A Singing Style Modeling System for Singing Voice Synthesizers

Keijiro Saino, Makoto Tachibana, Hideki Kenmochi

Corporate Research & Development Center, Yamaha Corporation, Japan
{keijiro_saino,makoto_tachibana,hideki_kenmochi}@gmx.yamaha.com

Abstract

This paper describes a method of modeling singing styles by a statistical method. In this system, singing expression parameters consisting of melody and dynamics which are derived from F0 and power are modeled by context-dependent Hidden Markov Models (HMMs). A modeling method of the parameters are optimized for dealing with them. Since parameters we focus on are essential but general ones for singing synthesizers, generated parameters from the trained models may be possible to be applied to many of them. In the experiment, we trained singing style models by using singing recording with much expressive style, then parameters were generated for songs not included in training data and actually applied to our singing synthesizer VOCALOID. As a result, the style was well perceived in the synthesized sound with good synthetic quality.

Index Terms: singing voice synthesis, singing style, HMM

1. Introduction

Singing voice synthesis is becoming more and more popular these days. Many people are now enjoying producing their original music by synthesizing vocal tracks by using existing singing synthesizers. However, synthesized sound provided by default settings from such vocal synthesizers usually has little singing expression. Therefore, for example in the case of our product VOCALOID [1], users adjust several controlling parameters which affect synthesis, such as pitch bend or dynamics in order to produce a sound having some particular style they would like. At this step is not so easy, many users take much time and effort for this parameter-adjustment. VocalListener [2], an automatic parameter estimation system using one’s singing voice as a reference, was proposed as a technique which can solve such problems, but the recording step required in the system may sometimes call for much effort.

To overcome these difficulties, we propose a method for modeling singing styles from a human’s singing, aiming at fully automatic parameter generation. It is reported that pitch is an effective parameter for singing style identification in [3]. Besides, VocalListener succeeded to synthesize a singing voice mimicking a user’s singing style by estimating control parameters of pitch and dynamics (power) so that the fundamental frequency (F0) and power of the synthesized sound approximate those extracted from one’s singing. Considering such results, we deal with pitch and dynamics as parameters which are feasible for expressive singing voice synthesis. In this system, less phoneme-dependent parameters are provided from F0 and power by removing their dependency on phoneme. These parameters are henceforth called “melody parameter” and “dynamics parameter,” respectively. Melody parameter can be separated into base melody parameter and vibrato parameters which include parameters of its “shape” and “rate.” We call these kinds of parameters “singing expression parameters.”

A singing voice synthesis system which models a singer’s characteristics by statistical means was already proposed in [4]. Though our modeling method is similar to that, we are focusing on singing expression parameters in particular. We therefore propose a modeling technique optimized for dealing with them. For example, they are not trained by phoneme HMMs but note HMMs assuming that the training parameters are less dependent on phoneme. Moreover, notes are supposed to be divided into up to three regions of different types according to their expressive behavior. We also propose to treat melody parameter as a relative pitch parameter against corresponding notes.

The rest of this paper is summarized as follows: Section 2 describes features we model and the modeling method itself. Results of subjective evaluation experiments are shown in Section 3. Finally, we conclude our proposal in Section 4.

2. Singing Style Modeling

Figure 1 shows the overview of the proposed system. At first, singing expression parameters are constructed from acoustic features such as F0 or power extracted from singing voice recording. They are modeled by context-dependent HMMs. The model parameters are estimated according to context-dependent labels which can be converted from musical score information excluding lyrics. Since we are dealing with features which represent “singing expression” directly, the modeling process is optimized for that.

2.1. Singing Expression Parameters

2.1.1. Melody Parameter and Dynamics Parameter

As mentioned in Section 1, we are dealing with pitch and dynamics as singing expression features for the singing style modeling. However, F0 and power should not be used directly because it may include variation which does not seem to be caused by singing expression, but phoneme. A pitch valley caused by a phoneme ‘z’ or unvoiced signal caused by a phoneme ‘s’ are examples that can be seen in Figure 2. Assuming that the variation is independent on singing expression, it is removed from acoustic features first. In this system, the removal process is done just by linear interpolation.
as seen in the bottom of Figure 2 assuming that such variation appears in every consonant section. Therefore, singing data for training have to be aligned by phonemes before the process. In this paper, it was done manually.

2.1.2. Vibrato Parameters

Next, vibrato parameters are extracted from the melody parameter. Vibrato is known as a periodic fluctuation of pitch and power, which is an important technique for singing expression. To make matters simple, in this paper, vibrato is considered to be constructed by only a fluctuation of pitch. Here we propose a novel vibrato feature which represents not only the extent of the fluctuation, but also its shape. Rate parameter is also extracted simultaneously. These are obtained by the following process (see Figure 3.)

At first, melody parameter is segmented by every section subjectively labeled as “sustained.” In Figure 3 is an example of melody parameter plot within one of the sections. Local peaks of the fluctuation within a low-passed melody parameter (solid line in [Cent]) are detected, then the vibrato phase is determined so that every local peak has a value of \( \pi \). As a result, dots in [Cent] are plotted. For each of the other frame phase parameter is given by linear interpolation and extrapolation. The solid line in [Cent] is the result of that. In [Cent], finally, an example of extracting vibrato shape parameter and rate parameter for the t-th frame is shown. One cycle of fluctuation having width of \( \hat{t}(t) \) frames is picked up by referring [Cent] . Vibrato phase parameter is obtained by applying an FFT to the one-cycle signal after both generalizations for phase and period. Phase of the one-cycle is modified so that it begins from \( 2n\pi \) [rad.] where \( n \) is an integer. Period is generalized by linear time stretch so as to be \( K \) frames, where \( K \) is a constant number which is set to 64 in our case. Since full dimensionality of this feature is high, only the lower \( d \) dimensions (excluding the 0-th) can be adopted as the parameter. At the same time, rate parameter is obtained by dividing \( K \) by \( \hat{t}(t) \). The 0-th dimension of the spectrum of the vibrato shape is adopted as a base melody parameter of the frame.

2.2. Training of Context-Dependent HMMs

Singing expression parameters mentioned in 2.1 are modeled by context-dependent HMMs. Context-dependent state duration models are also estimated simultaneously. MIDI note number (pitch in semitone) and duration (in 50msec units) are taken into account as contextual factors here.

2.2.1. Behavior-based Modeling

As the melody parameter is especially essential for modeling singing styles, and it normally behaves according to the pitch of given musical notes, our basic idea is modeling the parameters by note HMMs.

In addition, we propose one more idea: modeling by behavior-type HMMs. A strategy of this modeling is summarized in Figure 4. We suppose behavior of singing expression parameters within each note can be divided into up to three regions which represent types of behavior; “beginning,” “sustained,” and “ending.” In order to model such behavior more appropriately, singing expression parameters are modeled by behavior-type HMMs. Before this modeling starts, each type of behavior has to be segmented manually. Each note is defined to include up to 3 types of melodic behaviors, and the order of them is unchangeable. As a result, a pattern of behavior in each note can be classified into 7 possible patterns \{B, S, E, BS, BE, SE, BSE\}. This subjectively given behavior-type information is appended to the contextual factors described in the beginning of 2.2 and must be a critical help for dividing leaf nodes during decision-tree-based context clustering [5].

With this method of the modeling, behavior-pattern models are introduced to model appearance patterns of each type of behavior in every note. Context-dependent labels constructed by the same contextual factors as those described in the beginning of 2.2 are assigned to the observed patterns. Then they are clustered by a decision tree in the same manner used in [5]. At last, decision-tree-clustered context-dependent behavior-pattern models which have discrete probability distributions are obtained.

In addition, singing expression parameters observed within each note will become possible to be modeled with different numbers of states according to the pattern of its behavior which is actually observed.

2.2.2. Relative Melody Parameter

In the previous work of HMM-based singing voice synthesis [4], absolute F0 parameter is modeled by HMMs as a pitch parameter. We, on the other hand, suppose that relative pitch parameter is preferred for the modeling of behavior of the melody parameter. For example, if a singer sings songs always in a slightly lower pitch compared to the given note, that kind of style should be modeled more appropriately by using
relative pitch parameter rather than by using absolute one. Therefore, we subtract a frequency value given by a note number from the observed melody parameter in a logarithmic scale (cent.) Not only pitch behavior is expected to be modeled more appropriately, but any arbitrary pitches of notes can be produced by using the relative feature even if such notes do not appear in training songs.

2.2.3. An Individual Training of each HMM

In conventional methods of modeling speech by HMMs [6], a feature of every possible frame (time) in an utterance is used for the model parameter estimation. In our case, on the other hand, it may be more possible to misestimate state durations with this way of training. An example of that is like this: only one state might have high occupancy in case of a singer keeps pitch stable when he sings for consecutive notes in different contexts. For the purpose of avoiding this aspect, HMMs for singing expression parameters are trained individually, not being concatenated.

3. Experiment

3.1. Parameter Generation

Before the synthesis experiment, a process of parameter generation is described first.

At first, an arbitrarily given musical score is converted to a context-dependent label sequence. Then a song HMM is constructed by concatenating corresponding HMMs according to these labels. After state durations are determined by state duration models, singing expression parameters are generated from the HMM by using a parameter generation algorithm [7]. Since mean parameters of Gaussian for base melody have to represent absolute pitch during parameter generation, frequency values of corresponding notes in a logarithmic scale are added to them if relative melody parameters were used for the modeling.

After the parameter generation, vibrato is reproduced by the vibrato parameters, then it is added to the generated base melody parameter. The \( t \)-th frame of reproduced melody parameter, \( m(t) \), is derived from the following equation using the \( t \)-th frame of the generated base melody parameter \( m(t) \):

\[
m(t) = m(t) + \text{IDFT}
\left( s(t), \mod \left( \sum_{i=1}^{n} v_{\tau_k} s(\tau) K \right) \right) \quad \forall \theta \in \mathbb{R}{\quad (1)}
\]

In the above equation, \( v_i(t) \) denotes the generated vibrato rate parameter of the \( t \)-th frame, \( l(t) \) denotes the last frame before the \( t \)-th frame where \( v_i(t) \) has a discrete symbol of “non-vibrato,” \( s(t) \) denotes the generated vibrato shape parameter of the \( t \)-th frame having \( K \) dimensionality where only \( d \) dimensions (bins) excluding a DC component have non-zero values. IDFT\( (s(t)) \) provides a result of time \( f \) of the inverse DFT of spectrum \( s \), \( \mod(a,b) \) provides an integer what the remainder of the division of \( a \) by \( b \) is rounded to. The first item within \( \mod() \) in the equation (1) represent the progress of vibrato phase.

3.2. Experimental Conditions

Here we describe experimental conditions of the model training. Before the training, timings of MIDI events (note-on and note-off) of training songs, were revised according to phoneme alignment which was manually done (note-on timing was set to onset timing of the first vowel or syllabic nasal included in each note, note-off timing was set to offset timing of the last vowel or syllabic nasal.) Errors of F0 extraction were also fixed manually.

A summary of the singing voice database used for the experiment is shown in Table 1. We recorded a singing voice with a deep bendy style sung by a male professional singer. Five Japanese children’s songs were recorded for that style. The F0 contour seen in the top of Figure 2 comes from one of these recordings. The feature vector is composed by associating singing expression parameters as shown in Table 2. Feature vectors consist of not only static parameters, but also \( \Delta \) and \( \Delta^2 \) parameters. Since base melody and vibrato rate may have non-dimensional observation, they are modeled by MSD-
Table 5. Synthesis methods for subjective listening tests.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pitch parameter</th>
<th>Model unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>V (VOCALOID default)</td>
<td>(VOCALOID default)</td>
<td></td>
</tr>
<tr>
<td>Method A</td>
<td>Absolute</td>
<td>Note</td>
</tr>
<tr>
<td>Method B</td>
<td>Absolute</td>
<td>Behavior-type</td>
</tr>
<tr>
<td>Method C</td>
<td>Relative</td>
<td>Note</td>
</tr>
<tr>
<td>Method D</td>
<td>Relative</td>
<td>Behavior-type</td>
</tr>
</tbody>
</table>

HMMs (multi-space probability distribution HMMs) [8]. The process of making melody parameter relative was done only for the static feature. Table 3 shows topologies of two types of HMMs we used. The decision-tree based context clustering was applied to base melody, vibrato shape, vibrato rate, dynamics, and state duration, respectively. Appearance patterns of behavior type were also clustered when behavior-type HMMs were used. We used the MDL criterion to stop tree growth for all clustering process in the experiment. Table 4 shows an extraction condition of F0 and power.

3.3. Singing Synthesis Experiment with VOCALOID

From trained HMMs, we generated singing expression parameters for 5 Japanese children’s songs excluded from the training set. Then they were applied to VOCALOID for the experiments. Receiving pitch parameter, VOCALOID synthesizes a waveform with a suitable sample unit in its database shifting its pitch to given F0, but an unvoiced frame is kept unvoiced. The t-th frame of the generated dynamics parameter $D(t)$ was converted to multiplication factor $a(t)$ for the t-th frame by the following equation:

$$a(t) = 10^{\frac{D(t) - D_{\text{MAX}}}{20}},$$

where $D_{\text{MAX}}$ denotes the maximum value of dynamics parameter among all observed frames in the training data. $a(t)$ was multiplied to the amplitude of the t-th frame of the waveform synthesized by VOCALOID. In other words, $a(t)$ is a coefficient of amplification by $-(D_{\text{MAX}} - D(t))$ dB.

At last, subjective listening tests were conducted. Four phrases which have durations of about 10 seconds each are picked up from every synthesized song, then 10 phrases were randomly chosen and given to 10 subjects. As a result, every method had opportunities of being scored 100 times in total. Five synthesis methods that were tried are shown in Table 5.

First, style perception of synthesized sounds was scored on Mean Opinion Score (MOS) with a scale from 1 (poor) to 5 (good) to see if models work as they are desired to. A high score was rated when a subject perceived some singing style much in the synthesized sound (the singing style used for the training was not informed to subjects.) From the result of the evaluation shown in Figure 6, it can be seen that a singing style was successfully perceived in the result of the proposed method more than that of the VOCALOID default.

Next, naturalness of the synthesized sound was also evaluated on MOS in order to compare effectiveness among proposed methods. The result shown in Figure 7 shows that introducing relative pitch training and behavior-type models are both effective for a natural-sounding synthesis.

4. Conclusions

We proposed a statistic-based singing style modeling method in this paper. Assuming singing expression to be phoneme-independent, singing-expression parameters were derived from acoustic features by removing phoneme-dependent variations from them. A behavior-type-based modeling was preferred when singing expression parameters were modeled. The experimental results show the proposed system can generate parameters which enable singing synthesizers to provide a sound with an expressive singing style.

5. References