, we first extract various parameters to be used as training data from the waveform of a song in the singing voice database. Training data are mel-Cepstral coefficients, log fundamental frequencies \(F_0\), and vibrato parameters. Their dynamics features and them are used as the feature vector for training and these feature vectors are modeled by HMMs.

In this paper, vibrato is assumed as periodic fluctuations of
Figure 1: Overview of HMM-based singing voice synthesis system.

log $F_0$. Two parameters, fluctuation amplitude by cent and frequency by Hz, are used as vibrato parameter. Vibrato sections and vibrato parameters in those sections are estimated from a log $F_0$ sequence [8]. Then since the observation sequence of log $F_0$s is composed of one-dimensional continuous values and a discrete symbol representing an “unvoiced,” it is modeled together with spectrum parameters at once by the HMM-based on multi-space probability distribution (MSD-HMM) [9]. Furthermore, in this system, HMM is extended to a hidden semi-Markov model (HSMM) [10] in order to model duration explicitly. HSMM-based phoneme alignments performed by using WFST [11] are used for extracting the start timing of each note in the natural singing voice.  

Although each HMM models one phoneme in singing voice, same phonemes have different characteristics in connection with pitch, length of note, the relation to the previous or the next phoneme, etc. These variation factors are called “context.” The HMM considering contexts is used to model in more detail. Context-dependent labels can automatically be covered properly since it has a great impact on synthesized singing voices. Therefore, we proposed a technique for training HMMs by using pitch-shifted pseudo data. Since pitch is represented by the log $F_0$ parameter, the pitch-shifted pseudo data clustering are summarized as follows.

1. Prepare questions that can be answered with a yes or a no for contexts.
2. Tie all states per state position in HMMs.
3. Select one question that maximizes the likelihood of model parameters with respect to the training data and split the cluster by using this question.
4. Conduct step 3 again for each divided cluster.

The minimum description length (MDL) criterion [12] is used to stop these steps. HMM states that have similar acoustic features are tied. Unknown models corresponding to contexts that rarely appear in the training data can be generated by traversing the decision trees. However, models which have unseen pitches can generate unnatural sounds because an unseen pitch is substituted with the most similar pitch.

2.2. Synthesis Part

In the synthesis part, first a given musical score including lyrics to be synthesized is converted into a context-dependent label sequence. After that, a song HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. The HMM state durations of the song are then determined with respect to the score including lyrics, the state duration models, and the time-lag models. Next, the speech parameter generation algorithm [4] generates spectral, excitation, and vibrato parameters. Finally, a singing voice waveform is directly synthesized from the generated spectral, excitation, and vibrato parameters by using the Mel Log Spectrum Approximation (MLSA) filter [13].

2.3. Contextual factors

In the HMM-based text-to-speech synthesis system, contextual factors that may affect reading speech such as phoneme identity, part-of-speech, accent, and stress, etc., have been taken into account. However, contextual factors that affect singing voices should be different from those used in text-to-speech synthesis. Therefore, contextual factors that were specific to singing voices [3] were considered in this paper. MusicXML [14] was used for representing musical scores that include lyrics. MusicXML can describe various information such as tie, slur, dynamics, and tempo. Context-dependent labels can automatically be determined from the musical score including lyrics.

3. Pitch-shifted Pseudo Training Data

The performance of HMM-based speech synthesis systems depends strongly on training data because these systems are “corpus-based”. Because of this, HMMs corresponding to contextual factors that rarely appear in training data cannot be well-trained. To solve this problem, algorithms for designing speech database considering the balance among contextual factors have been proposed [15].

HMM-based singing voice synthesis system should also have database including various contextual factors. However, data have to be sparse because singing voices have more contextual factors than those used in reading speech synthesis, e.g., pitch, tempo, key, beat, and dynamics. Pitch in particular should be covered properly since it has a great impact on synthesized singing voices. Therefore, we proposed a technique for training HMMs by using pitch-shifted pseudo data. Since pitch is represented by the log $F_0$ parameter, the pitch-shifted pseudo data

---

1 Vibrato is sometimes modeled by log $F_0$ and the power of the singing voice waveforms.
2 Although HSMM is used in this system, HSMM is hereinafter called HMM for simplicity.
can be easily prepared by shifting \( F_0 \) up or down in half tone. This technique makes it possible to increase the amount of \( F_0 \) training data without recording a larger singing voice database.

Figure 2 and 3 show the distributions of pitch in training data (10 and 60 songs). Before the pitch-shift is applied, the distributions are sparse. As the amount of data increases, the distribution has been smooth after pitch-shift. Mel-Cepstral coefficients and vibrato parameters were added by copying the same data because they were affected by the small amount of pitch-shifting. Since context-dependent labels for context-clustering contain pitch information, we made new labels for pseudo data that contained information about shifted pitch.

The amount of training data increased threefold as a result of adding pitch-shifted pseudo data. Therefore, the size of decision trees increases when we use the MDL-based stopping criterion. Thus, context-clustering should be stopped at an appropriate size of each decision tree.

In the HMM-based singing voice synthesis system, the MDL criterion is used for deciding when to stop splitting. The MDL criterion is an information criterion that is good for selecting an appropriate model from among various probabilistic models. The change in the total description lengths before and after splitting is calculated, and splitting is conducted when the difference exceeds a threshold. This node splitting is carried out until there is no node that exceeds the threshold. The decision trees will be large if the threshold is set low (small if the threshold is set high). When node \( S \) is divided up into two nodes, \( S_{<} \) and \( S_{>} \), by a question \( q \), the change in the total description lengths caused by this split is calculated as follows:

\[
\Delta_q = L(S) - \{L(S_{<}) + L(S_{>})\} + \alpha \frac{N}{2} \log \Gamma(S_0)
\]

where, \( S_0 \) denotes the root node, \( \alpha \) is the heuristic weight for the penalty term of the MDL criterion, \( N \) is the number of parameters increased by this split, and \( \Gamma(\cdot) \) is the posterior probability. The heuristic weight, \( \alpha \), is used to control the size of the decision trees.

4. Experiments

To evaluate the performance of the proposed HMM-based singing voice synthesis system using pitch-shifted pseudo training data, subjective experiments were conducted.

4.1. Experimental Conditions

We used a singing voice database including 70 Japanese children’s songs sung by one female singer. This database also includes the musical scores of these songs represented by MusicXML. The waveforms of the songs were sampled at a rate of 16 kHz and windowed by a 25 ms Blackman window with a 5-ms shift. Feature vectors consisted of spectral, \( F_0 \) and vibrato features. The spectrum parameter vector consisted of 39 STRAIGHT [16] mel-Cepstral coefficients including the zeroth coefficient and their delta and delta-delta coefficients. The excitation parameter vector consisted of log \( F_0 \) and its delta and delta-delta. The vibrato parameter vector consisted of fluctuation amplitude, frequency, their deltas, and delta-deltas.

To evaluate the effectiveness of pitch-shifted pseudo training data, we conducted a subjective listening test for the baseline method without using pitch-shift and four proposed methods using pitch-shift (\( \alpha = 3.0, 4.0, 5.0, 6.0 \)). Although the decision tree based context-clustering technique was separately applied to distributions for spectrum, excitation, and state duration, the same \( \alpha \) was used. Only \( \alpha \) for the decision tree of vibrato parameter was kept constant. Ten songs not included in the training data were divided up every four to eight bars. As a result, 29 musical phrases were obtained. The singing voices were then synthesized from the HMMs trained with and without pitch-shifted pseudo training data. 10 subjects were asked to rate the naturalness of the synthesized singing voices on Mean Opinion Score (MOS) with a scale from 1 (poor) to 5 (good). For each subject, 15 randomly selected musical phrases were presented. Experiments were carried out in a sound-proof room.

4.2. Experimental Results

Figure 4 shows the experimental results with 10 songs for training and figure 5 shows ones with 60 songs used for training. The 95% confidence intervals were calculated by the t-test. These results are not comparable because there experiments were conducted independently. It can be seen from the figures that the highest MOS is obtained by the proposed technique in the case of \( \alpha = 5.0 \). Naturalness was improved by the proposed technique by adjusting the size of decision tree properly. As a result, the technique for training HMMs by using pitch-shifted pseudo data is effective under the appropriate context-clustering.

Table 1 and 2 list the number of leaf nodes. Although training data increased threefold, \( \alpha \) of the highest quality synthesized voices is not 3.0. Therefore, the best weight for the MDL criterion remains a matter of research.
5. Conclusions

In the present paper, an HMM-based singing voice synthesis system using pitch-shifted pseudo training data was proposed. To overcome the problem whereby pitches that hardly appear in training data cannot be properly generated, pitch-shifted pseudo training data was added. The naturalness of the synthesized singing voices was improved because the coverage of pitch training data was increased. Subjective listening test results showed that models for pitch can be trained effectively without a large amount of data in the proposed technique. Furthermore, the quality of synthesized voices exceeded that of the baseline when the sizes of the decision trees were properly adjusted. Future work will focus on evaluating the amount of pitch-shift.

6. Acknowledgements

The authors thank Dr. Shinni Sako for the construction of the database. The research leading to these results was partly funded from the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communications, Japan.

7. References


[16] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructur-