Transferring Vocal Expression of F0 Contour using Singing Voice Synthesizer

Yukara Ikemiya, Katsutoshi Itoyama, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University
606-8501 Sakyo, Kyoto, Japan
{ikemiya,itoyama,okuno}@kuis.kyoto-u.ac.jp

Abstract. A system for transferring vocal expressions separately from singing voices with accompaniment to singing voice synthesizers is described. The expressions appear as fluctuations in the fundamental frequency contour of the singing voice, such as *vibrato*, *glissando*, and *kobushi*. The fundamental frequency contour of the singing voice is estimated using the subharmonic summation in a limited frequency range and aligned temporally to chromatic pitch sequence. Each expression is transcribed and parameterized in accordance with designed rules. Finally, the expressions are transferred to given scores on the singing voice synthesizer. Experiments demonstrated that the proposed system can transfer the vocal expressions while retaining singer's individuality on two singing voice synthesizers: the *Vocaloid* and the *CeVIO*.

1 Introduction

Every singer has unique vocal expressions and singing style, which characterize his or her singing. The goal of our study is to create a library of vocal expressions and styles that can be applied to consumer-generated media and music information retrieval [1]. Such a library would enable the vocal expressions of favorite singers to be transferred other songs by using a singing voice synthesis system, such as the Vocaloid [2], and retrieval of songs based on singing style. A demonstration of our vocal expression transfer is available on-line[1].

This paper describes a system that transfers vocal expressions involving variation and fluctuation of the fundamental frequency (F0), such as *vibrato*, *kobushi*, and *glissando*, which are extracted from singing voices with instrumental accompaniment. Changes in F0 characteristics affect singing voice quality and individuality more than changes in spectral characteristics [3, 4]. Fig. 1 shows a typical template for each expression. Vibrato is a deliberate, periodic fluctuation in the F0 contour. Kobushi is short tremolo that appears in Japanese folk songs such as *enka* and *min-yo*. Glissando is generally separated into two types: *glissdown*, which is a glide down in pitch for an offset note, and *glissup*, which is a glide up in pitch for an onset note.

The proposed transfer system consists of three steps as follows:

[1] winnie.kuis.kyoto-u.ac.jp/members/ikemiya/demo/sst2013.html

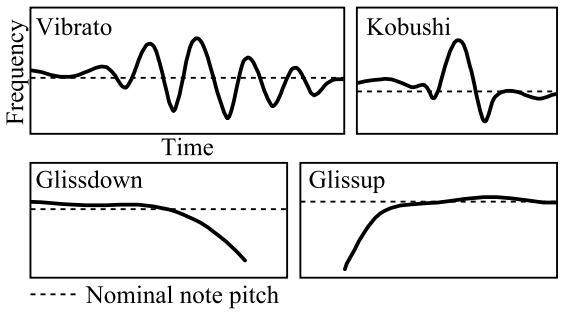


Fig. 1: Vocal expressions.

1. Estimation of singing voice F0 from polyphonic music.
2. Transcription of vocal expressions in F0 contour.
3. Transfer of extracted expressions into a new song.

We use a sequence of symbolized, chromatic pitches of the song, such as (G4, A#4, C5, ...), to achieve F0 estimation with high accuracy. Note that the sequence does not contain note values. The F0 contour is searched for in a limited frequency range by considering the smoothness.

About steps 2 and 3, we discuss problems with a number of existing studies. While some studies have been aimed at making synthesized singing voices more human-like by adding pitch fluctuations [5, 6], they use manually-tuned vibrato expressions and they do not represent singers' individuality. The VocaListener2 [7] transfers pitch, volume, and timbre from the user's singing to the Vocaloid system directly. The problem of the VocaListener2 that the transfer can be applied to the same song because the VocaListener2 simply extracts moment-to-moment fluctuations, not vocal expressions such as vibrato. Although a statistical model of vocal F0 fluctuation by the second-order transfer function has been proposed [8], this model also cannot represent separately each vocal expression. Singing voice synthesis systems based on the hidden Markov model (HMM) [9–11] learn singing styles as the distributions of the feature vectors and reconstruct them in other songs. A problem of the HMM-based systems is that they require many sets of singing voices without accompaniments and effects and corressponding scores for learning. Yasuraoka *et al.* have proposed a musical instrument sound synthesizer which learns the pitch, volume, and timbre of the instrument sounds such as guitar and reconstructs them in other melodies [12]. The objective of their system is very close to ours except for the target sounds. In addition, their system focuses on musical instrumental sounds, not vocal with accompanying sounds, as a target musical signal.

2 Estimating Fundamental Frequency

Our proposed transfer system requires a method for estimating the F0 contour of a singing voice with accompaniments with both high accuracy and high

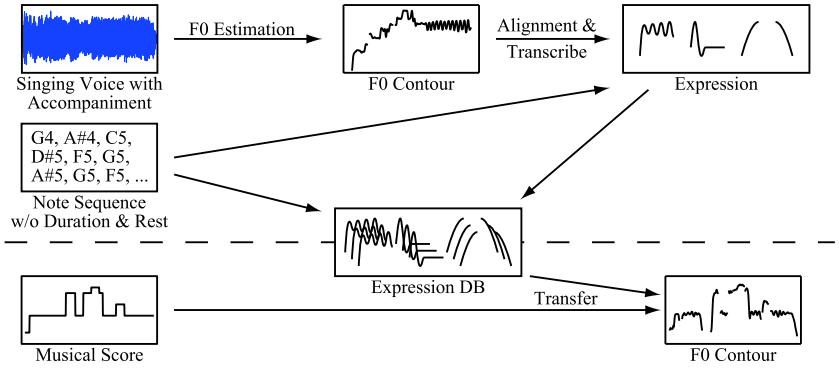


Fig. 2: Overview of proposed vocal expression transfer system.

frequency-resolution. The search range for F0 is limited, from $L-400$ to $H+400$ cent[2], where L and H are the lowest and highest pitches in the given pitch sequence.

We assume that the F0 of singing voices has the following properties:

1. A singing voice usually performs the predominant part.
2. A large movement in the F0 takes a long time.
3. It has inertia: a moving F0 continues to move and sometimes overshoots the desired pitch.

We developed an objective function that satisfies these properties to obtain an optimal F0 contour.

The objective function to be maximized is defined as

$$\hat{F} = \underset{F\in\{f_1,\ldots,f_T\}}{\arg\max} \left(\sum_{t=1}^{T} \log P_M(f_t) + \sum_{t=2}^{T} \log P_{\Delta F0}(f_t - f_{t-1}) + \sum_{t=3}^{T} \log P_{\Delta\Delta F0}(f_t - 2f_{t-1} + f_{t-2}) \right).$$

An optimal F0 contour is computed by using the Viterbi algorithm. Each term on the right-hand side corresponds to a property described above.

The first term is for property 1 on predominancy:

$$P_M(f_t) = \frac{\mathrm{SHS}(t,f)}{\int_{L-400}^{H+400} \mathrm{SHS}(t,f')\,df'},$$

$$\mathrm{SHS}(t,f) = \sum_{n=1}^{N} 0.84^{n-1}\,\mathrm{CQ}(t, f + 1200\log_2 n).$$

[2] Cent is the logarithmic unit of frequency, and a harftone is equal to 100 cent.

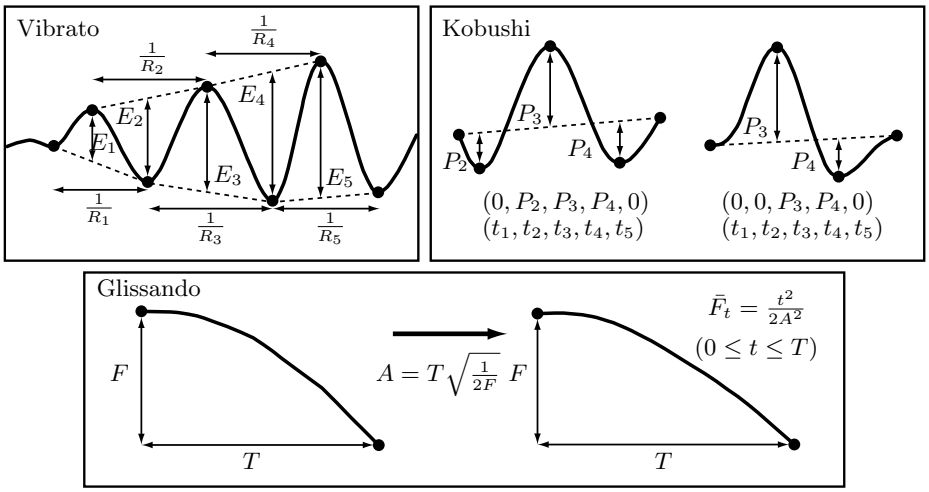


Fig. 3: Vocal expression parameterization.

We use the subharmonic summation (SHS) [13] to calculate the likeliness of the predominant F0 at time t and the N overtones to calculate frequency f. Here we set $N = 7$. $\mathrm{CQ}(t, f)$ is the spectrogram obtained the constant-Q transform [14].

The second term is for property 2 on F0 change:

$$P_{\Delta F0}(f) = \begin{cases} 1/200 & \text{if} \quad |f| < 100 \\ 0 & \text{otherwise} \end{cases}.$$

This term takes a uniform positive value when the F0 adjacent frame difference is less than a certain level or zero.

The third term is for property 3 on inertia. It is defined as a truncated Gaussian function:

$$P_{\Delta\Delta F0}(f) \propto \begin{cases} \exp(-f^2/(2 \times 50^2)) & \text{if} \quad |f| < 50 \\ 0 & \text{otherwise} \end{cases}.$$

This term takes a large value when the second-order derivative of F0 is close to zero.

3 Transcribing Vocal Expressions

This section describes a method for transcribing vocal expressions from the estimated F0 contour and storing them as parameters. We assume that the expressions are strongly connected to the notes on the score. In other words, the expressions do not exceed the note boundaries. We thus temporally align the F0 contour and the pitch sequence. This alignment is regarded as a problem of hidden state estimation. Let the observation and the hidden state sequence be

the F0 contour and pitch sequence, respectively. The alignment is regarded as a problem of state estimation for each observation. We thus define the state likeliness for the given F0 as the squared difference of the F0 and the pitch so that this problem can be solved using the Viterbi algorithm. When the F0 contour does not have values for a certain span, the span is detected as a rest, and the state changes to the next one. An overview of expression transcription is shown in Fig. 3.

3.1 Vibrato

We detect vibrato sections by using a previously proposed method [15] that uses short-time Fourier transform to find a sharp peak that corresponds to the vibrato rate (the number of vibrations per second). We uniquely restrict the range of the extent (the amplitude of vibration) to 30 – ∞ cent and that of the rate to 3 – 8 Hz because enka and min-yo songs have a vibrato with a much larger extent and a lower rate than the restrictions proposed previously [15].

A vibrato section is represented as a sequence of pairs of two parameters, rate R_i and extent E_i, such as $((E_1, R_1), (E_2, R_2), (E_3, R_3), \ldots)$. Let I be the number of peak points of the vibrato, and let f_i and t_i be the logarithmic frequency and time of the i-th peak point in the F0 contour ($i = 1, \ldots, I$). The extent and rate parameters $((E_1, R_1), \ldots, (E_{I-2}, R_{I-2}))$ are calculated as

$$R_i = \frac{1}{t_{i+2} - t_i} \quad \text{and}$$
$$E_i = \left| (f_{i+2} - f_i)(t_{i+1} - t_i) R_i + (f_i - f_{i+1}) \right| .$$

3.2 Glissando

Glissdown (glissup) sections are extracted by detecting a monotonic decrease (increase) of more than F_{least} cent from a phrase end (beginning). On the basis of the results of our preliminary experiments, F_{least} set to 200 cent.

A glissdown (glissup) is modeled as a parabola, and stored as the parameters of the parabola curve and its duration. Since they have bilateral symmetry, we describe only glissdown here. Let T s and F cent be the duration of the detected glissdown and the frequency decrease. Coefficient A of a parabola is calculated as

$$A = T\sqrt{\frac{1}{2F}}.$$

3.3 Kobushi

Before detecting kobushi sections, we extract all peak, valley, and point which cross the pitch of the corresponding note from the F0 contour. Although the pattern of kobushi is not well defined among professional singers, we have found by observing the F0 contour of enka and min-yo songs that kobushi follows three rules.

1. Kobushi sections do not overlap vibrato sections.
2. A kobushi section has only one peak greater than 150 cent (main peak).
3. In front of and behind the main peak, one or no small valley(s) (sub-peak) appears.

We define that a kobushi section contains the main peak and sub-peaks, and that the gradient between the peaks is more than V cent/s. Here we set $V = 1000$.

A kobushi section is stored as a quintuple: a starting point, a left sub-peak, a main peak, a right sub-peak, and an end point. Each element of the quintuple is a pair of an extent of the peak and its time. If a sub-peak does not exist, the extent of the corresponding element is set to zero. The extent P_i of the i-th peak is calculated as

$$P_i = f_i - \left(\frac{f_5 - f_1}{t_5 - t_1}(t_i - t_1) + f_1 \right),$$

where t_i and f_i denote the time and log-frequency of the i-th peak, respectively.

4 Transferring Vocal Expressions

This section describes the process of transferring vocal expressions to a vocal synthesizer by using a vocal expression library. The synthesizer is assumed to provide musical score information for synthesized vocal expressions and a mechanism for handling pitch.

4.1 Vocal Expression Library

A vocal expression library consists of sets of vocal expression parameters and note information for each vocal expression (vocal expression set). Note information includes four elements: pitch, duration, musical intervals, and label. Musical intervals are the differences in pitch from one note to the next. Note label represents whether a note is at the beginning, the middle, or the end of a musical phrase. If there is an unvoiced section of over 200 ms between two notes, the first note is labeled as the end of the phrase, and the second note is labeled as the beginning.

4.2 Preprocessing

The vocal pitch range of the input score may be completely different from that of the library. This makes it difficult to transfer vocal expressions on the basis of a simple rule. Thus, the pitch range in the library is adjusted by shifting the lowest pitch in the library to that of the input score. Furthermore, the note information for all notes in the score is acquired.

4.3 Transfer Rule

The following process is performed to each note in the input score. First, a set of vocal expressions matching four conditions are extracted from the library.

for all expressions Note label is the same as that of the target note.
for all expressions Difference between note pitch (note number) of vocal expression set and that of target note is smaller than M.
for kobushi and glissando Full length of vocal expression is shorter than the length of target note.
for kobushi Signs of note transitions are the same as that of the target note.

The smaller the M, the stricter the rule for transfer.

Second, from the extracted set, the set of notes nearest to the target note are chosen for each vocal expression. When no vocal expressions are extracted, no vocal expressions are applied to the target note. The nearness of notes is determined using two indices and priority.

1. Difference in note pitch
2. Difference in note length

If the full length of the selected vibrato is larger than the length of the target note, the vibrato is trimmed to end at the end of the note or the beginning of the glissdown, and if smaller, it is extended with the rate to the end of the note. Vocal expressions are transferred by resynthesizing the expression in accordance with the selected parameters and pasting it on the F0 contour of the target note.

5 Evaluation

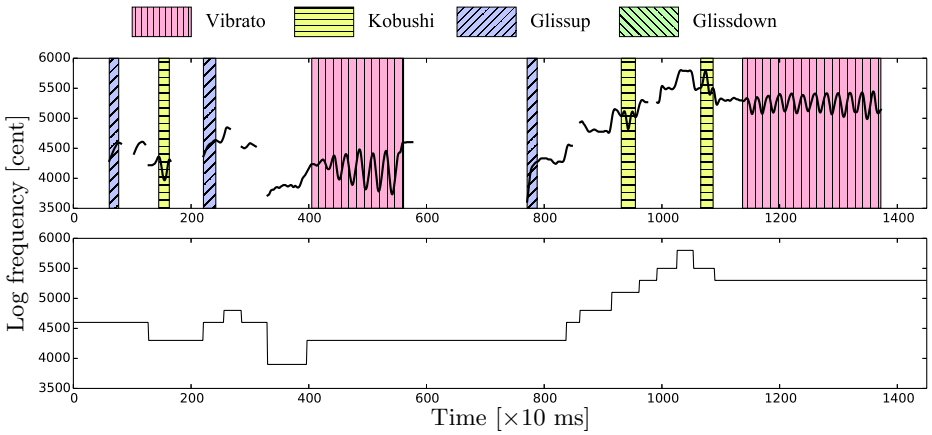
5.1 Experimental Settings

All musical pieces for our experiments were converted to a 16-kHz sampling rate with 16 bits per sample. A constant-Q spectrogram was calculated with a time resolution of 10 ms, a frequency resolution of 6 cent, a frequency range of 60 to 6000 Hz, and a Q value of $(1/(2^{0.01} - 1))/5$. Additionally, we postulated that voiced sections were detected in advance.

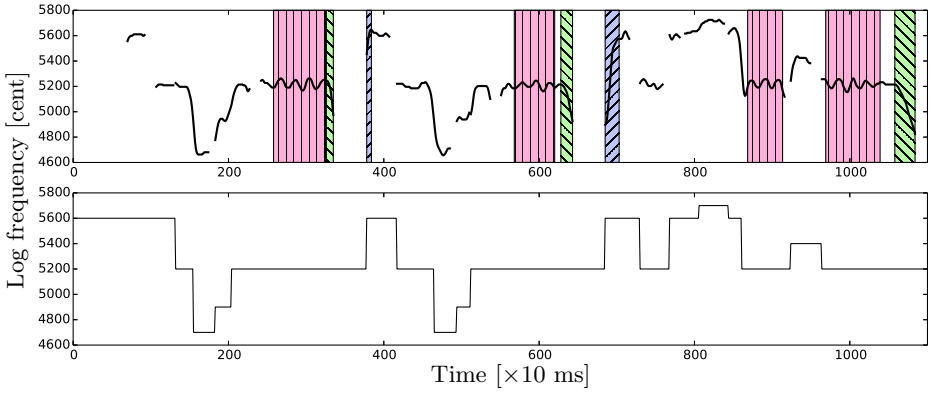
5.2 Transcription with Commercial Recordings

We applied our method to two commercial recordings, a verse part of "Jinsei Ichiro (by a Japanese famous singer, Misora Hibari)" and a chorus part of "Crispy (by a Japanese famous singer, Spitz)". The former is an enka song, while the latter is a Japanese pop song.

Fig. 4 shows the results of vocal expression identification. On the left (Fig. 4(a)), we can see that both long and short vibrato, kobushi characteristic of enka and glissup attached to strained singing, were identified. On the right (Fig. 4(b)), we can see that frequent glissdown best characterizes the singing style of the "Spitz"



(a) Jinsei Ichiro (Misora Hibari)



(b) Crispy (Spitz)

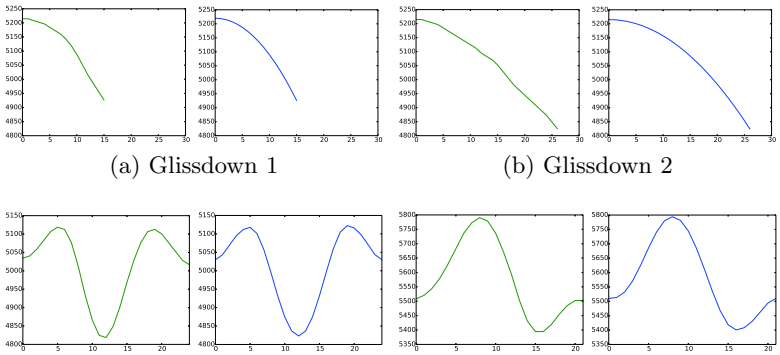
Fig. 4: Vocal expression identification. Upper figures show estimated F0 contour and identified expressions; bottom figures show alignment of note sequence with F0 contour.

vocal. Fig. 5 shows the result of resynthesis of the vocal expressions. Figs. 5(a)-(b) correspond to the second and third glissdowns in Fig. 4(b), Figs. 5(c)-(d) correspond to the second and third kobushi expressions in Fig. 4(a). The root mean square errors for glissdown and kobushi were 22.3 and 16.0 cent, respectively. Considering that a semitone is 100 cent, we can say that each expression was precisely resynthesized despite differences in scale and shape.

5.3 Transferring Expressions with Singing Voice Synthesizers

For transferring vocal expressions, we used two singing voice synthesizers: the *Vocaloid* and the *CeVIO*[3]. In Vocaloid, F0 can be controlled with pitch bend

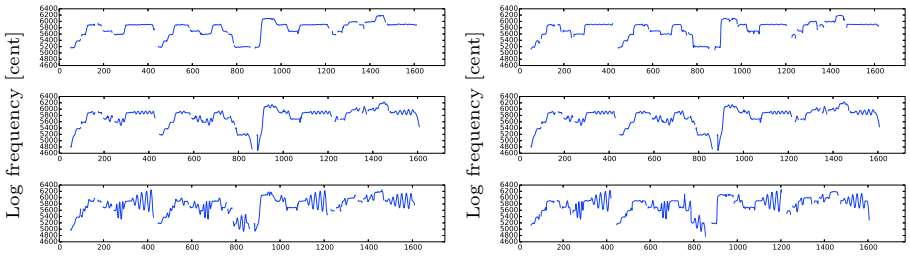
[3] `http://cevio.jp/`



(a) Glissdown 1 (b) Glissdown 2

(c) Kobushi 1 (d) Kobushi 2

Fig. 5: Resynthesized vocal expressions. Left figures (green) show original contour; right figures (blue) show synthesized contour.



(a) Vocaloid (b) Cevio

Fig. 6: F0 contour of synthesized singing voice with vocal expression. Top, center, and bottom figures are without expressions, with Spitz expressions, and with Misora Hibari expressions, respectively.

parameters, which change the pitch up and down from the chromatic regular pitch. In CeVIO, F0 can be controlled by directly setting the logarithmic values of the pitch.

Figure 6 shows the results of vocal expression transfer to the singing voice synthesizers. From the top, the figures of F0 contour with no vocal expression, with vocal expression of "Spitz", and with vocal expression of "Misora Hibari" are shown. We can confirm that vocal expression such as vibrato, kobushi, and glissando are transferred similarly in both synthesizers.

6 Conclusion

Our developed system for transferring vocal expressions achiveved to resynthesize expressions that reflect singer's individuality extracted from singing voices

with accompaniments. Vocal expressions are detected from the F0 contour and parametrized based on the designed rules, and then transferred to the input score using the Vocaloid and CeVIO singing voice synthesizers. Experimental results demonstrated that our method can transcribe vocal expressions from commercial songs and resynthesize them precisely. In future work, we intend to expand our method to other types of expressions. We also intend to apply the library of the expressions for retrieving musical pieces on the basis of the singing styles.

References

1. Downie, J.S.: Music information retrieval. Annu. Rev. Inf. Sci. Technol. **37** (2003) 295–340
2. Kenmochi, H., Ohshita, H.: Vocaloid - commercial singing synthesizer based on sample concatenation. In: INTERSPEECH 2007. (2007) 4009–4010
3. Saito, T., Goto, M.: Acoustic and perceptual effects of vocal training in amateur male singing. In: INTERSPEECH 2009. (Sep 2009) 832–835
4. Guzman, M.A., Dowdall, J., Rubin, A.D., Maki, A., Levin, S., Mayerhoff, R., Jackson-Menaldi, M.C.: Influence of emotional expression, loudness, and gender on the acoustic parameters of vibrato in classical singers. Journal of Voice **26**(5) (2012) 675–681
5. Stables, R., Athwal, C., Bullock, J.: Fundamental frequency modulation in singing voice synthesis. In: International Conference on Speech, Sound and Music Processing: Embracing Research in India. (2012) 104–119
6. Umbert, M., Bonada, J., Blaauw, M.: Generating singing voice expression contours based on unit selection. In: SMAC 2013. (Jul 2013)
7. Nakano, T., Goto, M.: VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics. In: ICASSP 2011. (2011) 453–456
8. Ohishi, Y., Kameoka, H., Mochihashi, D., Kashino, K.: A stochastic model of singing voice F0 contours for characterizing expressive dynamic components. In: Proc. INTERSPEECH. (Sep 2012)
9. Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., Tokuda, K.: Recent development of the HMM-based singing voice synthesis system - Sinsy. In: Proc. ISCA Tutorial and Research Workshop on Speech Synthesis. (Sep 2010) 211–216
10. Saino, K., Tachibana, M., Kenmochi, H.: A singing style modeling system for singing voice synthesizers. In: Proc. INTERSPEECH. (Sep 2010) 2894–2897
11. Lee, S.W., Ang, S.T., Dong, M., Li, H.: Generalized F0 modelling with absolute and relative pitch features for singing voice synthesis. In: Proc. ICASSP. (Mar 2012) 429–432
12. Yasuraoka, N., Abe, T., Itoyama, K., Takahashi, T., Ogata, T., Okuno, H.G.: Changing timbre and phrase in existing musical performances as you like. In: ACM Multimedia 2009. (2009) 10
13. Hermes, D.J.: Measurement of pitch by subharmonic summation. J. Acoust. Soc. Am. **83**(1) (Jan 1988) 257–264
14. Brown, J.C.: Calculation of a constant q spectral transform. J. Acoust. Soc. Am. **89**(1) (Jan 1991) 425–434
15. Nakano, T., Goto, M.: An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In: Proc. INTERSPEECH. (Sep 2006)