Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 308–322

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling☆

Takashi Nose[a,*], Misa Kanemoto[b], Tomoki Koriyama[b], Takao Kobayashi[b]

[a] *Graduate School of Engineering, Tohoku University, 6-6-05 Aramaki aza Aoba, Aoba-ku, Sendai 980-0011, Japan*
[b] *Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan*

Received 23 May 2014; received in revised form 28 March 2015; accepted 7 April 2015
Available online 16 April 2015

Abstract

This paper proposes a singing style control technique based on multiple regression hidden semi-Markov models (MRHSMMs) for changing singing styles and their intensities appearing in synthetic singing voices. In the proposed technique, singing styles and their intensities are represented by low-dimensional vectors called style vectors and are modeled in accordance with the assumption that mean parameters of acoustic models are given as multiple regressions of the style vectors. In the synthesis process, we can weaken or emphasize the intensities of singing styles by setting a desired style vector. In addition, the idea of pitch adaptive training is extended to the case of the MRHSMM to improve the modeling accuracy of pitch associated with musical notes. A novel vibrato modeling technique is also presented to extract vibrato parameters from singing voices that sometimes have unclear vibrato expressions. Subjective evaluations show that we can intuitively control singing styles and their intensities while maintaining the naturalness of synthetic singing voices comparable to the conventional HSMM-based singing voice synthesis.

Keywords: HMM-based singing voice synthesis; Singing style control; Multiple-regression HSMM; Pitch adaptive training; Vibrato modeling

1. Introduction

Speech synthesis is the key technologies for human computer interaction (HCI) systems, and the interactive robot is one of the most typical and important applications to be realized in HCI systems. Recently, a humanoid robot named HRP-4C (Kaneko et al., 2009) was developed whose appearance is quite close to that of a human (Nakaoka et al., 2009). For such a state-of-the-art interactive robot, more advanced speech synthesis with rich para-linguistic and non-linguistic information, e.g., affections, emotions, speaking styles, and speaker characteristics, is indispensable. In addition, a function of synthesizing not only speech but also singing voice is desirable to achieve HCI systems that is capable of making the speech communication more diverse and rich like a human. This is because in our daily life there are a variety of music pieces including singing voices that are capable of relaxing or exciting us and some people communicate their feeling to others by singing. If an interactive robot has a function of singing voice synthesis with

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author. Tel.: +81 22 795 7112.
E-mail addresses: tnose@m.tohoku.ac.jp (T. Nose), koriyama@ip.titech.ac.jp (T. Koriyama), takao.kobayashi@ip.titech.ac.jp (T. Kobayashi).

http://dx.doi.org/10.1016/j.csl.2015.04.001
0885-2308/

various singing styles, the application of the robot will expand not only to home entertainment but also to business showcase, exhibition, musical concert, and so on. Also in the education area, such an advanced and sophisticated interactive robot will give a good impact at the music class. These applications have a possibility of providing a new communication/interaction style between a human and a robot to our future life.

Singing voice synthesis is becoming an attractive application for speech synthesis in these days, and several products such as VOCALOID (Kenmochi and Ohshita, 2007) have become popular in the entertainment industry, especially in Japan. In the singing voice synthesis, users can easily create singing voices of certain singers or characters by inputting arbitrary musical notes (or MIDI codes) and lyrics, and this provides composers with an assistance method for adding original singing voices to their compositions. Recently, singing voice synthesizers have been utilized not only for hobby use but also for professional music production, a singing robot, karaoke, and live music, which shows the potential capability for entertainment and amusement applications (Tachibana et al., 2010; Kenmochi, 2012).

To develop a singing voice synthesis system, various approaches have been proposed (Cook, 1996; Rodet, 2002). The techniques based on speech production models, e.g., (Cook, 1993), and formant synthesis, e.g., (Sundberg, 2006), have an advantage that their model parameters have physical meanings and hence we are able to modify the synthetic singing voice by carefully controlling the parameters. However, the synthesis performance highly depends on the target voice, and the quality of output voice is not always satisfactory, which is the main drawback in singing voice synthesis. In terms of spectral reproducibility, concatenative synthesis, e.g, (Macon et al., 1997; Bonada et al., 2003; Kenmochi and Ohshita, 2007), outperforms the above techniques because recorded singing samples, such as diphones and sustained vowels, are directly used without physical modeling. A limitation of the concatenative synthesis is that singing samples, which are used as synthesis units, are recorded separately whereas we sing a song in a continuous manner beyond phonemes, syllables, and words. As a result, it is difficult to model the prosodic characteristics of singers and singing styles, and rule-based prosody generation is generally used. However, this heuristic approach is not always sufficient to synthesize singing voices with a wide variety of singing styles of various singers.

Singing voice synthesis based on hidden Markov models (HMMs) (Sako et al., 2004; Saino et al., 2006) is an alternative approach that enables simultaneous modeling of spectral and prosodic characteristics of singers and singing styles from a continuous singing voice corpus. The basic framework of the HMM-based singing voice synthesis is the same as that of HMM-based speech synthesis (Yoshimura et al., 1999). Although the baseline quality of the HMM-based singing voice synthesis has been steadily improved by introducing several techniques such as time-lag modeling (Saino et al., 2006), frame-based vibrato modeling (Oura et al., 2010), and pitch adaptive training (Oura et al., 2012), studies for diversifying singing voices are rather limited (Saino et al., 2010). By contrast, a variety of techniques, e.g., speaker adaptation (Tamura et al., 2001), style interpolation (Tachibana et al., 2005), and style control (Nose et al., 2007), have been proposed in HMM-based speech synthesis research area for adding or controlling various speaker and style characteristics (Yamagishi et al., 2009; Nose and Kobayashi, 2011). However, few studies have so far applied these techniques to singing voice synthesis (Sung et al., 2011).

In this study, we apply the style control technique of synthetic speech (Nose et al., 2007) to the HMM-based singing voice synthesis, which enables users to change the singing style expressivity in an intuitive and continuous manner.[1] In the proposed technique, multiple singing styles and their expressivity are represented by a low dimensional vector named a style vector and are simultaneously modeled using multiple-regression hidden semi-Markov models (MRHSMMs) (Niwase et al., 2005). The style vector is used as a explanatory variable of the MRHSMM where the mean parameter of each probability density function (pdf) is assumed to be given by a multiple regression of the style vector. In the model training, the parameters of MRHSMMs are estimated with training songs and the corresponding style vectors using maximum likelihood estimation with the EM algorithm. In the singing voice synthesis, we can control the singing style expressivity by changing the style vector.

To improve the modeling accuracy of pitch in the case of a limited amount of training data, we extend the model-space pitch adaptive training (Oura et al., 2012) into the feature-space one for the MRHSMM. In the proposed modeling, the observation features of static log fundamental frequency (F0) values are normalized using the pitch of the corresponding note in the parameter re-estimation process by the EM algorithm. By using pitch adaptive training, we can generate better F0 trajectories that closely follow the original pitch of given notes. In the singing voice synthesis, the vibrato modeling is also important for natural sounding synthetic singing voices (Maher and Beauchamp, 1990). However, the

[1] Part of this work was presented at INTERSPEECH 2013 (Nose et al., 2013).

singing voices sometimes have unclear vibrato expressions, making conventional vibrato modeling techniques such as (Oura et al., 2010) unsuitable. To alleviate this problem, we also propose a technique for the robust vibrato parameter extraction using zero-crossing and energy with a moving average filter. Through subjective evaluation tests, we show that the intuitive singing style control is well achieved while maintaining the naturalness of the synthetic voices.

2. Brief overview of conventional HMM-based singing voice synthesis

We first briefly outline the HMM-based singing voice synthesis that is the basis of this study. Most processes of model training and parameter generation are the same as those of the HMM-based speech synthesis. Specifically, we record singing voices of multiple songs as training data in advance. From the musical scores of these songs, the phonemes and their contextual information, e.g., pitch, duration, and position of preceding/current/succeeding note, are extracted and converted into context-dependent labels. Context-dependent HMMs are then trained from the training data using the EM algorithm and parameter tying by tree-based context clustering (Young et al., 1994). The spectral and F0 features are simultaneously modeled using multi-space probability distribution HMMs (Tokuda et al., 1999). To model the state duration of HMM appropriately, a hidden semi-Markov model (HSMM) (Zen et al., 2007) is generally used that has an explicit duration distribution in each state.

In the synthesis phase, the input musical score is converted into context-dependent labels and spectral and prosodic feature trajectories are estimated from the corresponding context-dependent HMM sequence using the parameter generation algorithm (Tokuda et al., 1995). Finally, the singing voice waveform is synthesized from the generated spectral and prosodic parameters using vocoding methods such as cepstral vocoding with MLSA filter (Imai, 1983) and STRAIGHT (Kawahara et al., 1999).

3. Singing style control for HMM-based expressive singing voice synthesis

3.1. Modeling of singing styles using multiple regression HSMM (MRHSMM)

To model multiple singing styles and their intensities, we use multiple-regression HSMM (MRHSMM) (Nose et al., 2007). In the MRHSMM, mean parameters, $\boldsymbol{\mu}_i$ and m_i, of output and state-duration probability density functions (pdfs) of the i-th state are assumed to be given by a multiple regression of a low dimensional vector, i.e., a style vector, $\boldsymbol{v}$ as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} \tag{1}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{2}$$

$$\boldsymbol{\xi} = [1, v_1, v_2, \ldots, v_L]^\top \tag{3}$$

$$= [1, \boldsymbol{v}^\top]^\top \tag{4}$$

where L is the dimension of the style vector, and v_n is the intensity/expressivity of the n-th singing style. $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are $M \times (L+1)$ and $1 \times (L+1)$ regression matrices, respectively, and M is the dimension of the feature vector. The space where the style vector is defined is called style space, and the style space is determined by the style vectors associated with training data. Fig. 1 shows an example of one- and three-dimensional style spaces.

The pdfs at state i are thus expressed as

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \boldsymbol{\Sigma}_i) \tag{5}$$

$$p_i(d) = \mathcal{N}(d; \boldsymbol{H}_{p_i}\boldsymbol{\xi}, \sigma_i^2) \tag{6}$$

where $\boldsymbol{o}$ and $\boldsymbol{\Sigma}_i$ are the observation vector and covariance matrix of the output pdf, respectively, and d and σ_i^2 are the state duration and variance of state-duration pdf, respectively.

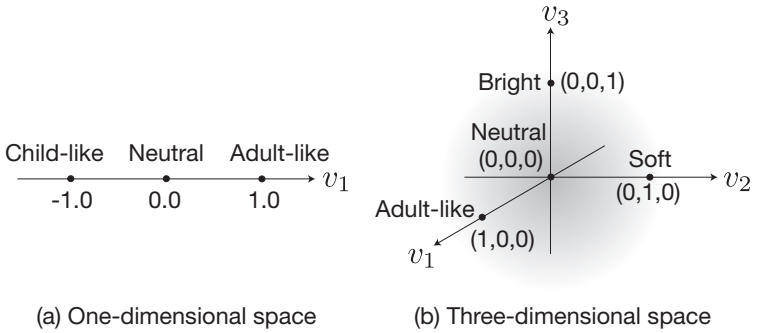


Fig. 1. Example of style spaces. Style vectors for respective singing styles of training utterances are shown.

In the model training, different style vectors are set to the training singing voices in accordance with the singing style expressivity appearing in each voice. The optimum model parameter set $\boldsymbol{\lambda}^*$ including regression matrices is estimated using the maximum likelihood criterion as follows:

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \prod_{k=1}^{K} P(\boldsymbol{o}^{(k)}|\boldsymbol{\lambda}, \boldsymbol{v}^{(k)}) \tag{7}$$

where $\boldsymbol{o}^{(k)}$ and $\boldsymbol{v}^{(k)}$ are the k-th singing voice and the corresponding style vector. We use style-independent variance parameters, $\boldsymbol{\Sigma}_i$ and σ_i^2, estimated using all training data in the training of MRHSMMs, which is the same condition as our previous study (Nose et al., 2007). Details of the estimation formulas for regression matrices and variance parameters can be found in (Nose et al., 2007).

3.2. Singing style control of synthetic singing voice based on MRHSMM

Fig. 2 shows the outline of the process of synthesizing a singing voice.

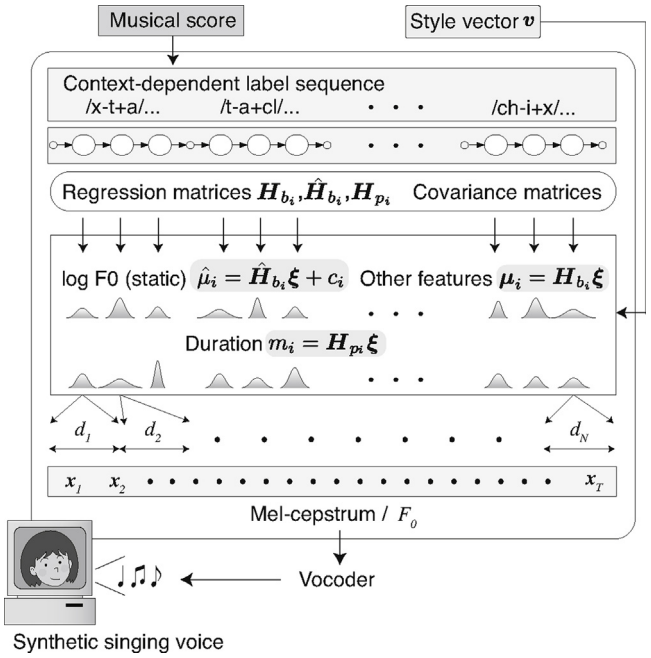


Fig. 2. Outline of synthesis process of singing style control based on MRHSMMs.

In the synthesis phase, users give a style vector that corresponds to their intended singing style expressivity in the style space. Mean parameters of output and state-duration pdfs are calculated from the given style vector and regression matrices of trained MRHSMMs using Eqs. (1) and (2). As a result, an ordinary HSMM sequence is obtained, and singing voice parameters are then generated from the HSMM sequence using the speech parameter generation algorithm as described in Section 2. Users can continuously control the singing style and its expressivity appearing in the synthesized singing voice by changing the input style vector since the style vector can have continuous values.

4. Robust pitch modeling for singing style control

4.1. Pitch modeling using feature-space pitch adaptive training for MRHSMM

As described in Section 3, singing style expressivities can be expected to be controlled by using MRHSMMs when sufficient training data is available. However, the reproducibility of each acoustic feature strongly depends on the training data because the HMM-based singing voice synthesis is a corpus-based approach. As for the pitch feature, which is one of the most important features in singing voice synthesis, it is difficult to generate a desirable F0 contour that closely follows the notes when the pitch contexts of the training data have poor coverage. This is also a crucial problem in the singing style control. In the conventional MRHSMM-based style control of synthetic speech (Nose et al., 2007), there is no framework for avoiding the coverage problem of pitch contexts, and hence the robust pitch modeling technique is strongly required in the case of singing voice synthesis. For this purpose, we here propose a pitch modeling technique by extending the pitch adaptive training (Oura et al., 2012) into the training of MRHSMMs in a feature space.

In the conventional pitch adaptive training of HMM/HSMM (Oura et al., 2012), the mean parameter μ_i of the pdf for the static feature of log F0 in the i-th state is given by a shift from an F0-normalized mean parameter $\hat{\mu}_i$ as

$$\mu_i = \hat{\mu}_i + c_i \tag{8}$$

where c_i is the log F0 value of the note corresponding to the i-th state. In this case, $\hat{\mu}_i$ represents not an absolute but a relative log F0 mean based on the note information, and Eq. (8) can be viewed as model-space normalization of the static log F0 feature. To simplify the extension of the pitch adaptive training to MRHSMM, instead we use the following feature-space normalization that is equivalent to Eq. (8) as

$$\hat{o}_t = o_t - c_i \tag{9}$$

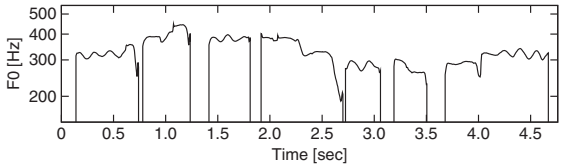
where o_t and $\hat{o}_t$ are log F0 observations of voiced frames at time t before and after the F0 normalization, respectively. Note that the variance parameters do not vary since the transformation is only the shifting. Given the note sequence $\boldsymbol{c}^{(k)}$ of the k-th singing voice, the optimum model parameter set $\boldsymbol{\lambda}^*$ after the F0 normalization is estimated as follows:

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \prod_{k=1}^{K} P(\boldsymbol{o}^{(k)}|\boldsymbol{\lambda}, \boldsymbol{v}^{(k)}, \boldsymbol{c}^{(k)}). \tag{10}$$

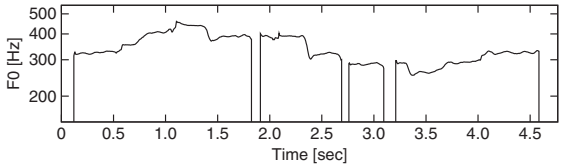
The estimation formula of regression matrices $\hat{\boldsymbol{H}}_{b_i}$ at the i-th state for the static feature of the normalized log F0 is obtained using Eq. (9) as follows:

$$\hat{\boldsymbol{H}}_{b_i} = \left(\sum_{k=1}^{K} \sum_{t=1}^{T^{(k)}} \sum_{d=1}^{t} \gamma_t^d(i) \left[\sum_{\tau=t-d+1}^{t} (o_\tau^{(k)} - c_i) \right] \boldsymbol{\xi}^{(k)\top} \right) \cdot \left(\sum_{k=1}^{K} \sum_{t=1}^{T^{(k)}} \sum_{d=1}^{t} \gamma_t^d(i) \cdot d \cdot \boldsymbol{\xi}^{(k)} \boldsymbol{\xi}^{(k)\top} \right)^{-1} \tag{11}$$

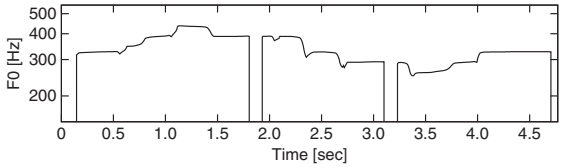
where $o_\tau^{(k)}$ is the log F0 value of the k-th observation sequence $\boldsymbol{o}^{(k)}$ at time t, and $T^{(k)}$ is the number of frames of $\boldsymbol{o}^{(k)}$. $\gamma_t^d(i)$ is the probability of being in state i at a period of time from $t-d+1$ to t given $\boldsymbol{o}^{(k)}$. $\boldsymbol{v}^{(k)}$ is the style vector corresponding to $\boldsymbol{o}^{(k)}$, and $\boldsymbol{\xi}^{(k)} = [1, \boldsymbol{v}^{(k)\top}]^\top$. Compared to the case of the conventional model-space approach, the derivation of the pitch adaptive training for MRHSMM can be very simple by using the feature-space approach. Note that this



(a) Original voice (clear vibrato)



(b) Original voice (unclear vibrato)



(c) Synthetic voice (without vibrato modeling)

Fig. 3. Example of vibrato expressions by different singers and singing styles.

approach is different from the data-level F0 normalization (Saino et al., 2010) because the normalization is conducted not in the data preparation phase but within the parameter estimation process. In the data-level F0 normalization, the alignment between feature vectors and musical notes are fixed in advance whereas the alignment is not fixed in the forward-backward step of the EM algorithm.

In the parameter generation, first the normalized mean parameter of each state is calculated using Eq. (1). Then, the normalized mean parameter is transformed into the mean parameter of each state using the input note and Eq. (8). Finally, a log F0 contour is generated from the pdf sequence in an ordinary manner of the HMM-based singing voice synthesis.

4.2. Unclear vibrato problem

Another issue with pitch modeling in singing voice synthesis is a vibrato modeling problem. Vibrato is a kind of singing expression mainly consisting of a regular and pulsating change of pitch and is typically represented by a periodic F0 variation in singing voice synthesis. When the vibrato is not explicitly modeled in the HMM-based singing voice synthesis, the periodicity and phase characteristics of the vibrato can not be taken into account in the F0 model training, and the resultant F0 contour is highly flattened in long-tone segments and vibrato expression in the original singing voice is not reproduced. In this paper, we define *segment* as a phone unit. When the segment is a vowel and have long tone, the segment is called long-tone segment. To overcome this problem, a frame-based vibrato modeling technique was proposed (Oura et al., 2010). In this technique, the F0 contour is assumed to be a periodically time-variant sequence for vibrato segments, and vibrato parameters are calculated from the amplitude and the interval between two peaks of the sinusoidal F0 contour. The parameters are extended to frame-level features using interpolation. These features are added to the observation vectors and are modeled simultaneously by context-dependent HMMs.

The above frame-based vibrato modeling technique is effective when the singer is professional and the vibrato expression is very clear. However, the vibrato expression highly depends on a target singer or a target singing style, and the vibrato expression is not always very clear.

In Fig. 3(a) and (b) are F0 contours extracted from the same part of the same song used in (Oura et al., 2010) and from an adult-like style song used in the experiments in Section 5, respectively. Note that (a) and (b) were sung in different singing styles, i.e., classical and pop, respectively, by different singers. For comparison, a synthetic F0 contour generated using the training data of the adult-like style without vibrato modeling is also shown in (c). In the F0 contour of (a), we can see a clear vibrato expression in the long-tone segments. By contrast, the amplitude is much

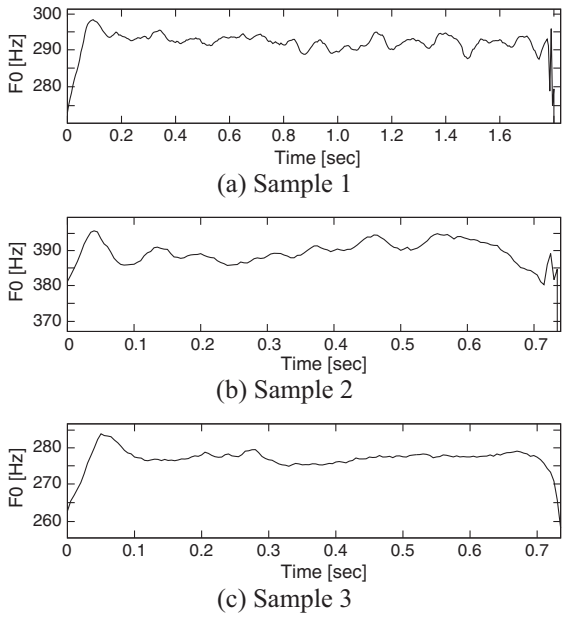


Fig. 4. Example of unclear vibrato expressions in the long-tone segments of natural singing songs.

smaller for the vibrations of (b) than for those of (a), and the vibrato periodicity is not always clear. Other detailed samples of unclear vibrato in the different long-tone segments are shown in Fig. 4. It is apparent that the clear periodic F0 vibration can not be found in these samples and reliable vibrato parameters is difficult to be extracted using the conventional approach. When comparing (b) and (c), the F0 contour of the long-tone segments is highly flattened in the synthetic voice and the fluctuation appearing in the original voice has disappeared, which degrades the naturalness of the synthetic singing voice. Therefore, a vibrato modeling technique applicable to such an unclear vibrato expression is important to synthesize songs with various speakers and singing styles.

4.3. Robust vibrato modeling

We propose a vibrato modeling technique that can be used even when the vibrato expression is not very clear. From a preliminary analysis, we defined the long-tone segment as a segment that has a single vowel and its duration is more than 600 ms. In each long-tone segment, the vibrato is modeled by a sine wave in which the fundamental period and the amplitude were assumed to be constant. The vibrato parameter vector consists of a vibrato rate parameter r in Hz and a vibrato amplitude parameter a in cent of a vibrato function $g(t)$ for log F0. $g(t)$ is given by

$$g(t) = a \cdot C \cdot \sin(r \cdot t) \tag{12}$$

where $C = \log 2/1200$ is a constant for scaling between the values of log F0 and F0 in cent. In the actual log F0 contour $f(t)$ of a long tone, the vibrato appears around an underlying gradual F0 movement. To obtain the vibrato pattern $\hat{f}(t)$ where the gradual movement is removed, we apply a moving average filter to $f(t)$ and subtract the output from $f(t)$. Then, the vibrato rate parameter is calculated from the zero-crossing rate n_z of $\hat{f}(t)$ as

$$r = \pi n_z \tag{13}$$

where the zero-crossing rates of $\hat{f}(t)$ and $g(t)$ were assumed to be equal. In this study, we restrict the range of vibrato frequency from 5 to 8 Hz on the basis of the previous study (Oura et al., 2010). By using a similar manner, the vibrato amplitude parameter a is given by

$$a = \sqrt{\frac{\sum_{t=0}^{T-1} \left[f(T_s\, t) \right]^2}{\sum_{t=0}^{T-1} \sin^2(\pi\, n_z T_s\, t)}} \tag{14}$$

where the energy between long-tone segments of $\hat{f}(t)$ and $g(t)$ were assumed to be equal. T and T_s [s] are the length of the long tone and the frame period, respectively. Although a vibrato amplitude was restricted from 30 to 150 cent in (Oura et al., 2010), we set the lower limit to zero to take into account a smaller vibrato expression. Note that the vibrato parameters are not determined and not modeled in the not-long-tone segments in this study.

A single set of vibrato rate and amplitude parameters is determined for a long-tone segment. We construct two-dimensional vibrato feature vectors by combining the rate and amplitude parameters. In the model training, vibrato parameter vectors are modeled using context-dependent single Gaussian pdfs. In the synthesis stage, first a log F0 contour is generated from HSMMs and long-tone segments are determined in accordance with phone durations. For the long-tone segments, vibrato is added by calculating a superposition of the generated log F0 contour and the vibrato function $g(t)$ determined by the vibrato parameters. We use linear interpolation so that the resultant log F0 sequence smoothly changes in the first and last 50 ms.

5. Experiments

5.1. Experimental conditions

We used twenty-seven Japanese traditional songs sung by a female voice actress in two different singing styles. As the styles, we chose age-related singing styles, i.e., child-like and adult-like styles each of which can be though to be a variation of singing styles. In the recording of the singing voices, we directed the singer to sing like a child or an adult and named the resultant styles "child-like" and "adult-like" styles, respectively. To examine whether listeners can perceive the age difference for both styles of the recorded singing voices, we conducted a preliminary listening test. We cut out a singing voice phrase that lasted eight seconds from the beginning of each song. Six phrases were randomly chosen for each participant. Fifteen participants listened to two singing voice samples of different styles for each phrase and were asked which sample they perceived to be younger. Participants could choose "indeterminable" if they could not perceive any age difference. The results showed that all phrases of the child-like style were perceived to be younger than those of the adult-like style, which shows that there is a clear style difference in terms of the perceived age.

In the experiments, twenty-five songs and two songs were chosen as training data and test data, respectively. Singing voice signals were sampled at a rate of 16 kHz and the frame shift was 5 ms. We used STRAIGHT analysis (Kawahara et al., 1999) for acoustic feature extraction and extracted spectral envelope, F0, and aperiodicity features. The spectral envelope was then converted into mel-cepstral coefficients using a recursion formula. The aperiodicity feature was also converted into average values for five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz. As a result, the feature vector consisted of 39 mel-cepstral coefficients including the zeroth coefficient, log F0, five-band aperiodicity values, and their delta and delta-delta coefficients. The total dimensionality was 138.

For the model training, phoneme segmentation was conducted manually for the training data. The topology of MRHSMM was a five-state left-to-right model with no skip. The output distribution in each state was modeled with a single Gaussian pdf, and covariance matrices were assumed to be diagonal. In the context clustering for parameter tying, decision trees were automatically constructed on the basis of the minimum description length (MDL) criterion (Shinoda and Watanabe, 2000). In the vibrato modeling, parameters of all Gaussian pdfs were tied as the result of the tree-based context clustering. A possible reason for the global parameter tying is that the vibrato is a segment-based feature, i.e., a set of parameters consisting of rate r in Hz and amplitude a in cent as described in Sect 4.3, and the number of training samples for the vibrato modeling was much smaller than that of the frame-based feature, e.g., mel-cepstrum. We used one-dimensional style space in Fig. 1(a). Style vectors, −1.0 and 1.0, were given for the training songs of child-like and adult-like styles, respectively. The contextual factors used for MRHSMMs are listed in Table 1. These contextual factors were automatically determined from MusicXML[2] data. Fig. 5 shows the histograms of pitch and pitch difference appearing in the training and test data. From the figure, the note pitch contexts of the training data well covered those of the test data.

The coverage of these contextual factors in the training data is 100 % for the test data. The window length of the moving average filter was set to 20 frames on the basis of a preliminary experimental result.

[2] MusicXML Definition, http://musicxml.org/.

Table 1
Contextual factors used in the experiments.

- {preceding, current, succeeding} phoneme
- absolute pitch (MIDI note number) of {preceding, current, succeeding} note
- pitch difference between {preceding, succeeding} note and current note
- relative pitch of {preceding, current, succeeding} note based on the key of the current measure
- length of {preceding, current, succeeding} note
- position of {preceding, current, succeeding} note in the current measure

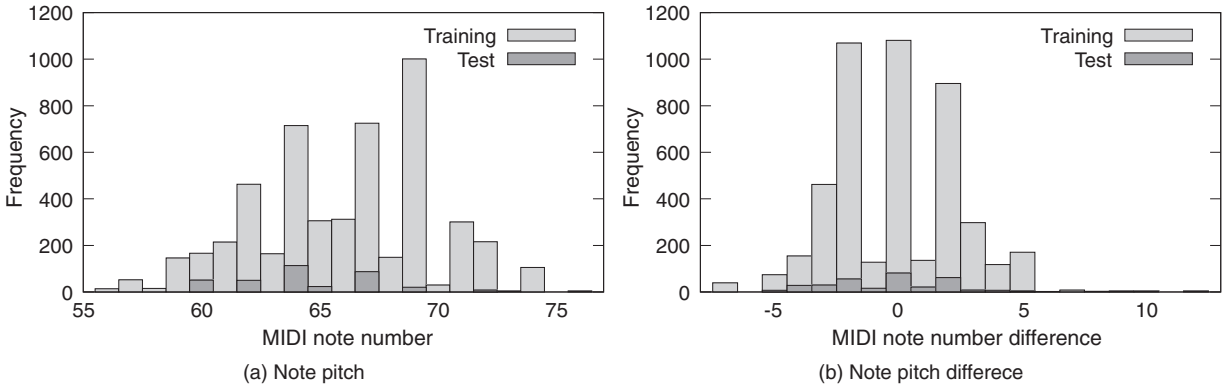


Fig. 5. Histograms of pitch of current notes and pitch difference between current and next notes for training and test data.

To improve the perceptual quality of the synthetic voice, the parameter generation algorithm considering global variance (Toda and Tokuda, 2007; Nose and Kobayashi, 2013) was applied to the acoustic features except for the F0. In the subjective evaluations, test songs were split into short phrases, each of which consists of four measures. Consequently, eleven different phrases were used. The mean duration of the eleven phrases was 4.9 s. The participants were twelve Japanese graduate/undergraduate students, and six phrases were randomly chosen for each participant.

5.2. Comparison of naturalness between style-dependent HSMM and MRHSMM

We compared naturalness of the synthetic singing voices of the proposed MRHSMM-based singing voice synthesis and the conventional style-dependent HSMM-based one. In this experiment, neither pitch adaptive training nor vibrato modeling was used. A single speaking style was chosen from the two singing styles and was used in model training of style-dependent HSMMs. The training data was the same as the target singing style data used in the training of MRHSMMs. We conducted a five-scale mean opinion score (MOS) test on naturalness for the synthetic singing phrases. In MRHSMM, the style vectors of -1.0 (corresponding to the child-like style) and 1.0 (corresponding to the adult-like style) were given in the synthesis process. Fig. 6 shows the average scores of all participants with confidence intervals of 95%.

From the figure, we found that the scores of style-dependent HSMM and MRHSMM were close and the proposed technique can synthesize singing voices with the similar performance to the conventional technique in terms of the naturalness. It is noted that the purpose of this study is not to outperform the conventional style-dependent HSMM with a fixed singing style but to add variability of singing style intensity to the conventional one with an acceptable naturalness level.

5.3. Effect of pitch adaptive training and vibrato modeling

We conducted a subjective test to evaluate the effect of introducing the proposed pitch adaptive training and unclear vibrato modeling into the baseline MRHSMM-based system, respectively.

Table 2 and 3 summarizes the following subjective experimental results of the paired-comparison tests on naturalness and XAB tests on similarity using synthetic singing voice samples with three different types of training methods. In the

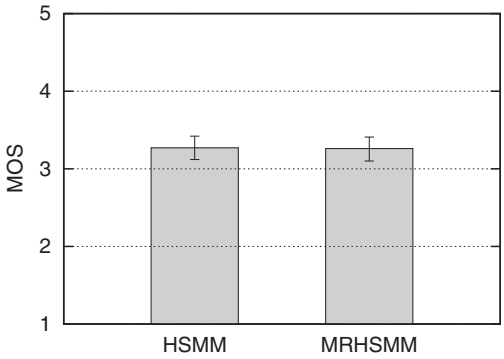


Fig. 6. Comparison of naturalness of synthetic singing voices of style-dependent HSMM and MRHSMM.

Table 2
Preference scores of naturalness on the use of pitch adaptive training (PAT) and vibrato modeling (VM) for baseline MRHSMM-based system (MR).

Preference score (%)			*p*-value
MR	MR + PAT	MR + PAT + VM	(*Z*-test)
51.5	48.5	–	0.728
–	36.4	**63.6**	0.002

Scores with statistically significant preference at $p < 0.05$ level are in the bold font.

Table 3
Preference scores of singing style similarity on the use of pitch adaptive training (PAT) and vibrato modeling (VM).

Preference score (%)			*p*-value
MR	MR + PAT	MR + PAT + VM	(*Z*-test)
47.9	52.1	–	0.617
–	26.4	**73.6**	<0.001

Scores with statistically significant preference at $p < 0.05$ level are in the bold font.

tables, the ratios of the samples that the participants chose by comparing two different methods, e.g., MR vs MR+PAT, are shown as preference scores (%). We tested the rate difference of the preference scores between the two methods using the statistical *Z*-test under the condition that the number of samples (result data) was relatively large. In the table, *p*-values of the *Z*-test are also shown. We set the value of the input style vectors to −1.0 and 1.0 in this experiment.

First, we evaluated the effect of pitch adaptive training by comparing naturalness of the synthetic singing voice samples generated with and without pitch adaptive training. From Table 2, we found that there was no statistically significant difference between the scores, which indicates that the pitch adaptive training is comparable to the baseline when the contextual coverage of note pitch and note pitch difference in the training data is sufficient for the test data as shown in Fig. 5. Note that these results are consistent with the previous study for the pitch adaptive training in HMM-based singing voice synthesis (Oura et al., 2012).

We also compared similarity of the synthetic singing voices to the original ones of the target singer's target singing style with and without the proposed pitch adaptive training using an XAB test. For A and B, we used the same synthetic samples used in the naturalness evaluation. The reference sample X was the vocoded singing voice of the original sample of the target singer and singing style. Participants first listened to X and then listened to A and B in random order. Participants were asked which sample sounded more similar to the reference sample X. Table 3 shows the result. We conducted a statistical *Z*-test and there was not a statistically significant difference between the two techniques ($p = 0.617$). The pitch adaptive training should be used for synthesizing the songs having wide or unknown contextual coverage.

Next, we evaluated the effect of the robust vibrato modeling. We compared naturalness of the synthetic singing voices with and without the vibrato modeling technique proposed in Section 4.3. In this experiment, we applied pitch

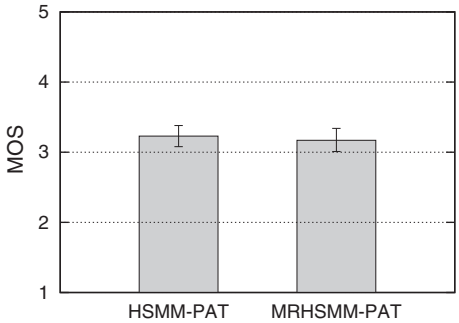


Fig. 7. Comparison of naturalness of synthetic singing voices with conventional pitch adaptive training for HSMM (HSMM-PAT) and proposed pitch adaptive training for MRHSMM (MRHSMM-PAT).

adaptive training for both cases. The results in Table 2 show that the proposed vibrato modeling technique significantly improved the naturalness of synthetic singing voices. We also compared similarity of the synthetic singing voices to the original ones of the target singer's target singing style with and without the vibrato modeling using an XAB test. For A and B, we used the same synthetic samples used in the naturalness evaluation. The reference sample X was the vocoded singing voice of the original sample of the target singer and singing style. Twelve participants first listened to X and then listened to A and B in random order. Participants were asked which sample sounded more similar to the reference sample X. Table 3 shows the result. We conducted a statistical Z-test and there was a statistically significant difference between the two techniques at a 5% significance level. The above results indicates that both naturalness and reproducibility were improved by introducing the proposed vibrato modeling to the singing voice synthesis.

5.4. Comparison of naturalness between conventional and proposed pitch adaptive modeling

Here, we compared the performance of the conventional pitch adaptive training for style-dependent HSMM (Oura et al., 2012) and the proposed one for MRHSMM in terms of the naturalness of synthetic singing voices. In this experiment, vibrato modeling techniques were not used. The training data for style-dependent HSMMs was the same as that in Section 5.2. The training data was the same as the target singing style data used in the training of MRHSMMs. A five-scale MOS test was used for evaluating the naturalness of the synthetic singing phrases. In MRHSMM, the style vectors of -1.0 (corresponding to the child-like style) and 1.0 (corresponding to the adult-like style) were given in the synthesis process. Fig. 7 shows the average scores of all participants with confidence intervals of 95%.

From the figure, there was not a significant difference between conventional and proposed pitch adaptive training techniques, which indicates that the proposed feature-space pitch adaptive training can be easily applied to MRHSMM with a small formula change as shown in Eq. (11) as well as keeping the similar naturalness of the synthetic singing voices to the conventional model-space pitch adaptive training for HSMM.

5.5. Comparison of naturalness between conventional and proposed vibrato modeling

In this experiment, we compared the performance of the conventional frame-based vibrato modeling (Oura et al., 2010) and the proposed segment-based vibrato modeling in terms of the naturalness of synthetic singing voices. We conducted a paired-comparison test. The vibrato parameters of the frame-based vibrato modeling were extracted for the long-tone segments as the similar way to (Oura et al., 2010). Two parameters, amplitude in cents and frequency in Hz, were used for training and synthesis. The ranges of the parameters were set to the same as those for the proposed technique. The training data of the style-dependent HSMMs was the same as the target singing style data used in the training of MRHSMMs. In MRHSMM, the style vectors of -1.0 and 1.0 were given in the synthesis process as described in Sect 5.2. Table 4 shows the result. In the table, p-values of statistical Z-test are also shown.

From the table, there was a statistically significant difference between the two techniques at a 5% significance level. This indicates that the frame-based vibrato modeling does not always work well in the case of using singing voice data with unclear vibrato expressions. In the frame-based vibrato modeling, the vibrato parameters are determined by

Table 4
Comparison of naturalness of synthetic singing voices with conventional and proposed vibrato modeling.

Preference score (%)		p-value
Conventional	Proposed	(Z-test)
41.7	**58.3**	0.046

Scores with statistically significant preference at $p < 0.05$ level are in the bold font.

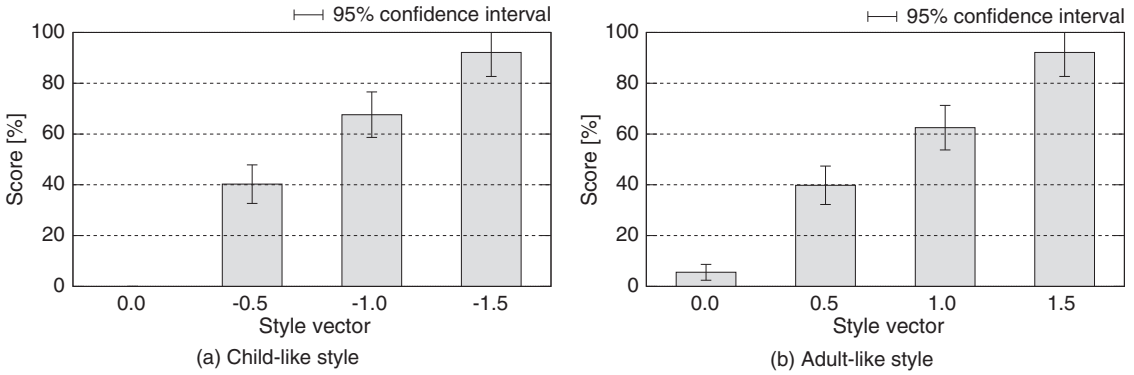


Fig. 8. Perceived variation of the intensity of singing style expressivity when changing a style vector.

periodic peaks of F0 contours, however in the unclear vibrato case, the peaks and periodicity are both not clear and amplitude is much smaller than the case of clear vibrato as shown in Fig. 4.

5.6. Controllability of singing style expressivity

We examined whether users can intuitively control the expressivity of singing styles. We synthesized singing voice samples by changing the input style vector from 0.0 to −1.5 and from 0.0 to 1.5 in child-like and adult-like styles, respectively, with an increment/decrement of 0.5. Twelve participants listened to all possible pairs from the four types of synthetic samples in both singing styles and judged which sample had a stronger expressivity of the singing style without any reference samples. Fig. 8 shows the average scores of all participants with confidence intervals of 95%.

In the figure, the scores become larger in accordance with the intensity of the perceived expressivity of the singing styles. We conducted one-way ANOVA and there is a statistically significant difference among the samples synthesized with four different style vectors at a 5% significance level ($p \ll 0.05$ in both styles), which indicates that participants could perceive the difference in the style expressivity appearing in the synthetic singing voices when the given style vector was changed. Note that the score for the style vector of 0.0 in the child-like style was zero. When comparing two singing styles, the difference in the expressivity was clearer in the child-like style than the adult-like style, which indicates that the perceived difference depends on the singing styles.

5.7. Naturalness of synthetic singing voices

Finally, we evaluated the naturalness of the synthetic singing voices when changing the input style vector. We conducted a MOS test on naturalness for the synthetic singing phrases given a style vector from −1.5 to 1.5 with an increment of 0.5. Twelve participants listened to the samples and rated the naturalness on a five-point scale: “1” for bad to “5” for excellent. Fig. 9 shows the results with confidence intervals of 95%.

From the results, the scores of the naturalness of the synthetic singing voices are close to four (good) when the value of the style vector was set between −1.0 and 1.0. This means that we can control the expressivity of singing styles without degradation of the naturalness in the case where we synthesize the singing voices having intermediate expressivity of the representative singing styles. Although there was a slight degradation of the naturalness when the

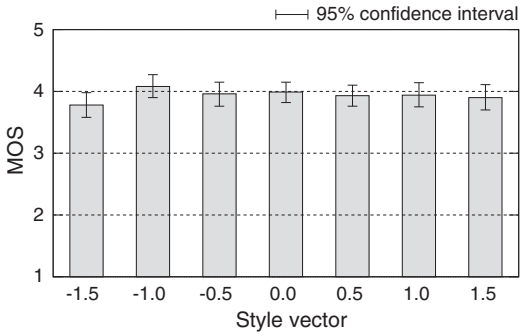


Fig. 9. Naturalness of synthetic singing voices for different input style vectors.

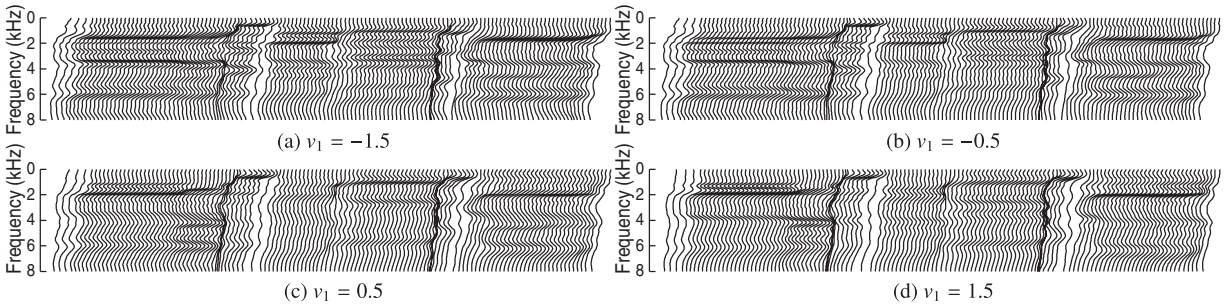


Fig. 10. Example of running spectra for different input style vectors.

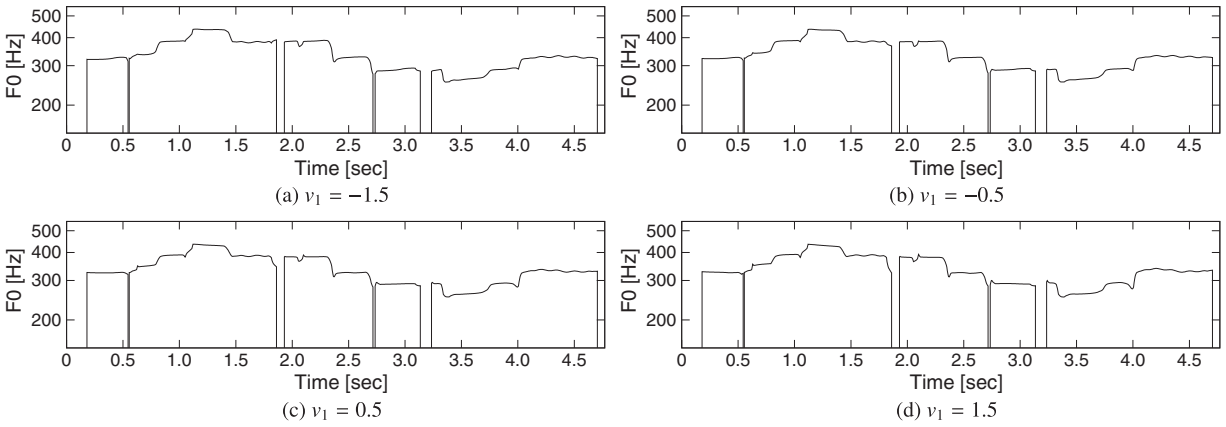


Fig. 11. Example of generated F0 contours for different style vectors.

value of the style vector was set to −1.5 and 1.5, which means that the expressivities of singing styles were emphasized, we concluded that the style expressivity of synthetic singing voices was well controlled with acceptable quality.

Figs. 10 and 11 show examples of running spectra and F0 contours given different style vectors, respectively. We can see that the spectral envelopes and F0 contours vary depending on the value of the input style vector and the vibrato expression is reproduced in the long-tone segments compared to the case of no vibrato modeling in Fig. 3(c).

6. Conclusions

In this paper, we proposed a novel singing voice synthesis technique that enables users to control singing styles intuitively and continuously. The technique is based on MRHSMM, and feature-space pitch adaptive training was proposed for the MRHSMM training to model pitch precisely in the singing style control. We also proposed a robust

vibrato modeling technique that can be used even for singing voices with unclear vibrato expressions. The subjective evaluation results showed that the naturalness of the synthetic singing voices with representative styles were comparable between the conventional style-dependent HSMM and the proposed MRHSMM. The proposed pitch adaptive training in the feature space can be applied easily to the MRHSMM and had the similar performance to the conventional pitch adaptive training in the model space for the HSMM. In addition, the proposed segment-based vibrato modeling was shown to be more effective on naturalness than the conventional frame-based vibrato modeling. Finally, we showed that users can control the style expressivity of synthetic singing voices with acceptable quality by changing the input style vector. Future work will focus on modeling and controlling local singing style variations that often appear in singing voices of professional singers. The proposed technique also needs to be evaluated with a larger amount of singing voice data including more singing style variations with multiple dimensional style vectors.

Acknowledgments

The authors thank Dr. Shinji Sako of Nagoya Institute of Technology for the use of MusicXML data, a part of singing voice data, MIDI data, and lyric data. Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071.

References

Bonada, J., Loscos, A., Kenmochi, H., 2003. Sample-based singing voice synthesizer by spectral concatenation. In: Proc. Stockholm Music Acoustics Conference (SMAC 03).

Cook, P.R., 1993. SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. Comput. Music J. 17 (1), 30–44.

Cook, P.R., 1996. Singing voice synthesis: history, current work, and future directions. Comput. Music J. 20 (3), 38–46.

Imai, S., 1983. Cepstral analysis synthesis on the mel frequency scale. In: Proc. ICASSP 83, pp. 93–96.

Kaneko, K., Kanehiro, F., Morisawa, M., Miura, K., Nakaoka, S., Kajita, S., 2009. Cybernetic human HRP-4C. In: Proc. 9th IEEE-RAS International Conference on Humanoid Robots, pp. 7–14.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27 (3–4), 187–207.

Kenmochi, H., 2012. Singing synthesis as a new musical instrument. In: ICASSP 2012, pp. 5385–5388.

Kenmochi, H., Ohshita, H., 2007. Vocaloid-commercial singing synthesizer based on sample concatenation. In: Proc. INTERSPEECH 2007, pp. 4010–4011.

Macon, M.W., Jensen-Link, L., Oliverio, J., Clements, M.A., George, E.B., 1997. A singing voice synthesis system based on sinusoidal modeling. In: Proc. ICASSP-97, pp. 435–438.

Maher, R., Beauchamp, J., 1990. An investigation of vocal vibrato for synthesis. Appl. Acoust. 30 (2), 219–245.

Nakaoka, S., Kanehiro, F., Miura, K., Morisawa, M., Fujiwara, K., Kaneko, K., Kajita, S., Hirukawa, H., 2009. Creating facial motions of Cybernetic Human HRP-4C. In: Proc. 9th IEEE-RAS International Conference on Humanoid Robots, pp. 561–567.

Niwase, N., Yamagishi, J., Kobayashi, T., 2005. Human walking motion synthesis with desired pace and stride length based on HSMM. IEICE Trans. Inf. Syst. E88-D (11), 2492–2499.

Nose, T., Kanemoto, M., Koriyama, T., Kobayashi, T., 2013. A style control technique for singing voice synthesis based on multiple-regression HSMM. In: Proc. INTERSPEECH 2013, pp. 378–382.

Nose, T., Kobayashi, T., 2011. Recent development of HMM-based expressive speech synthesis and its applications. In: Proc. APSIPA ASC 2011, http://www.apsipa.org/proceedings_2011/pdf/APSIPA189.pdf

Nose, T., Kobayashi, T., 2013. An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. Speech Commun. 55 (2), 347–357.

Nose, T., Yamagishi, J., Masuko, T., Kobayashi, T., 2007. A style control technique for HMM-based expressive speech synthesis. IEICE Trans. Inf. Syst. E90-D (9), 1406–1413.

Oura, K., Mase, A., Nankaku, Y., Tokuda, K., 2012. Pitch adaptive training for HMM-based singing voice synthesis. In: Proc. ICASSP 2012, pp. 5377–5380.

Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., Tokuda, K., 2010. Recent development of the HMM-based singing voice synthesis system-Sinsy. In: Proc. 7th ISCA Workshop on Speech Synthesis (SSW7).

Rodet, X., 2002. Synthesis and processing of the singing voice. In: Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), pp. 15–31.

Saino, K., Tachibana, M., Kenmochi, H., 2010. An HMM-based singing style modeling system for singing voice synthesizers. In: Proc. 7th ISCA Workshop on Speech Synthesis (SSW7).

Saino, K., Tachibana, M., Kenmochi, H., 2010. A singing style modeling system for singing voice synthesizers. In: INTERSPEECH, pp. 2894–2897.

Saino, K., Zen, H., Nankaku, Y., Lee, A., Tokuda, K., 2006. An HMM-based singing voice synthesis system. In: Proc. INTERSPEECH 2006-ICSLP, pp. 1141–1144.

Sako, S., Miyajima, C., Tokuda, T., Kitamura, T., 2004. A singing voice synthesis system based on hidden Markov model. Trans. Inform. Process. Soc. Jpn. (Japanese Edition), 719–727.

Shinoda, K., Watanabe, T., 2000. MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Jpn. (E) 21 (2), 79–86.

Sundberg, J., 2006. The KTH synthesis of singing. Adv. Cogn. Psychol. 2 (2–3), 131–143.

Sung, J.S., Hong, D.H., Kang, S.J., Kim, N.S., 2011. Factored MLLR adaptation for singing voice generation. In: Proc. INTERSPEECH 2011, pp. 2789–2792.

Tachibana, M., Nakaoka, S., Kenmochi, H., 2010. A singing robot realized by a collaboration of VOCALOID and Cybernetic Human HRP-4C. In: Proc. InterSinging.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. E88-D (11), 2484–2491.

Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T., 2001. Text-to-speech synthesis with arbitrary speaker's voice from average voice. In: Proc. EUROSPEECH, pp. 345–348.

Toda, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inf. Syst. E90-D (5), 816–824.

Tokuda, K., Kobayashi, T., Imai, S., 1995. Speech parameter generation from HMM using dynamic features. In: Proc. ICASSP-95, pp. 660–663.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In: Proc. ICASSP-99, pp. 229–232.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Lang. Process. 17 (1), 66–83.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. EUROSPEECH, pp. 2347–2350.

Young, S.J., Odell, J., Woodland, P.C., 1994. Tree-based state tying for high accuracy acoustic modeling. In: Proc. ARPA Human Language Technology Workshop, pp. 307–312.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2007. A hidden semi-Markov model-based speech synthesis system. IEICE Trans. Inf. Syst. E90-D (5), 825–834.