An HMM-based Singing Voice Synthesis System

Keijiro Saino, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Department of Computer Science and Engineering
Nagoya Institute of Technology, Nagoya, Japan
{k-saino,zen,nankaku,ri,tokuda}@lavender.ics.nitech.ac.jp

Abstract

The present paper describes a corpus-based singing voice synthesis system based on hidden Markov models (HMMs). This system employs the HMM-based speech synthesis to synthesize singing voice. Musical information such as lyrics, tones, durations is modeled simultaneously in a unified framework of the context-dependent HMM. It can mimic the voice quality and singing style of the original singer. Results of a singing voice synthesis experiment show that the proposed system can synthesize smooth and natural-sounding singing voice.

Index Terms: singing voice synthesis, HMM, time-lag model.

1. Introduction

In recent years, various applications of speech synthesis systems have been proposed and investigated. Singing voice synthesis is one of the hot topics in this area [1–5]. However, only a few corpus-based singing voice synthesis systems which can be constructed automatically have been proposed.

Currently, there are two main paradigms in the corpus-based speech synthesis area: sample-based approach and statistical approach. The sample-based approach such as unit selection [6] can synthesize high-quality speech. However, it requires a huge amount of training data to realize various voice characteristics. On the other hand, the quality of statistical approach such as HMM-based speech synthesis [7] is buzzy because it is based on a vocoding technique. However, it is smooth and stable, and its voice characteristics can easily be modified by transforming HMM parameters appropriately. For singing voice synthesis, applying the unit selection seems to be difficult because a huge amount of singing speech which covers vast combinations of contextual factors that affect singing voice has to be recorded. On the other hand, the HMM-based system can be constructed using a relatively small amount of training data. From this point of view, the HMM-based approach seems to be more suitable for the singing voice synthesizer. In the present paper, we apply the HMM-based synthesis approach to singing voice synthesis.

Although the singing voice synthesis system proposed in the present paper is quite similar to the HMM-based text-to-speech synthesis system [7], there are two main differences between them. In the HMM-based text-to-speech synthesis system, contextual factors which may affect reading speech (e.g. phonemes, syllables, words, phrases, etc.) are taken into account. However, contextual factors which may affect singing voice should be different from those used in text-to-speech synthesis. Therefore, in the proposed system, contextual factors which may affect singing speech (i.e. tones, lyrics, and durations) are used. Another difference is introduction of an additional model to control start timing of each musical note. In human singing voices, there are differences between start timings of musical notes and speech as shown in Figure 1. In the present paper, we call them "time-lags." The time-lag can have negative values if its start timing is earlier than that of corresponding musical note. Since this could be an important factor to synthesize natural-sounding singing voice, they are modeled explicitly by time-lag models in the proposed system.

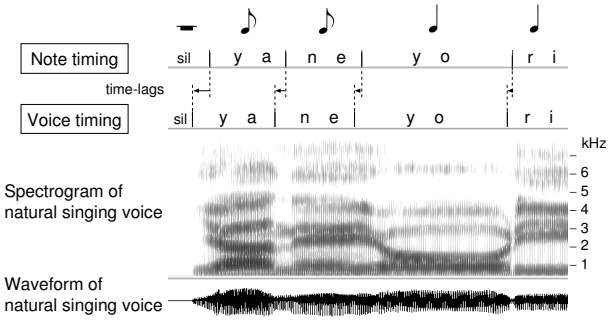

Figure 1: *An example of "time-lag."*

The rest of the present paper is organized as follows: Section 2 describes the HMM-based singing voice synthesis system. Details of time-lag models are described in Section 2.3. Section 3 shows results of a singing voice synthesis experiment. Conclusions and future plans are shown in the final section.

2. HMM-based Singing Voice Synthesis System

2.1. System Overview

Figure 2 shows the overview of the HMM-based singing voice synthesis system. It consists of training and synthesis parts.

In the training part, first we extract spectral (e.g., mel-cepstral coefficients [8]) and excitation (e.g., fundamental frequencies) parameters from a singing voice database and then they are modeled by context-dependent HMMs. Context-dependent state duration models and time-lag models are also estimated.

In the synthesis part, first an arbitrarily given musical score in-

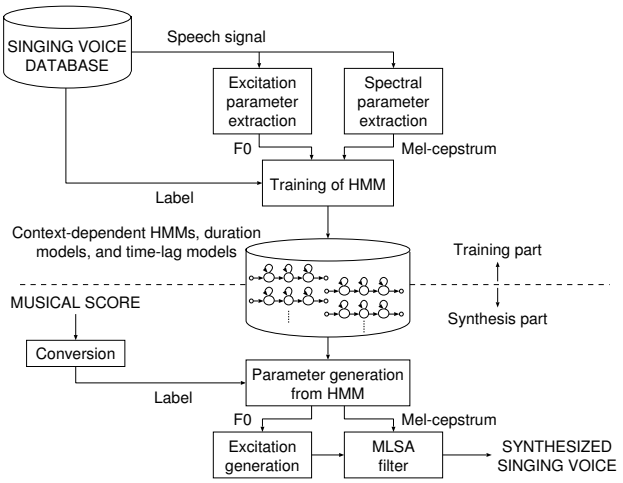

Figure 2: *The overview of the HMM-based singing voice synthesis system.*

cluding lyric to be synthesized is converted to a context-dependent label sequence. Secondly, according to the label sequence, a song HMM is constructed by concatenating the context-dependent HMMs. Thirdly, state durations of the song HMM are determined with respect not only to the state duration models but also to the time-lag models. Fourthly, spectral and excitation parameters are generated by the speech parameter generation algorithm [9]. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using Mel Log Spectrum Approximation (MLSA) filter [10].

This system is quite similar to the HMM-based text-to-speech synthesis system [7]. However, there are two main differences between them: contextual factors and time-lag models. In the following, details of them are described.

2.2. Contextual Factors

In the HMM-based text-to-speech synthesis system, contextual factors which may affect reading speech such as phoneme identity, part-of-speech, accent, stress, etc. have been taken into account [7]. However, contextual factors that affect singing voice should be different from those used in text-to-speech synthesis. In the present paper, the following contextual factors were considered:

phoneme: The preceding, current, and succeeding phonemes.

tone: The musical tones of the preceding, current, and succeeding musical notes (e.g. “A4”, “C5#”, “B3♭”, etc.).

duration: The durations of the preceding, current, and succeeding musical notes (in 100 ms unit).

position: The positions of the preceding, current, and succeeding musical notes in the corresponding musical bar (in triplet thirty-second note).

These contexts can automatically be determined from the musical score including lyric.

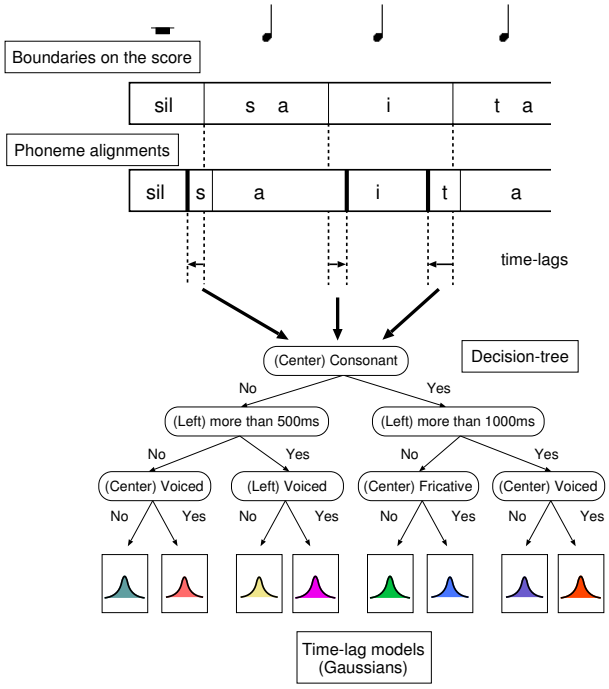

Figure 3: *Time-lag modeling.*

2.3. Time-lag Modeling

Another unique feature in the proposed system is time-lag modeling. In the case of singing voice synthesis, we have to obey rhythm or tempo of the music. Therefore, start timing of musical notes or phoneme durations in each musical note should be determined according to the musical score. However, if the score is strictly followed, the synthetic singing voice will be unnatural because there are time-lags between start timings of musical score and real human speech (see Fig. 1).

To model this phenomenon, we introduce “time-lag models.” After training context-dependent HMMs, the forced alignment is performed to the training data to obtain time-lags between musical notes and real human speech. We assign context-dependent labels similar to those described in 2.2 to the obtained time-lags. Then they are clustered by a decision tree in the same manner used for clustering spectral, fundamental frequency and duration parameters [7, 11]. As a result, we obtain decision-tree-clustered context-dependent time-lag models (one-dimensional Gaussians). Figure 3 shows its overview.

In the synthesis stage, first we determine the duration of each musical note from the given score including lyric to be synthesized. Then, the time-lags of musical notes and state durations of the song HMM are determined simultaneously so as to maximize the joint probability of time-lags and state durations:

$$P(\boldsymbol{d}, \boldsymbol{g} \mid \boldsymbol{T}, \Lambda) = P(\boldsymbol{d} \mid \boldsymbol{g}, \boldsymbol{T}, \Lambda) \cdot P(\boldsymbol{g} \mid \Lambda) \quad (1)$$

$$= \prod_{k=1}^{N} P(\boldsymbol{d}_k \mid T_k, g_k, g_{k-1}, \Lambda) \cdot P(g_k \mid \Lambda), \quad (2)$$

where N denotes the total number of musical notes in this song, $\boldsymbol{d}_k$ denotes the state durations in the k-th musical note, T_k denotes the duration of the k-th musical note determined from the given musical score, and g_k denotes the time-lag of the start timing on the k+1-th musical note. Note that $g_0 = g_N = 0$ since the corresponding boundaries are the beginning and the end of the musical score. The time-lags $\boldsymbol{g}$ and state durations $\boldsymbol{d}$ which maximize Eq.(1) can be obtained by solving the following a set of linear equations:

$$\boldsymbol{A}\boldsymbol{g} = \boldsymbol{b} \tag{3}$$

$$\boldsymbol{d}_k = \boldsymbol{\mu}_{d_k} + \rho_k \cdot \mathrm{diag}^{-1}\left(\boldsymbol{\Sigma}_{d_k}\right), \tag{4}$$

where

$$a_{i,i} = -1 - \frac{\sum_{t=1}^{K\cdot n_{i+1}} \sigma^2_{d_{(i+1),t}}}{\sum_{t=1}^{K\cdot n_i} \sigma^2_{d_{i,t}}} - \frac{\sum_{t=1}^{K\cdot n_{i+1}} \sigma^2_{d_{(i+1),t}}}{\sigma^2_{g_i}}, \tag{5}$$

$$a_{(i+1),i} = \frac{\sum_{t=1}^{K\cdot n_{i+1}} \sigma^2_{d_{(i+1),t}}}{\sum_{t=1}^{K\cdot n_i} \sigma^2_{d_{i,t}}}, \tag{6}$$

$$a_{i,(i+1)} = 1, \tag{7}$$

$$a_{i,j} = 0 \qquad (j \neq i \pm 1, j \neq i), \tag{8}$$

$$b_i = \frac{\sum_{t=1}^{K\cdot n_{i+1}} \sigma^2_{d_{(i+1),t}}}{\sum_{t=1}^{K\cdot n_i} \sigma^2_{d_{i,t}}} \left(T_i - \sum_{t=1}^{K\cdot n_i} \mu_{d_{i,t}}\right) - \left(T_{i+1} - \sum_{t=1}^{K\cdot n_{i+1}} \mu_{d_{(i+1),t}}\right) - \frac{\sum_{t=1}^{K\cdot n_{i+1}} \sigma^2_{d_{(i+1),t}}}{\sigma^2_{g_i}} \mu_{g_i}, \tag{9}$$

$$(1 \leq i \leq N{-}1,\ 1 \leq j \leq N{-}1)$$

$$\rho_k = \frac{(T_k - g_{k-1} + g_k) - \sum_{t=1}^{K\cdot n_k} \mu^2_{d_{k,t}}}{\sum_{t=1}^{K\cdot n_k} \sigma^2_{d_{k,t}}}. \tag{10}$$

In the above equations, n_k denotes a number of phonemes in the k-th musical note, K denotes the number of states in each phoneme HMM, $a_{i,j}$ denotes the (i, j)-th element of $\boldsymbol{A}$, b_i denotes the i-th element of $\boldsymbol{b}$, $\mu_{d_{k,t}}$ and $\sigma^2_{d_{k,t}}$ denotes the mean and the variance of the duration model of the t-th state duration in the k-th musical note, respectively, $\boldsymbol{\mu}_{d_k}$ and $\boldsymbol{\Sigma}_{d_k}$ denotes the mean vector and the diagonal covariance matrix of the duration model in the k-th musical note, respectively. Since $\boldsymbol{A}$ becomes an asymmetric tri-diagonal matrix, Eq.(3) can easily be solved by a fast algorithm.

Table 1: *Singing voice database.*

Singer	1 male (non-professional)
Songs	60 Japanese children's songs (about 72minutes in total)
Sampling Rate	44.1kHz
Quantization	16bit

Table 2: *Mel-cepstral analysis condition.*

Sampling Rate	16kHz
Frame Shift	5ms
Window Length	25ms
Window Function	Blackman Window
Spectral Feature	24 mel-cepstral analysis [8]

3. Experiment

3.1. Experimental Conditions

Although a variety of reading speech databases were available, we could not find any appropriate singing voice database. Therefore, we recorded a singing voice database by ourselves. The overview of this database is summarized in Table 1. To improve the quality of the database, phoneme boundaries, fundamental frequencies (F_0), and musical scores were manually corrected.

The speech analysis conditions are shown in Table 2. Each feature vector consisted of spectrum and F_0 parameter vectors: each spectrum parameter vector consisted of 0–24th mel-cepstral coefficients, their delta and delta-delta coefficients, and the F_0 parameter vector consisted of log F_0, its delta and delta-delta. We used the five-state left-to-right with no-skip HMM structure. Thirty-six Japanese phonemes including silence and pause were used. The decision-tree based context clustering was applied to spectrum, F_0, duration and time-lag models, separately. We used the MDL criterion to stop tree growth. The resultant trees of spectrum, F_0, duration and time-lag models had 4264, 2477, 197, and 317 leaf nodes, respectively.

3.2. Singing Voice Synthesis Experiment

By using estimated HMMs, we synthesized singing voices. As a result, smooth and natural-sounding synthetic singing voice was obtained. Samples of synthesized singing voice are available at [12]. It shows that the HMM-based speech synthesis system was successfully applied to the singing voice synthesis. Figure 4 plots an example of F_0 pattern of a synthetic singing voice. It shows that generated F_0 pattern was slightly lower than the F_0 pattern defined by the musical score. This is because the original singer had a tendency to sing a little flat, and it was confirmed that the tendency was successfully reflected to the synthetic voice by the system.

Effects of the use of time-lag models can be seen in Fig. 5. It is known that human singing voice has a tendency to start consonants a little earlier than the timing of musical note if the note start with consonants. Figure 5 shows that the start timings of each musical

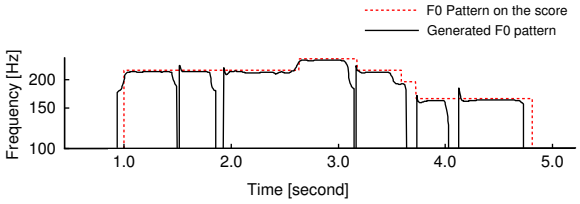

Figure 4: *An example of generated F_0 pattern.*

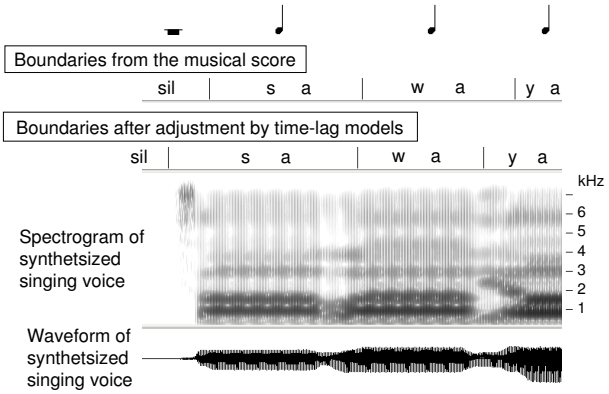

Figure 5: *An example of synthetic singing voice waveform.*

note have been controlled according to the tendency by the time-lag models.

To evaluate the effectiveness of time-lag models, we conducted a subjective listening test. Ten songs not included in the training data were divided every four to six musical bars. As a result, we obtained 27 musical phrases. Then we synthesized singing voices from the HMMs with and without using the time-lag models. Fourteen subjects were asked to rate the naturalness of synthesized singing voices on Mean Opinion Score (MOS) with a scale from 1 (poor) to 5 (good). For each subject, randomly selected 15 musical phrases were presented. Experiments were carried out in a sound-proof room. Speech samples were played with a click for every quoter note synchronized to the corresponding musical score. Figure 6 shows the experimental results. It can be seen from the figure that the introduction of the time-lag models improved the naturalness of synthetic singing voice. Interestingly, many listeners said that the characteristics of the original singer had been found in the synthetic voice.

4. Conclusions

In the present paper, a corpus-based singing voice synthesis system based on hidden Markov models (HMMs) was proposed. This system employed the HMM-based speech synthesis method to synthesize singing voice. Musical information such as lyrics, tone, duration was modeled simultaneously in a unified framework of the context-dependent HMM. Experimental results showed that the proposed system could mimic the voice quality and singing style of the original singer and natural-sounding singing voice could be

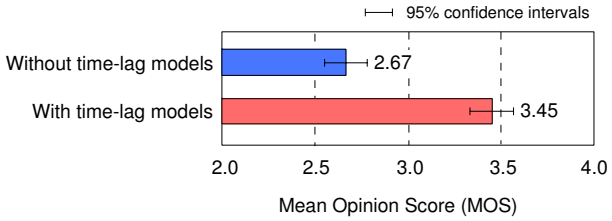

Figure 6: *Evaluation of naturalness for each methods.*

synthesized.

In the present paper, only four contextual factors were considered. However, other kinds of contextual factors such as dynamics should be included to improve the ability of this system. Future work will focus on the use of these kinds of contextual factors.

5. Acknowledgments

This work was partly supported by MEXT e-society project. The authors would like to thank Masanori Ito, Chisato Ishikawa, and Hiroaki Kuwabara for preparing and correcting database, and Shinji Sako for designing the base system.

6. References

[1] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George, "Concatenation-based MIDI-to-singing voice synthesis," in *Proc. 103rd Meeting of the AES*, New York, 1997.

[2] M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *Proc. ICASSP*, vol.1, pp.435–438, 1997.

[3] K. Lomax, "The development of a singing synthesizer," in *Proc. SPECOM*, pp.146–150,1996.

[4] Lu H.-L., "Toward a high-quality singing synthesizer with vocal texture control," *PhD thesis, Stanford University*, 2002.

[5] Ortolà J., "Musical and phonetic controls in a singing voice synthesizer," *Master's thesis, Polytechnics University of Valencia*, 2001.

[6] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP* , vol. 1, pp. 373–376, 1996.

[7] 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*, vol.5, pp.2347–2350, 1999.

[8] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol.1, pp.137–140,1992.

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, vol.3, pp.1315–1318, 2000.

[10] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*, vol.1, pp.93–96, 1983.

[11] J. J. Odell, "The use of context in large vocabulary speech recognition," *PhD thesis, Cambridge University*, 1995.

[12] `http://kt-lab.ics.nitech.ac.jp/~k-saino/music/`.