TRANSCRIBING VOCAL EXPRESSION FROM POLYPHONIC MUSIC

Yukara Ikemiya, Katsutoshi Itoyama, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

A method for transcribing vocal expressions such as *vibrato*, *glissando*, and *kobushi* separately from polyphonic music is described. The expressions appear as fluctuation in the fundamental frequency contour of the singing voice. They can be used for search and retrieval of music and for expressive singing voice synthesis based on singing style since they strongly reflect the individuality of the singer. The fundamental frequency contour of the singing voice is estimated using the Viterbi algorithm with limitation from a corresponding note sequence. Next, the notes are aligned with the fundamental frequency sequence temporally. Finally, each expression is identified and parameterized in accordance with designed rules. Experiments demonstrated that this method can transcribe expressions in the singing voice from commercial recordings.

Index Terms— Singing voice analysis, Vocal expression identification / transcription, F0 estimation.

1. INTRODUCTION

Every singer has a unique singing style, which characterizes his or her singing. The goal of our study is to create a library of singing styles that can be applied to consumer-generated media (CGM) and music information retrieval (MIR) [1]. Such a library will enable us the singing style of favorite singers to be transferred to singing voice synthesis systems, such as the Vocaloid [2], and the retrieval of songs based on singing style. A demonstration of singing-style transfer is available on-line [1].

While some systems have been aimed at making synthesized singing voices more human-like by adding pitch fluctuations [3, 4, 5], they do not take into account singer individuality. Singing voice synthesis systems based on the hidden Markov model (HMM) [6, 7, 8] learn singing style as the feature vectors of HMM; however, it is difficult to create a singing style library for various singers since such a library requires many sets of the singing voices and corresponding scores for learning. Although a statistical model of fundamental frequency (F0) contour of singing voices was proposed [9], the model parameters distinguish only whether the generated fluctuations exist in the change between two consecutive notes or on a note. Some studies aim to transcript particular characteristics from singing voice F0 contour [10, 11].

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

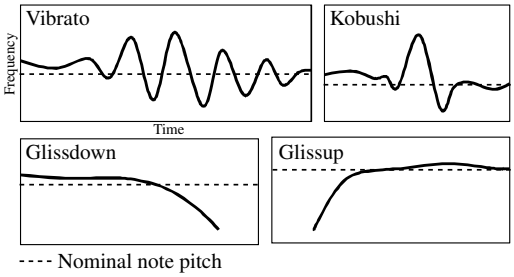


Fig. 1: Vocal expressions.

While existing systems use only the singing voice for input, we target polyphonic music since users would naturally want to analyze commercial recordings. We break down singing style into parameters to use for singing style transfer or retrieval. In this paper, we present a method for transcribing vocal expressions (*vibrato*, *kobushi*, and *glissando*) separately. Figure 1 shows a typical template for each expression. *Kobushi* is short tremolo that appears in Japanese folk songs such as *enka* or *min-yo*. *Glissando* is generally separated into two types: *glissdown*, which glides down pitch in an offset note, and *glissup*, which glides up pitch in an onset note. They are observed as fluctuations in the F0 contour and affect perception of singing voice individuality [12, 13]. Figure 2 shows a block diagram of our method. Two problems need to be addressed.

1. Estimation of singing voice F0 from polyphonic music
2. Identification and parameterization of vocal expressions in F0 contour

While several studies have targeted the former [14, 15, 16, 17], the methods proposed did not achieve highly precise estimation or were time-consuming. We propose a simple and precise method using the corresponding note sequence. Although some studies use a score to improve F0 estimation [18, 19, 20, 21], we use only note numbers and the order. The F0 contour is searched for in a limited frequency range by considering the smoothness. The latter problem is a step in vocal expression transcription. Each vocal expression is identified and parameterized in accordance with designed rules on the presupposition that the expressions do not overlap temporally.

2. F0 ESTIMATION

Singing voice F0 estimation requires high accuracy and high-frequency resolution since vocal expressions are directly ex-

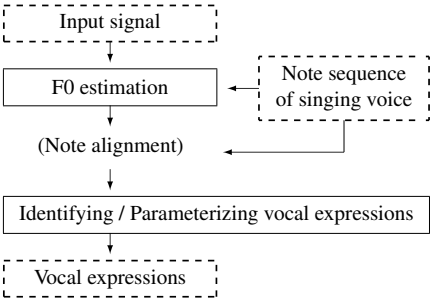


Fig. 2: Block diagram of vocal expression transcription.

tracted from F0. We thus use the note sequence, which represents the sequential order of notes in singing. The search range for F0 is limited from 400 [*cent*] lower than the minimal value to 400 [*cent*] higher than the maximal value of the sequence, where "*cent*" is the logarithmic unit of frequency and is calculated from linear frequency [*Hz*]: $C = 1200 * \log_2 \frac{H}{L}$. The H and L denote the original linear frequency and the lowest frequency in log scale, respectively. In the next section, we describe the formularization for F0 estimation.

2.1. Formularization

The estimation of F0 can be considered to be a time sequence search problem in time-frequency domain. To represent the likelihood that an F0 f is the most predominant F0 in the t-th spectrum, we introduce measurement cost function $P_M(t, f)$. In the simplest method, F0 is estimated to be the maximum value of $P_M(t, f)$. However, this causes errors of half/double pitch or other instrument pitches.

To prevent this, we introduce smoothness cost functions $P_{\Delta F0}(f)$ and $P_{\Delta\Delta F0}(f)$ to represent the probability functions of the first and secondary difference of F0, respectively. Intuitively, these are the features that ensure respectively that singing voice F0 does not change drastically and that it changes smoothly. With these, F0 estimation comes down to the maximization problem represented by

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

We effectively compute this using the Viterbi algorithm.

2.2. Design of cost functions

The measurement cost function is set to the subharmonic summation (SHS) spectrogram [22], which is normalized for each time frame. This function can be calculated easily and quickly:

$$SHS(t, f) = \sum_{n=1}^{N} (0.84)^{n-1} CQ(t, f + 1200 \log_2 n) , \quad (2)$$

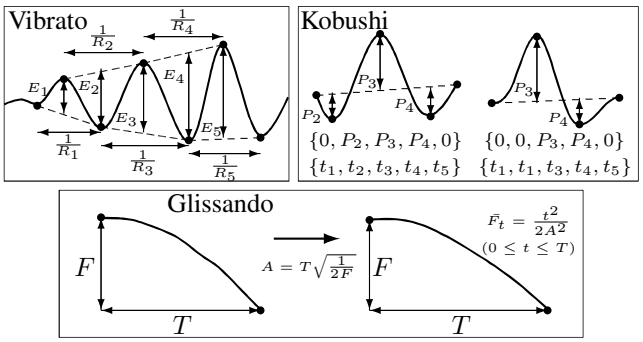


Fig. 3: Vocal expression parameterization.

$$P_M(t, f) = \frac{SHS(t, f)}{\int_{f_l}^{f_u} SHS(t, f')df'} , \quad (3)$$

where $CQ(t, f)$ denotes a constant-Q spectrogram [23] for which f and t are log frequency [*cent*] and time, respectively. The N is the number of harmonics being considered, and f_l and f_u are the lower and upper limits of the search range for F0.

The smoothness cost functions are

$$P_{\Delta F0}(f) = \mathscr{U}(-100, 100) \quad (4)$$

$$P_{\Delta\Delta F0}(f) = \begin{cases} \mathscr{N}(f|0, 50^2) & (-50 < f < 50), \\ 0 & (elsewise) \end{cases} \quad (5)$$

where $\mathscr{U}(L, U)$ denotes a uniform distribution with minimum and maximum values of L and U [*cent*], and $\mathscr{N}(f|\mu, \sigma^2)$ denotes a normal distribution with mean and standard deviation of μ and σ [*cent*]. The time frame width is 10 [*msec*].

Although a Laplace or normal distribution could be used for $P_{\Delta F0}(f)$ [16], such a pointed cost function for $\Delta F0$ can cause excess smoothness of the F0 contour [24]. This flattens the peaks in *vibrato*, *kobushi*, and so on, which results in degraded vocal expression identification. For this reason, we use $P_{\Delta F0}(f)$ only to limit the transition range and use $P_{\Delta\Delta F0}(f)$ to achieve smoothness. This not only on improves F0 estimation accuracy but also removes fine fluctuations [9] unrelated to the singer's individuality.

3. IDENTIFYING AND PARAMETERIZING VOCAL EXPRESSIONS

Before identifying expressions, we temporally align the note sequence with the F0 contour by minimizing the squared errors using the Viterbi algorithm. In this alignment, a note changes to the next one when there is a rest. We identify vocal expressions in each note section. In the following subsections, we describe how we identify *vibrato*, *glissando*, and *kobushi*. The parameterization of expressions is shown in Figure 3.

3.1. Vibrato

Vibrato has two parameters, extent and rate, that respectively represent the amplitude and speed of pitch variation.

Table 1: F0 estimation accuracy under four conditions: maximizing MCF (N), MCF with frequency range limitation (FRL), MCF with smoothness cost functions (S), and MCF with FRL and S (FRL-S). "FRL-S" for SHS represents our method.

MCF	PreFEst-core				SHS			
MPE	N	FRL	S	FRL-S	N	FRL	S	FRL-S
50	70.32	75.18	73.64	76.84	48.81	81.65	51.76	**85.93**
25	60.80	64.35	63.87	66.38	45.29	74.36	47.16	**77.91**

MCF: measurement cost function
MPE: maximum permissible error [*cent*]

Identification To identify *vibrato*, we use a previously proposed method [25] that calculates the closeness between the F0 contour and the sine wave by short-time Fourier transformation. We found that *enka* and *min-yo* songs have a vibrato with a much larger extent and a lower rate than the values proposed previously. We thus uniquely set the range of the extent to 30 – ∞ [*cent*] and that of the rate to 3 – 8 [*Hz*].

Parameterization Let f_i and t_i be the log-frequency and time of the i-th peak point (zero cross point). Extent E_i and rate R_i are calculated using:

$$E_i = |(f_{i+1} - f_{i-1})\frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}} + (f_{i-1} - f_i)|,$$
$$R_i = \frac{1}{t_{i+1} - t_{i-1}}, \quad 1 \leq i \leq I - 1, \tag{6}$$

where I is the number of peak points. Vibrato can be resynthesized by using a time-varying sine wave and a set of $\{i, E_i, R_i\}$.

3.2. Glissando

Identification *Glissdown* (*glissup*) is identified as 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} is set to 200.

Parameterization *Glissdown* and *glissup* are stored as the parameters of a parabola. Since they are considered to have bilateral symmetry, we describe only *glissdown* here. From time width T [*sec*] and log-frequency width F [*cent*] of the observed *glissdown*, the coefficient A of a parabola is calculated:

$$A = T\sqrt{\frac{1}{2F}}. \tag{7}$$

Glissdown can be resynthesized as a parabola determined by A and T.

3.3. Kobushi

Before identifying *kobushi*, we obtain feature points that are simply defined as zero cross points and rising and trailing points in the F0 contour.

Identification Although the pattern of *kobushi* is not well defined, we have found by listening to *enka* songs and observing the pitch changes that *kobushi* follows three rules.

1. *Kobushi* sections do not overlap with *vibrato* sections.
2. A *kobushi* section has only one peak with a height greater than 150 [*cent*] (main peak).

Table 2: Vocal expressions per singer.

Singer	Music numbers	Total of vocal expressions		
		Vibrato	Glissdown	Glissup
1.Ogata Tomomi	7,28,52	37	6	41
2.Yoshii Hiromi	17,34,69	106	3	27
3.Morimoto Kosuke	38,39,45	32	13	37

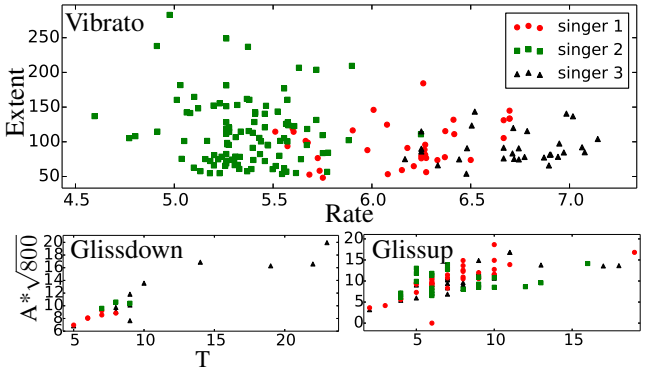


Fig. 4: Vocal expression parameters.

3. In front of and behind the main peak, there is one low peak (sub-peak) of the opposite sign or nothing.

We define a peak as a feature point with a gradient from the previous feature point of more than V [*cent / sec*]. Here, $V = 1000$.

Parameterization We store *kobushi* as 5-length vectors of peak extents and time indices. The vector has five values in order: a starting point, a left sub-peak, a main peak, a right sub-peak, and an end point. If a sub-peak does not exist, the extent of the point is set to zero. The extent P_i of the i-th peak is calculated using

$$P_i = f_i - (\frac{f_5 - f_1}{t_5 - t_1}(t_i - t_1) + f_1), \tag{8}$$

where t_i and f_i denote the time [*sec*] and log-frequency [*cent*] of the i-th peak, respectively. *Kobushi* can be resynthesized as follows. We derive the polynomial whose extreme-value points are the time indices. Next, the extreme values are scaled to the peak extents.

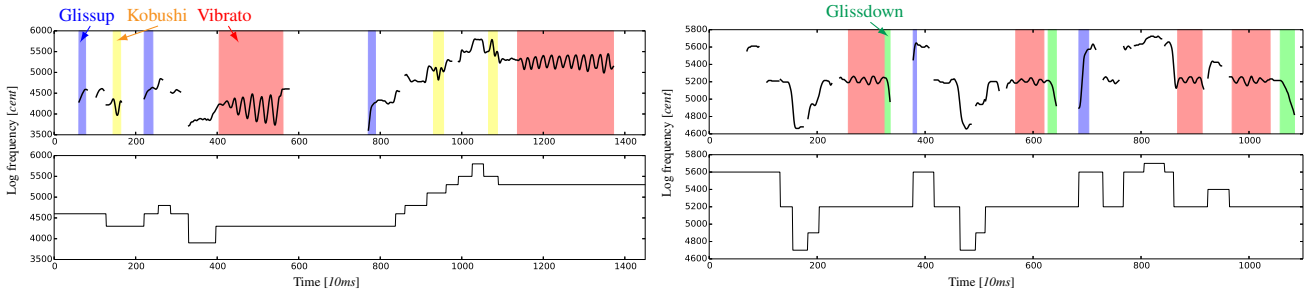
4. EVALUATION

4.1. 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 [*msec*], a frequency resolution of 6 [*bit*], 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.

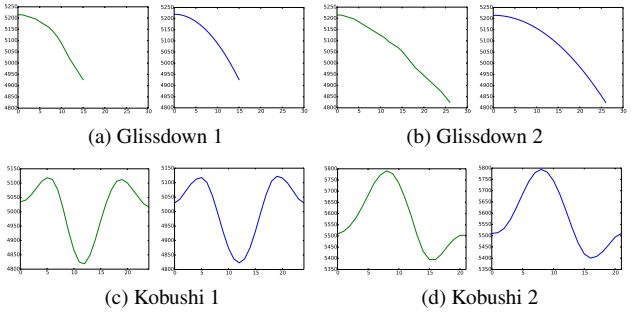
4.2. F0 Estimation Accuracy

Our F0 estimation method was evaluated using the following data and ground truths.



(a) Jinsei Ichiro (Misora Hibari) (b) Crispy (Spitz)

Fig. 5: Vocal expression identification. Upper figures show estimated F0 contour and identified expressions; bottom figures show alignment of note sequence with F0 contour.



(a) Glissdown 1 (b) Glissdown 2

(c) Kobushi 1 (d) Kobushi 2

Fig. 6: Resynthesized vocal expressionsLeft figures (green) show original contour; right figures (blue) show synthesized contour.

Data

We used 96 popular songs from the "RWC Music Database: Popular Music" (RWC-MDB-P-2001) [26]. The songs were divided into 2001 fragments containing from 10 to 20 melody notes.

Ground truths

We used F0 annotation data [27] as the F0 ground truths. F0 accuracy was calculated for maximum permissible errors of 50 and 25 [*cent*] since F0 errors must be small enough for vocal expressions to be transcribed. The accuracy rate was defined as the number of correctly estimated frames divided by the total number of voiced frames.

Our method uses the SHS spectrogram as the measurement cost function (MCF). It was compared with the F0's PDF estimated in the core step of PreFEst [14], a competitive method for melody extraction. The results are shown in Table 1. Our method achieved accurate and precise performance compared with the MCF of PreFEst. While SHS causes half/double pitch errors, the frequency range limitation and smoothness cost functions of our method dramatically suppressed the errors.

4.3. Vocal Expression Individuality

Table 2 shows the results of singing style identification for 3 singers using RWC-MDB-P-2001. Now, we excluded *kobushi*s since the data music includes only popular songs and most of *kobushi*s identified were misidentification of short *vibrato*s.

Figure 4 plots the expression parameters. The vibrato rate and extent were averaged for each vibrato. The parameters for each singer were clustered, especially for *vibrato* and *glissdown*.

This suggests that vocal expressions are useful for singer identification. For example, by combining existing features for singer identification with the vocal expression parameters, improvement of accuracy can be expected.

4.4. Performance with Commercial Recordings

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

Figure 5 shows the results of vocal expression identification. On the left (Fig. 5(a)), we can see that large/slow *vibrato*, *kobushi* characteristic of *enka* and *glissup* attached to strained singing, were identified. On the right (Fig. 5(b)), we can see that frequent *glissdown* best characterizes the singing style of the "Spitz" vocal. Figure 6 shows the result of resynthesis of the vocal expressions. Figs. 6(a)-(b) correspond to the second and third *glissdown*s in Fig. 5(b), Figs. 6(c)-(d) correspond to the second and third *kobushi*s in Fig. 5(a). The root mean square errors of *glissdown* and kobushi are 22.3 and 16.0 [*cent*], respectively. When we consider that a semitone is 100 [*cent*], it can be said that each expression was precisely resynthesized despite differences in scale and shape.

5. CONCLUSION

In our proposed method for vocal expression transcription, the singing voice F0 contour is accurately estimated by considering smoothness. Vocal expressions are then identified from the contour and parameterized in accordance with designed rules. Testing 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 more expressions. We also intend to improve our algorithm to enable more accurate F0 estimation and vocal expression identification. This research was partially supported by KAKENHI (S) No. 24700168.

6. REFERENCES

[1] J. S. Downie, "Music information retrieval," *Annu. Rev. Inf. Sci. Technol.*, vol. 37, pp. 295–340, 2003.

[2] H. Kenmochi and H. Ohshita, "Vocaloid - commercial singing synthesizer based on sample concatenation," *Proc. INTERSPEECH*, pp. 4009–4010, 2007.

[3] R. Stables, C. Athwal, and J. Bullock, "Fundamental frequency modulation in singing voice synthesis," *Proc. of international conference on Speech, Sound and Music Processing: embracing research in India*, pp. 104–119, 2012.

[4] M. Umbert, J. Bonada, and M. Blaauw, "Generating singing voice expression contours based on unit selection," *Proc. SMAC*, Jul 2013.

[5] T. Nakano and M. Goto, "VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," *Proc. ICASSP*, pp. 453–456, May 2011.

[6] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system - Sinsy," *Proc. ISCA Tutorial and Research Workshop on Speech Synthesis*, pp. 211–216, Sep 2010.

[7] K. Saino, M. Tachibana, and H. Kenmochi, "A singing style modeling system for singing voice synthesizers," *Proc. INTERSPEECH*, pp. 2894–2897, Sep 2010.

[8] S. W. Lee, S. T. Ang, M. Dong, and H. Li, "Generalized F0 modelling with absolute and relative pitch features for singing voice synthesis," *Proc. ICASSP*, pp. 429–432, Mar 2012.

[9] Y. Ohishi, H. Kameoka, D. Mochihashi, and K. Kashino, "A stochastic model of singing voice F0 contours for characterizing expressive dynamic components," *Proc. INTERSPEECH*, Sep 2012.

[10] J. C. Devaney, M. I. Mandel, and I. Fujinaga, "Characterizing singing voice fundamental frequency trajectories," *WASPAA*, pp. 73–76, 2011.

[11] S. S. Miryala, K. Bali, R. Bhagwan, and M. Choudhury, "Automatically identifying vocal expressions for music transcription," *Proc. ISMIR*, pp. 239–244, 2013.

[12] T. Saito and M. Goto, "Acoustic and perceptual effects of vocal training in amateur male singing," *Proc. INTERSPEECH*, pp. 832–835, Sep 2009.

[13] M. A. Guzman, J. Dowdall, A. D. Rubin, A. Maki, S. Levin, R. Mayerhoff, and M. C. Jackson-Menaldi, "Influence of emotional expression, loudness, and gender on the acoustic parameters of vibrato in classical singers," *Journal of Voice*, vol. 26, no. 5, pp. 675–681, 2012.

[14] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, Sep 2004.

[15] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," *Proc. ISMIR*, vol. 15, 2006.

[16] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search," *Proc. ICASSP*, vol. 5, pp. 253–256, May 2006.

[17] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing melody extraction in polyphonic music by harmonic tracking," *Proc. ISMIR*, pp. 373–374, Oct 2007.

[18] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Evaluation of a score-informed source separation system," *Proc. ISMIR*, pp. 219–224, 2010.

[19] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," *Proc. ICASSP*, pp. 45–48, 2011.

[20] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," *Proc. ISMIR*, pp. 277–282, 2012.

[21] S. Ewert and M. Müller, "Using score-informed constraints for nmf-based source separation," *Proc. ICASSP*, pp. 129–132, 2012.

[22] Dik J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, Jan 1988.

[23] Judith C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, Jan 1991.

[24] H. Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Trans. on SMC*, vol. smc-13, no. 3, pp. 208–214, Mar 1983.

[25] T. Nakano and M. Goto, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," *Proc. INTERSPEECH*, Sep 2006.

[26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," *Proc. ISMIR*, pp. 287–288, Oct 2002.

[27] M. Goto, "AIST annotation for the RWC music database," *Proc. ISMIR*, pp. 359–360, Oct 2006.