**LETTER**

**Building a Measurement Model for Simulating Naturalness of Vibrato Based on Subjective Evaluation**

**Takahiro MIYAZAKI**\(^1\), Nonmember and **Masanori MORISE**\(^{1,†,††}\), Member

**SUMMARY** This work introduces a measurement model to estimate the naturalness of vibrato. We carried out a subjective evaluation using a mean opinion score (MOS). We then built a measurement model by using two-dimensional Gaussian functions. We found that three Gaussian functions can measure naturalness with an error of 4.0%.

**key words:** singing, vibrato, measurement model, subjective evaluation

1. Introduction

Studies on singing voice analysis and synthesis have been conducted [1] for various purposes relating to singing synthesis applications. A complexity arises in that the vibrato is different from a speech in terms of the acoustic features of the pitch and timbre. Pitch difference is observed as a fluctuation of the fundamental frequency \(F_0\). Fluctuation in \(F_0\) includes intentional expressions [2] such as vibrato. Overshoot and fine fluctuation [3] are also observed. Vibrato, a common singing expression, is defined as a singing technique in which the \(F_0\) contour vibrates quasi-periodically. Its acoustic information is known to affect singing ability [4] and perception of singing voice [5]. Vibrato is commonly used in VOCALOID and other synthesis systems [5]. Vibrato is widely used as the main parameter in singing voice analysis and synthesis.

We focus on the vibrato as an ornament singing expressions typically used in singing synthesizers. In the generation of vibrato using a vocal synthesizer, the user listens to the synthesized result and evaluates the naturalness. The user is required to repeat this as many times as necessary to generate a satisfactory vibrato expression. To reduce the manipulation cost, we attempt to build a measurement model of the vibrato naturalness. If the measurement can be performed with high accuracy without reproducing singing voice, the user would be free from the bothersome manipulation. We attempted to estimate the naturalness of vibrato from the vibrato rate and vibrato extent.

In this study, we first carried out a subjective evaluation utilizing vibrato and non-vibrato singing voices based on a mean opinion score (MOS). After that, we constructed a measurement model by performing non-linear fitting on the experimental results using two-dimensional Gaussian functions.

2. Related Works

Vibrato parameters have been utilized in several studies on singing voice analysis and synthesis. Migita et al. pointed out that vibrato parameters are effective for expressing the individualities of singing voices [6]. The vibrato database used in their study contains the mimicked and normal singing voice of four professional singers. They found that when a singer imitates another singer, the vibrato parameters are consciously controlled. A study on SingBySpeaking [5], a singing voice synthesis system that converts speech into singing, reported a vibrato \(F_0\) control model using vibrato parameters.

Commercial singing synthesizers such as VOCALOID, Auto-Tune, and Melodyne come with a function for vibrato expression. Artistically expressed vibrato provides listeners with a unique emotional experience, and users have attempted to make an attractive vibrato by using these applications.

2.1 Related Works on Vibrato

The vibrato rate and vibrato extent are calculated by [7]

\[
\frac{1}{\text{rate}} = \frac{1}{N} \sum_{n=1}^{N} R_n, \tag{1}
\]

\[
\text{extent} = \frac{1}{2N} \sum_{n=1}^{N} E_n, \tag{2}
\]

where \(N\) represents the total number of peaks and dips extracted from the \(F_0\) contour. Vibrato extent is calculated from \(F_0\) converted to cent as

\[
f_{\text{cent}} = 1200 \log_2 \left( \frac{f_{\text{Hz}}}{f_c} \right) + 4800. \tag{3}
\]

Figure 1 shows an example of \(R_n\) [s] and \(E_n\) [cent] extracted from the vibration section of the \(F_0\) contour. The horizontal and vertical axes represent the time and frequency, respectively. Equation (3) provides a cent value of 4,800 for the frequency \(f_c\) of C4 (around 261.62 Hz).
the subjective evaluation. However, another study ters including such temporal changes should be included in perfect vibrato measurement model is required, all paramet-
ralness and vibrato parameters. After that, we carry out a
result within a reasonable amount of time.
As discussed above, studies on analyzing and synthesizing
the vibrato have been carried out in several aspects. In this
study, we focus on the auditory perception in vibrato and
on the relationship between vibrato parameters and auditory
perception. Our objective is to build a measurement model
that estimates the naturalness of vibrato based on a subject-
ive evaluation.
Previous studies [9, 10] have pointed out that the vi-
brato rate and the extent vary temporally. In cases where a
perfect vibrato measurement model is required, all para-
ters including such temporal changes should be included in
the subjective evaluation. However, another study [11] has
suggested that the use of constant rate and extent would not
affect the naturalness significantly. We therefore use only
the temporally constant vibrato parameters so as to complete
the evaluation within a reasonable amount of time.
Our intent is to clarify the relationship between natu-
ralness and vibrato parameters. After that, we carry out a
non-linear fitting to build a model.

3. Subjective Evaluation

A subjective evaluation was carried out to determine the re-
lationship between vibrato parameters and perceived natu-
ralness. In this evaluation, we used singing voices with vi-
brato synthesized by VOCALOID. Since vibrato is a singing
expression in F0, we can design it by using a vocoder [12].
However, the vibrato of a human singing voice changes in
terms of not only the F0 but also the power and timbre [13].
We used VOCALOID in this evaluation because it can con-
trol the vibrato parameters and synthesize natural vibrato.

3.1 Generation of Singing Voice with Vibrato

We used 390 singing voices in the evaluation, including six
that were non-vibrato. These singing voices were single sus-
tained tones of 2 seconds consisting of the vowel /a/. The in-
formation in the singing voice was presented to the subjects
before the experiment. Table 1 represents vibrato para-
ters used in the evaluation, and a total of 384 vibrato singing
voices (8 vibrato rates × 8 vibrato extents × 3 pitches × 2
singers) were used in the evaluation. The sampling fre-
quency and quantization bit were 44.1 kHz and 16 bits, re-
spectively.
The ratio between the time length of the vibrato section
and that of the singing section is defined as the vibrato dura-
tion. With VOCALOID, we can set the vibrato duration in
one note as a percentage from 0 to 100, and the range from
0 to 127 is used as the specific parameters of the vibrato rate
and extent.
The vibrato duration was based on prior analysis re-
results of a singing database [6]. Since specific parameters
of VOCALOID, which indicate the vibrato rate and extent,
are different from the physical quantity, we first generated a
singing voice with the specific parameters’ maximum value.
After that, we calculated the maximum physical quantity re-
lated to the parameter based on the F0 contour estimated
from the generated singing voice, and the upper limit was
determined based on the calculated result. In the same way,
the lower limit of the vibrato rate was determined. In the
lower limit of the vibrato extent, we divided the determined
maximum pitch into equal parts. The vibrato rate and extent
were set to a total of 8 patterns, as shown in Table 1.
In VOCALOID, the vibrato rate and extent are gen-
erated using specific parameters that have, no linear cor-
respondence to actual physical quantities. Therefore, to
exploit these physical quantities, we obtained the physical
quantities of vibrato rate and extent using the method de-
scribed in Sect. 2.1. Harvest [14] was used to extract the F0
contour. We looked for specific parameters of VOCALOID
that would have an error of less than 5% for the target phys-
ical quantity and judged it to be a vibrato corresponding to
the target physical quantity. We conducted a preliminary
investigation of perceptual differences between each target
parameter and parameter with 5% difference and found no
significant difference.

3.2 Experiment Condition

The evaluation conditions are listed in Table 2. We used
a sound-proof room with the A-weighted SPL of 18 dB.
Twenty individuals with normal hearing ability participated
in the evaluation. The sound stimuli were reproduced
through a set of headphones (SENHEISER HD650).
Among the various singer instances of VOCALOID, we
used Kaito V3 and Hatsune Miku V4X. Three pitches in
one octave were used to measure the dependence of pitch on
naturalness. Three pitches (C3 (around 261.6 Hz), F#4, and
C4) were used in Hatsune Miku V4X. In Kaito V3, we used
C2, F#2, and C3 based on the cover ranges of each singer.

| Table 1 | Vibrato parameters used in the evaluation. |
|--------|--------------------------------------------|
| Vibrato rate | 2.4, 3.4, 4.4, 5.4, 6.4, 7.4, 8.4, 9.4 Hz |
| Vibrato extent | 15, 30, 45, 60, 75, 90, 105, 120 cent |
| Vibrato duration | 1.44 s (72 %) |

Fig. 1  Example of an F0 contour and the definition of parameters used to calculate the vibrato rate and extent. This figure was cited from [8].
Table 2 Evaluation conditions.

| Parameter          | Description                              |
|--------------------|------------------------------------------|
| Number of subjects | 20 persons (17 men and 3 women)          |
| Environment        | Sound-proof room                         |
| Background noise   | 18 dB (A-weighted SPL)                   |
| Headphones         | SENNHEISER HD650                         |
| Audio I/O          | Roland QUAD-CAPTURE                      |
| Sampling           | 44.1 kHz / 16 bit                        |

This is the definition in VOCALOID and different from the international pitch notation.

The subjects first evaluated a singing voice for perceptible pitch fluctuations. For those that were perceptible, they evaluated in terms of naturalness of vibrato for that voice by using the MOS evaluation on a scale of one (bad) to five (excellent). This procedure was performed for all singing voices. To evaluate naturalness of vibrato, it is necessary to perceive pitch fluctuations similar to the vibrato.

### 3.3 Results of Subjective Evaluation

Figure 2 shows a heatmap of vibrato naturalness. The most natural vibrato parameters were 5.4 Hz and 60 cent in most conditions. This result suggests a certain commonality between perceptual parameters in the naturalness of vibrato and parameter ranges in vibrato rate and extent.

### 4. Building a Measurement Model for Naturalness

We built the measurement model for the naturalness of vibrato on the basis of the subjective evaluation results. We used two-dimensional Gaussian functions to build the model. The input parameters are the MOS values for the vibrato rate and extent. Non-linear fitting was performed to model the naturalness of the vibrato.

#### 4.1 Non-Linear Fitting by Using Two-Dimensional Gaussian Functions

We built the model using two-dimensional Gaussian functions. The 64 data items we obtained from the heatmap were used. Gaussian function was given by

$$
\begin{align*}
f(x, y) &= \frac{\alpha}{2\pi\sigma_x\sigma_y \sqrt{1 - \rho^2}} \times \\
&\exp\left( -\frac{1}{2(1 - \rho^2)} \frac{(x - \mu_X)^2}{\sigma_x^2} + \frac{(y - \mu_Y)^2}{\sigma_y^2} - \frac{2\rho (x - \mu_X)(y - \mu_Y)}{\sigma_x\sigma_y} \right),
\end{align*}
$$

where \( x \) and \( y \) represent vibrato rate and extent, respectively. \( \mu, \sigma, \) and \( \rho \) represent average, standard deviation, and correlation coefficient, respectively. \( \alpha \) is a parameter to control the amplitude of the function. Since \( \mu \) and \( \sigma \) have two parameters (e.g., \( \mu_X \) and \( \mu_Y \)), one function consists of six parameters. We attempted to build the measurement model by using several functions. Although the Gaussian function has a value from 0 to \( \alpha \), MOS is given by a value from 1 to 5. A normalization by subtracting one was conducted before fitting. The initial values of parameters \( \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho, \) and \( \alpha \) were set to 5.4, 60, 2.7, 30, 0, and 1, respectively. The initial values of \( \mu \) and \( \sigma \) were set based on the peak value of the heatmap in Fig. 2. We have made an optimization using linear least squares.

The number of variations in the vibrato rate and extent was 64 (8×8). When the number of Gaussian functions was set to more than ten, the total number of parameters for constructing the Gaussian function exceeds 64. We conducted a fitting by using between one and ten Gaussian functions. Figure 3 shows the models built using one, three, five, and seven Gaussian functions. The horizontal and vertical axes represent the vibrato extent and vibrato rate, respectively.

Next, we evaluated the model by using the following error index defined as,
The use of interpolation in the heatmap in Fig. 2 may achieve the objectives of Sect. 1 to some extent. However, it is not appropriate to use a heatmap that only compares each vibrato when analyzing the impact of other factors such as pitch on vibrato naturalness. It is necessary to model the overall distribution in a comparable way to account for future demand. Therefore, to build a measurement model for the naturalness of vibrato, we hypothesized that 2-dimensional Gaussian functions would be effective as a basic function for non-linear fitting, and results showed that at least three Gaussian functions could approximate the subjective score at 4.0%.

To improve the flexibility of the model, an interpolation would be effective. It is possible to interpolate the parameters in each Gaussian function obtained in the evaluation, which would enable the model to approximate the heatmaps in interpolated pitch. Since singing synthesizers are required to display the range of parameters in each pitch, this interpolation would help the user to better control the vibrato parameters.

6. Conclusion

In this study, we performed a subjective evaluation using MOS on the naturalness of vibrato parameters. The results showed that there was a parameter range perceived as natural without depending on the listener, which suggests that the measurement model can be built on the basic of vibrato parameters. We also found that the appropriate number of two-dimensional Gaussian functions is three. Our model shows good potential as a tool to support singing synthesis by displaying a guide that always gives a natural vibrato.

In our evaluation, we calculated the average under all vibrato conditions. In the future work, expanding this model to optimize the model for the singer and the pitch of the singing would be useful. Also, subjective evaluation experiments should be conducted with higher-quality vibrato singing voices converted using a high-quality vocoder.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP19K12736 and JST PRESTO Grant Number JP-MJPR18J8, Japan.

References

[1] M. Goto, “Singing information processing,” 2014 12th Int. Conf. Signal Process. (ICSP), pp.2431–2438, 2014.
[2] K. Kojima, M. Yanagida, and I. Nakayama, “Variability of vibrato—a comparative study between Japanese traditional singing and bel canto—,” Proc. Speech Prosody 2004, pp.151–154, 2004.
[3] M. Akagi and H. Kitakaze, “Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours,” Proc. ICSP 2000, pp.458–461, 2000.
[4] T. Saitou and M. Goto, “Acoustic and perceptual effects of vocal training in amateur male singing,” Tenth Annual Conference of the International Speech Communication Association, pp.832–835, 2009.
[5] T. Saitou, M. Goto, M. Unoki, and M. Akagi, “Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices,” Proc. WASSPA 2007, pp.215–218, 2007.
[6] N. Migita, M. Morise, and T. Nishiura, “A study of vibrato features to control singing voices,” Proc. ICA 2010, pp.23–27, 2010.
[7] T. Nakano, M. Goto, and Y. Hiraga, “An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features,” Proc. INTERSPEECH 2006, pp.1706–1709, 2006.
[8] T. Miyazaki and M. Morise, “Estimation of vibrato parameters from singing waveform and its evaluation,” IPSJ Journal, vol.61, no.12,
pp.1936–1946, 2020.
[9] E. Prame, “Measurements of the vibrato rate of ten singers,” STL–QPSR, vol.33, no.4, pp.73–86, 1992.
[10] J. Sundberg and J. Bretos, “Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos,” TMH-QPSR, vol.43, no.1, pp.37–44, 2002.
[11] C. Suzuki, H. Banno, K. Asahi, and M. Morise, “Investigation of effect on naturalness of vibrato with fine fluctuation,” IEEJ Transactions on Electronics, Inf. & Syst., vol.139, no.2, pp.182–183, 2019.
[12] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. & Syst., vol.E99-D, no.7, pp.1877–1884, July 2016.
[13] H. Kawahara, Y. Hirachi, M. Morise, and H. Banno, “Procedure senza vibrato: a key component for morphing singing,” Proc. IC-SLP2004, vol.5, pp.934–937, 2004.
[14] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” Proc. INTERSPEECH 2017, pp.2321–2325, 2017.