PITCH ADAPTIVE TRAINING FOR HMM-BASED SINGING VOICE SYNTHESIS

Keiichiro Oura, Ayami Mase, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Computer Science, Nagoya Institute of Technology, Japan

ABSTRACT

A statistical parametric approach to singing voice synthesis based on hidden Markov Models (HMMs) has been growing in popularity over the last few years. The spectrum, excitation, vibrato, and duration of singing voices in this approach are simultaneously modeled with context-dependent HMMs and waveforms are generated from the HMMs themselves. HMM-based singing voice synthesis systems are heavily based on the training data in performance because these systems are "corpus-based." Therefore, HMMs corresponding to contextual factors that hardly ever appear in the training data cannot be well-trained. Pitch should especially be correctly covered since generated F_0 trajectories have a great impact on the subjective quality of synthesized singing voices. We applied the method of "speaker adaptive training" (SAT) to "pitch adaptive training," which is discussed in this paper. This technique made it possible to normalize pitch based on musical notes in the training process. The experimental results demonstrated that the proposed technique could alleviate the data sparseness problem.

***Index Terms*—** singing voice synthesis, hidden Markov model, pitch adaptive training

1. INTRODUCTION

A statistical parametric approach to speech synthesis based on hidden Markov models (HMMs) has been growing in popularity over the last few years [1]. Context-dependent HMMs are estimated from speech databases in this approach and speech waveforms are generated from the HMMs themselves. This framework makes it possible to model different voice characteristics, speaking styles, or emotions without having to record large speech databases. For example, adaptation, interpolation, and eigenvoice techniques have been applied to this system, which has demonstrated that voice characteristics could be modified. A singing voice synthesis system has also been proposed by applying the HMM-based approach [2].

The quality of synthesized singing voices strongly depends on training data because HMM-based singing voice synthesis systems are "corpus-based." Therefore, HMMs corresponding to contextual factors that rarely appear in training data cannot be well-trained. Although databases including various contextual factors should be used in the HMM-based singing voice synthesis systems, it is almost impossible to cover all possible contextual factors since singing voices involve a huge number of contextual factors, e.g., keys, lyrics, dynamics, note positions, durations, pitch, etc. Pitch should be properly covered particularly, since it has a great impact on the subjective quality of synthesized singing voices[1]. A technique using pitch-shifted pseudo-data [3] is one solution to this problem. However, there are various other problems such as large computational costs. Although data-level pitch normalization [4] has been proposed, there are still some other problems such as the inconsistency between data and training. This paper proposes pitch adaptive training for HMM-based singing voice synthesis to overcome these problems. The differences between log F_0 sequences extracted from waveforms and the pitch of musical notes can be modeled in the proposed training.

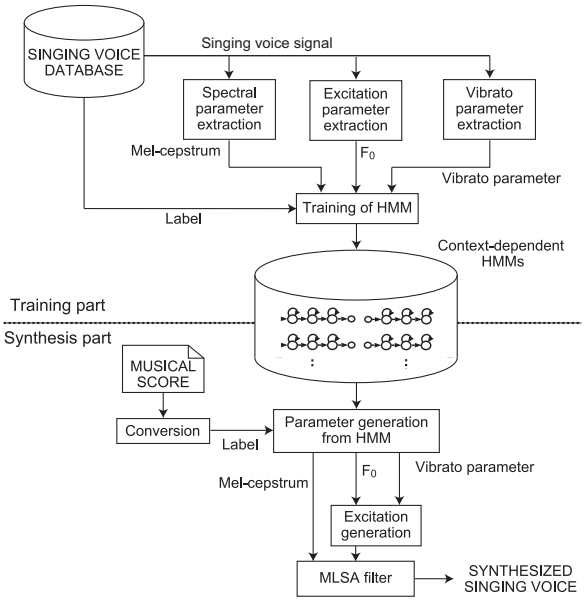

Fig. 1. *Overview of HMM-based singing voice synthesis system.*

The rest of this paper is organized as follows. Section 2 overviews the HMM-based singing voice synthesis system. Details on pitch adaptive training for HMM-based singing voice synthesis are described in Section 3. Section 4 discusses subjective experiments and Conclusions are drawn in Section 5.

2. HMM-BASED SINGING VOICE SYNTHESIS SYSTEM

The HMM-based singing voice synthesis system is quite similar to the HMM-based text-to-speech synthesis system. However, there are distinct differences between them. Fig. 1 gives an overview of the HMM-based singing voice synthesis system [2]. It consists of training and synthesis parts. The spectrum (e.g., mel-cepstral coefficients), excitation, and vi-

[1] F_0 modeling for HMM-based speech synthesis has also a great impact. Many techniques has peen proposed [5, 6].

brato are extracted from a singing voice database in the training part and they are then modeled with context-dependent HMMs. Context-dependent models of state durations are also estimated. An arbitrarily given musical score including the lyrics to be synthesized is first converted to a context-dependent label sequence in the synthesis part. Second, according to the label sequence, an HMM corresponding to the song is constructed by concatenating the context-dependent HMMs. Third, the state durations of the song HMM are determined with respect to the state duration models. Fourth, the spectrum, excitation, and vibrato parameters are generated by an algorithm to generate the speech parameters. Finally, a singing voice is synthesized directly from the generated spectrum, excitation, and vibrato parameters by using a Mel Log Spectrum Approximation (MLSA) filter.

3. PITCH ADAPTIVE TRAINING FOR HMM-BASED SINGING VOICE SYNTHESIS

HMM-based systems for speech synthesis heavily depend on training data in performance because these systems are "corpus-based." Therefore, HMMs corresponding to contextual factors that hardly ever appear in the training data cannot be well-trained. Algorithms used for designing speech databases that take into consideration the balance between contextual factors have been proposed [7] to solve this problem. Databases including various contextual factors should also be used in HMM-based singing voice synthesis systems. However, data have to be sparse because singing voices involve numerous contextual factors, e.g., pitch, tempo, key, beat, and dynamics, in addition to those used in reading speech synthesis. Pitch should especially be correctly covered since generated F_0 trajectories have a great impact on the subjective quality of synthesized singing voices.

A technique using pitch-shifted pseudo-data [3] is one solution to this problem. Since pitch is represented by a log F_0 parameter, pitch-shifted pseudo-data can easily be prepared by shifting log F_0 up or down in halftones. This technique makes it possible to increase the amount of F_0 training data without having to record large amounts of singing voice data. Spectrum and vibrato parameters are added by copying the same data assuming that they have not been affected by small amounts of pitch-shifting. The amount of training data is increased threefold by adding pitch-shifted pseudo-data. Therefore, decision trees increase in size when the minimum description length (MDL)-based criterion is used. Thus, context-clustering should be stopped when decision trees reach an appropriate size. The MDL criterion is used in the HMM-based singing voice synthesis system to determine when splitting nodes should be stopped. The heuristic weight is used to control the size of the decision trees. The quality of synthesized voices is improved when the sizes of the decision trees are properly adjusted.

Although the technique using pitch-shifted pseudo-data improves the quality of synthesized voices, this technique has four other problems:

1) The features included in a specific pitch range cannot be modeled since the pitch contexts are mixed due to pitch-shifted pseudo-data.

2) Whether or not singing voices can be synthesized outside the pitch range of the training data depends on the amount of pitch-shifting.

3) Pitch-shifted pseudo-data increase the computational cost in HMM training.

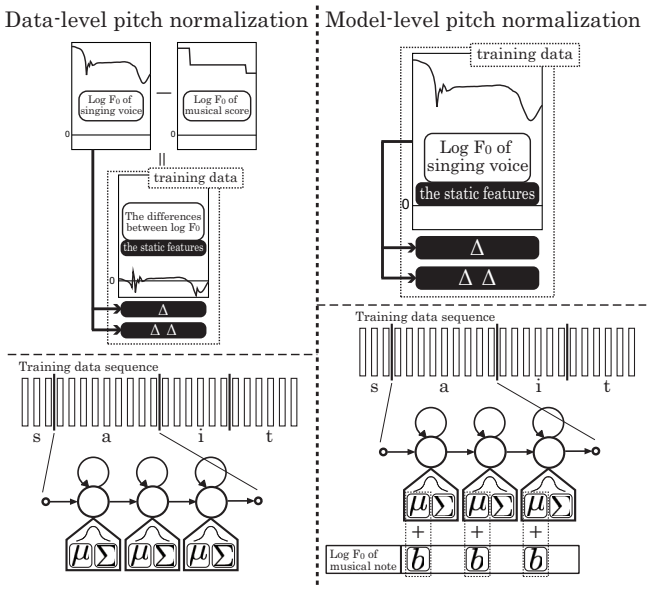

Fig. 2. *Comparison of data-level and model-level pitch normalizations.*

4) The amount of pitch-shifting and heuristic weights to control the number of parameters have to be properly adjusted manually.

Pitch normalization techniques are required to overcome these four problems.

Data-level pitch normalization [4] is one solution to these problems, where the differences between log F_0 sequences extracted from waveforms and the pitch of musical notes are used for training data. However, there are still two other problems:

5) Alignments of musical note levels are required to prepare the training data to calculate the differences.

6) If the alignments of musical note levels are fixed in HMM training, no parameters (spectrum, excitation, vibrato, or state duration) can be simultaneously estimated. However, if they are not fixed in HMM training, the differences cannot be estimated well because there are inconsistencies between the data and training.

This paper proposes model-level normalization of pitch to overcome these problems. The method of "speaker adaptive training" (SAT) [8] is applied to "pitch adaptive training." Fig. 2 compares between data-level and model-level pitch normalizations. The difference between the training speaker's voice and an average voice is assumed to be expressed in the SAT algorithm as a simple linear regression function of the mean vectors of state output distributions:

$$\boldsymbol{\mu}_i^{(f)} = \boldsymbol{W}_i^{(f)} \boldsymbol{\xi}_i \quad (1)$$

$$\boldsymbol{W}_i^{(f)} = \left[\boldsymbol{\zeta}_i^{(f)}, \boldsymbol{\epsilon}_i^{(f)} \right] \quad (2)$$

$$\boldsymbol{\xi}_i = \left[\boldsymbol{\mu}_i^{\top}, 1 \right]^{\top} \quad (3)$$

where $\boldsymbol{\mu}_i^{(f)}$, $\boldsymbol{W}_i^{(f)}$, and $\boldsymbol{\xi}_i$ correspond to the mean vectors of state i for training speaker f, a transformation matrix that

indicates the difference between training speaker f and an average voice, and extended mean vectors of the average voice. The SAT algorithm simultaneously estimates both the parameter set of HMMs and the set of transformation matrices for each training speaker so that the likelihood is maximized. However, mean $\hat{\mu}_i$ of static features of log F_0 in state i is defined in the ptich adaptive training algorithm as

$$\hat{\mu}_i = \boldsymbol{W}_i \boldsymbol{\xi}_i \quad (4)$$

$$\boldsymbol{W}_i = [1, b_i] \quad (5)$$

$$\boldsymbol{\xi}_i = [\mu_i, 1]^\top \quad (6)$$

where μ_i is the mean of the difference between log F_0 extracted from the waveform and pitch of a musical note. The b_i is log F_0 of a musical note that includes state i. Since the transformation matrices are fixed by the musical score, pitch adaptive training only estimates the parameter set of HMMs. As a result, all problems with the conventional method can be solved with the proposed training for 1-6 below:

For 1) The features included in a specific pitch range can be modeled because all training data have correct pitch contexts.

For 2) The differences between log F_0 extracted from the waveform and pitch of a musical note are modeled. Therefore, pitch that does not appear in the training data can be synthesized.

For 3) The total number of training data is not increased. The computational cost is also not increased.

For 4) The MDL criterion can be used to automatically control the number of parameters.

For 5) Alignments of musical note levels are not required for HMM training because of model-level normalization of pitch.

For 6) There is no need to fix alignments in HMM training. All parameters can be simultaneously estimated. There are no inconsistencies between data and training.

4. EXPERIMENTS

Subjective experiments were conducted to evaluate the performance of pitch adaptive training for HMM-based singing voice synthesis.

4.1. Experimental conditions

Seventy Japanese children's songs (total of 71.8 min) by a female singer were used. Singing voice signals were sampled at 48kHz and windowed with a 5-ms shift. The feature vectors consisted of spectrum, excitation, and vibrato parameters. The spectrum parameter vectors consisted of 49 STRAIGHT mel-cepstral coefficients including the zero coefficient, their delta, and delta-delta coefficients. The excitation parameter vectors consisted of log F_0, its delta, and delta-delta. The vibrato parameter vectors consisted of amplitude (cent) and frequency (Hz), their delta, and delta-delta.

A seven-state (including beginning and ending null states), left-to-right, no-skip structure was used for the hidden semi-Markov model (HSMM). The spectrum stream was modeled with single multi-variate Gaussian distributions. The excitation and vibrato streams were modeled with multi-space probability distribution HSMM (MSD-HSMM). The state durations of each model were modeled with a five-dimensional (equal to the number of emitting states in each model) multi-variate Gaussian distribution. The decision tree-based context-clustering technique was separately applied to distributions for the spectrum, excitation, vibrato, state duration, and timing [9]. The MDL criterion was used to control the size of the decision trees. A algorithm to generate speech parameters that took into consideration context-dependent global variance (GV) without silence was used for generating the spectrum parameters.

A baseline method, a conventional method using pitch-shift, and the proposed method using pitch adaptive training were evaluated. Note that, to be fair, heuristic weights to control the number of parameter used in the conventional and proposed methods were manually adjusted based on the baseline method because the MDL criterion could not be used for the conventional method. The range of pitch-shifted pseudo-data was $\pm$ a halftone in the conventional method.

Ten songs not included in the training data were used for the evaluation. Ten subjects were asked to rate the naturalness of the synthesized singing voices on a Mean Opinion Score (MOS) with a scale from 1 (poor) to 5 (good). Fifteen randomly selected musical phrases were presented to each subject. The experiments were carried out in a sound-proof room.

4.2. Experimental results

Three subjective listening tests were conducted. Figs. 3 and 4 present the results for the 10 songs that were used for training. The pitch range of the test songs (6.4 min, C4-D5) in Fig. 3 was included in the pitch range of the 10 training songs (7.1 min, C4-F5). The key of the test songs in Fig. 4, on the other hand, was transposed up to a half octave (G♭4-A♭5). Fig. 5 has the results for the 60 songs (65.4 min, G3-F5) that were used for training. The pitch range of the test songs in Fig. 5 was included in the pitch range of the 60 training songs. We can see from Fig. 4 that the conventional and proposed methods achieved better subjective scores than the baseline method when the key of the test songs was transposed up to a half octave. Compared with the conventional method, the proposed method could model pitch that did not appear in the training data because of model-level normalization of pitch. It can be seen from Figs. 3 and 5 that the features included in a specific pitch range could be modeled with the proposed method when the total amount of training data was increased.

Table 1 summarizes the computation time for HMM training. The conventional method spent a long time on computation because the amount of training data was increased threefold by adding pitch-shifted pseudo-data. The computation time for the proposed method was almost the same as that for the baseline method.

The log F_0 sequences of synthesized /a/ phonemes (A4, C5, E5, and G5) were plotted to evaluate the performance of excitation modeling. Fig. 6 shows the results. The solid line indicates a log F_0 sequence calculated with the musical score and the broken lines indicates the log F_0 sequences gener-

Table 1. *Computation time for HMM training.*

Method	Number of songs for training	
	10 songs	60 songs
Baseline	9.3h	85.5h
Conventional	25.3h	274.0h
Proposed	9.5h	93.0h

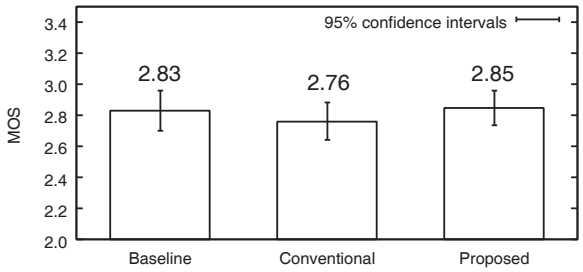

Fig. 3. *Subjective evaluation results: 10 songs were used for training. Pitch range of test songs was included in pitch range of 10 training songs.*

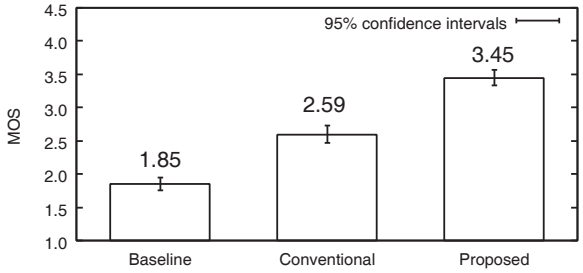

Fig. 4. *Subjective evaluation results: 10 songs were used for training. Key of test songs was transposed up to half octave.*

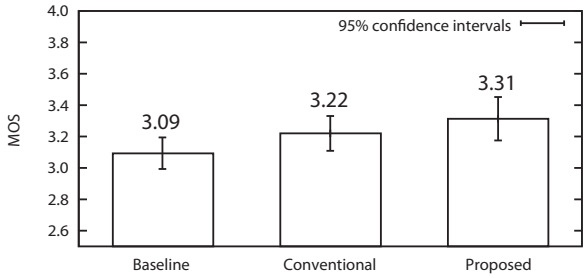

Fig. 5. *Subjective evaluation results: 60 songs were used for training. Pitch range of test songs was included in pitch range of 60 training songs.*

ated by all the methods. Note that vibrato models were not used to enable better visualization. Musical note "E5," which hardly ever appeared in the training data, could not be generated in the baseline method. It can be seen that the conventional method cannot generate log F_0 sequences for musical note "G5." The proposed method could generate all pitches even if a specific pitch did not appear in the training data.

5. CONCLUSIONS

We applied the method of "speaker adaptive training" to "pitch adaptive training" to the research discussed in this paper. This technique made it possible to normalize pitch based on musical notes in the training process. The experimental results revealed that the proposed technique could alleviate the data sparseness problem. Future work include evaluations which speech parameters affect the singing voice quality of the proposed technique.

6. ACKNOWLEDGEMENTS

The authors would like to thank Mr. Shinji Sako for constructing the database. The research leading to these results was partly funded by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication, Japan.

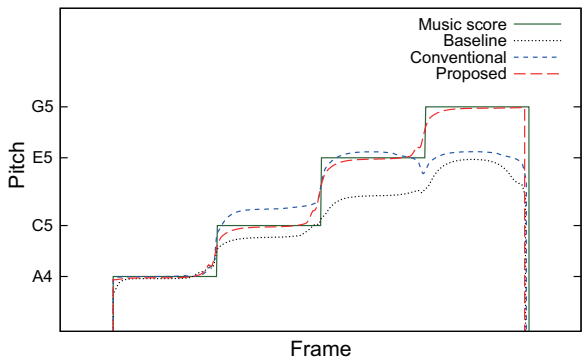

Fig. 6. *Log F_0 sequences of synthesized singing voices.*

7. REFERENCES

[1] 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. of Eurospeech*, 1999, pp. 2347–2350.

[2] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system —Sinsy," in *Proc. the 7th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2010, pp. 211–216.

[3] A. Mase, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-based singing voice synthesis system using pitch-shifted pseudo training data," in *Proc. of Interspeech*, 2010, pp. 845–848.

[4] K. Saino, M. Tachibana, and H. Kenmochi, "An HMM-based singing style modeling system for singing voice synthesizers," in *Proc. the 7th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2010, pp. 252–257.

[5] Yao Qian, Hui Liang, and Flank K. Soong, "Generating natural F0 trajectory with additive trees," in *Proc. of Interspeech*, 2008, pp. 2126–2129.

[6] Heiga Zen and Norbert Brauns chweiler, "Context-dependent additive log F0 model for HMM-based speech synthesis," in *Proc. of Interspeech*, 2009, pp. 2091–2094.

[7] A. Kuramatsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kawabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," in *Speech Communication*, 1990, vol. 9, pp. 357–363.

[8] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E-90D, no. 2, pp. 533–543, 2007.

[9] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proc. of ICSLP*, 2006, pp. 1141–1144.