Melody Transcription Framework using Score Information for Noh Singing

Katunobu Itou*[†], Rafael Caro Repetto[†], Xavier Serra[†]
* Faculty of Computer and Information Sciences, Hosei University, Tokyo, Email: it@fw.ipsj.or.jp
[†] Music Technology Group, Universistat Pompeu Fabra, Barcelona

***Abstract*—Not only do novice listeners have difficulty enjoying Noh singing but researchers also have difficulty treating it formally. A major reason is the huge difference between the score information and the acoustics of its execution. This paper proposes melody transcription for Noh singing using score information. The method's design is based on comparative observation among score information, commentary, and the acoustic signals of its execution by multiple performers. The calculation is based on global polynomial regression and modification within the note using score information. According to visual judgment of the plots, the resultant transcriptions fitted the pitch contours well. In addition, the possibility of discovering new unprescribed ornaments using melody transcription was suggested.**

Keywords–Noh singing, Melody transcription, Speaker adaptation, Phone segmentation, Music

I. Introduction

Noh is a traditional Japanese performance art consisting of music, drama, and dance.

In Japan, one can see a Noh performance nearly every day, with many audience members being repeaters. For a first-timer, watching the dance and listening to the music is comprehensible, however understanding the drama and the songs is difficult. Indeed performers' phonation differs from that of modern Japanese. As for the musical aspect, even the principle of the scale is completely different from that of modern music. For these reasons, enjoying Noh requires a certain amount of experience and familiarity.

Moreover, Noh singing seldom becomes a subject of research. One main reason is the greater difference between the acoustic signal of singing and score information, because the pitches of scale notes are not absolute and are changeable even within a single phrase. In previous Noh music research, several studies were concerned with the acoustics of the singing voice [1]–[3]. For the melody of Noh singing, previous research has only dealt with the interpretation of score information contained in vocal books [4], [5]. In the research about the acoustics of melody, there was only a comparison study [6].

Therefore, we propose a framework of melody transcription for Noh singing using score information to help to bridge the gap between the acoustic signal and the score. Visualization of the resultant transcription may help novice listeners to interpret the musical aspect of Noh singing, and researchers may use it to ascertain the melodic line. In addition, without such a framework, importing the recent improvement of computer music research, for example, music information retrieval, and synthesis is difficult.

In section II, the musical aspects of Noh singing are described. In section III, the available information to be utilized for transcribing the melody is described. In section IV, the proposed melody transcription method is described. In section V, the results of the preliminary evaluation are presented, followed by a discussion. Finally in section VI, we conclude the paper.

II. Musical aspects of Noh Singing

Noh music consists of vocal and instrumental parts. For the vocal part, the main actor (*shite*), the second actor(*waki*) and a few subsidiary actors sing the main melody and the chorus is sung by the accompaniment singers(*ji*).

The melody of Noh singing has three modes: the melodic mode (*yowagin*), the dynamic mode (*tsuyogin*), and the speech mode (*kotoba*). The lyrics (also known as script or words) are written in classic Japanese.

The melodic and dynamic modes have their own scale. Both scales have two or three main notes, and each main note has its auxiliary notes. The skeleton of the melody is composed of the main notes, and the next notes are strictly limited by composition rules. In addition, in a phrase, many notes stay at the same pitch. Consequently, the variation of the types of the melodic line on the score is more limited than that of other music.

In contrast, in the execution of the score prescribed melodic line, the scale is not absolute. Even within a single phrase, the same pitch notes under the score information can be sung with a different actual pitch (f0, fundamental frequency), and the difference between different pitch notes is not absolute either. An actual pitch varies depending on singers and varies even within a single piece depending on singing styles. Actual pitch also varies depending on the characters in the drama.

In Noh singing, ornaments are huge and heavily used. Ornaments are categorized into two types: one is prescribed in the score and the other is not prescribed. Examples of prescribed ornaments are *"hon-yuri"*, a type of melisma at the end of phrase, and abrupt rising and falling pitch (float). Examples of unprescribed ornaments are vibrato, sliding from a lower pitch at the beginning of a phrase, and stress at the end of a phrase.

For these reasons, in interpreting Noh music, an expression and/or personality included in actual execution of the melody is more important than the melodic line prescribed in the score. Such an expression and/or personality deviates from the exact execution prescribed by the score. According to these characteristics, for the Noh melody, there is no standard for measuring the accuracy of performances in acoustics. In this study, we propose melody transcription aimed to be utilized as a standard of the melodic line to measure expression or individuality. In order to transcribe from acoustic signals, score information along with the knowledge of interpretation is used.

III. Information for transcribing

There are five *shite* schools and they publish their own vocal books(*utai bon*), which contain all the lyrics of a piece, and the parts of the melodic and dynamic modes have melodic annotations. These annotations prescribe the melodic line similar to scores in Western music. However, some information is implicit and loose, so these annotations have to be interpreted as context dependent or based on common knowledge within Noh music. Vocal books are used for amateur performers' practice; however, they found melodic interpretation difficult, so commentaries were also published [7].

In the commentary, to assist interpretation, graphical notation of the melodic line is used. However, this notation does not assist understanding of the melodic line in the way it does in Western musical scores, because of the difficulty in estimating the exact line from acoustic signals. The commentary includes the graphical notation of 300 phrases from 55 pieces. Figure 1 shows an example of a graphical notation [7] and Figure 2 shows a pitch (f0) contour of its execution. This sample is a phrase in the dynamic mode.

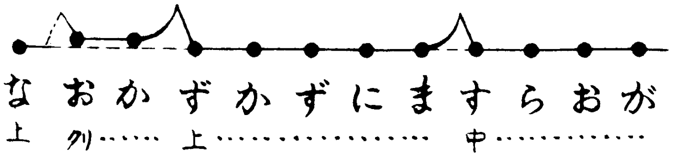

Figure 1. A graphical notation of a melodic information in a Noh vocal book

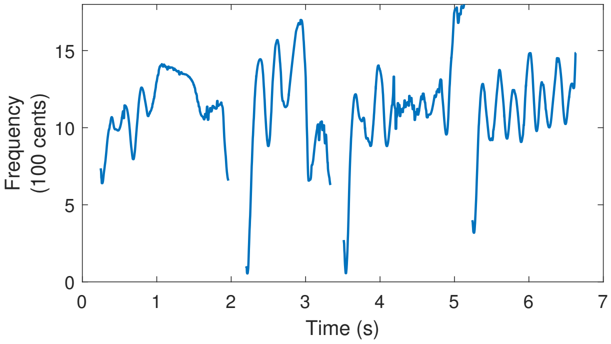

Figure 2. Pitch contours of Noh singing

In Figure 1, the first row is the graphical notation. Bullets show onsets, and lines and curves show pitch transience. In this notation, the different vertical positions indicate different pitches similar to Western music notation. However, this notation is continuous unlike the discrete Western notation. Therefore, ornaments indicated using smaller symbols, e.g, a grace note, are indicated using the combination of curves. The second row contains the lyrics in Japanese *kana*. The last row contains note names. Figure 2 illustrates the difficulty of finding onsets and transitions of notes because of the extent of certain vibrato being greater than the transition pitch difference. By contrast, the extent of vibratos in Western music hardly exceeds 200 cents. Cent is a unit for musical pitch. One semitone is 100 cents. In this paper, n cents for an fHz pitch is $n = 1200 \cdot \log_2 \frac{f}{131}$. As seen in this example,fitting a melodic line to a pitch contour is difficult because a new melodic line notation, closer to the acoustic signal, is required.

To design the new notation, we collected speech signals of phrases in the commentary from a cappella parts in commercial Noh singing compact discs (CDs) and observed their spectrograms and pitch contours. Consequently, we categorized the pitch transition patterns within a single note into the following three: maintaining same pitch (staying), transition to a different pitch, and abrupt rising and falling pitch (floating). In every pitch transition, many ornaments–vibratos, pitch sliding, and others–were observed.

In the next section, we describe the proposed melody transcription method on the basis of the collection and the categorization above. This method uses score information, and its input is assumed to be an a cappella signal.

IV. Noh Melody Transcription using Score Information

A. Input signal

The proposed method assumes a single phrase as an input. Normally, Noh-singing CDs are edited as a whole piece that runs from 10 to 40 min as several tracks or a single track. However, in vocal books, phrases are divided with punctuations, which correspond to rest marks of music. In Noh-singing signals, these punctuations almost correspond to pauses, and such pauses are automatically detected when using a voice activity detection technique of speech recognition.

B. Score information

Each note corresponds to a single syllable. In Japanese, syllable is classified in two types: CV, which means a consonant succeeding a vowel, V, which means just a vowel. Hence, each note has one or two phone fields. In addition, each phone field has pitch transition information, which consists of three values at most. The first is an original and mandatory pitch. The second is a first-transited pitch, and the third is a second-transited pitch. Pitch is expressed as an integer whose value difference refers approximately to a halftone of Western music. Table I is an example of the first three notes in Figure 1.

TABLE I. Score information

syllable	*phone*	*pitch*		
na	n	5		
	a	5	10	6
o	o	6		
ka	k	6		
	a	6	10	5

In Western music, pitch transition occurs at the onset. In Noh singing, pitch transition does not occur quickly at the onset, and the execution is like a grace note [1]. In addition, unlike Western music or Japanese popular music, consonants last longer. Reflecting these aspects, a note must be divided into phones, i.e., a note must contain a phone boundary.

C. Pitch contour estimation

A pitch contour is estimated as an f0 contour using Melodia algorithm [8]. To be adjusted to Noh singing, the following condition is changed. Noh singing has a very low pitch, whose lower values can be less than 100 Hz; hence, we do not apply the lower part of the equal loudness filter [9]. In addition, we do not use voicing detection because the voice quality of Noh singing differs from that of Western music singing [2] and musical instruments.

D. Score alignment

Score alignment is a technique to link score information to audio signals of a score's performance. For a monophonic source, in order to link, the detected onset of note [10] or the shape of pitch contour [11] is used. However, these techniques are not suitable for Noh singing. Many onsets of Noh singing are blurred because each syllable is not uttered separately. As mentioned above, the melodic line of Noh singing often stays at the same pitch, and such a situation, of course, does not change pitch contour shapes. In addition, vibrato depth in Noh singing is not only much greater than that in Western music but is sometimes greater than the note transition of the phrase.

We used phone segmentation based on forced alignment using hidden Markov model(HMM)-based phone models for speech recognition. Using this method, the onset of a note is the beginning time of its syllable. Moreover, the boundary between the consonant and the vowel in the note can be estimated during the slower start of Noh melody transition, which makes this method suitable. Accurate segmentation requires the preparation of acoustic phone models matched with Noh-Singing signals. However, phonation of Noh singing differs from that of Japanese speech [1], and in preliminary experience, segmentation accuracy is not good when using the acoustic phone model for standard Japanese adult speakers [12].

To prepare a phone model that matches Noh singing, we used the maximum likelihood linear regression (MLLR) speaker adaptation technique [13]. The speaker adaptation technique can fit a general-purpose acoustic model to a specific speaker using only a small amount of data from the target speaker. MLLR speaker adaptation estimates linear transformations for parameters of HMM phone models. In this method, the entire transcription is available from a vocal book. Part of the target data can be used as a supervised adaptation data.

Figure 3 exemplifies the phonetic segmentation result of the pitch contour of the first 4 s of Figure 2. Vertical lines indicate onsets. This figure was plotted using manual labeling for a later explanation.

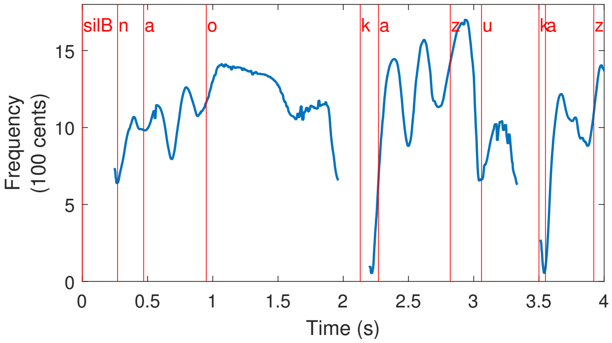

Figure 3. Phone segmentation of pitch contours

E. Melody transcription

Transcribing a melody requires assigning a pitch contour segment to a certain pitch. The relative pitch has to be shifted; however, a shift is not adequate for Noh singing because pitch also changes. A pitch frequency histogram is often used for tonic detection, but, histograms of Noh singing do not have appropriately sharp peaks. Deep vibratos extending wider than the difference between pitches and changeable pitch make the peaks of the histogram broader. Furthermore, unlike Western music, pitch does not change in steps. Therefore, in this study, to transcribe the melody, curves are estimated by polynomial approximation to fit the pitch contour using score information.

This study assumes the following: pitches in a melody are changeable in parallel following the same polynomial coefficients.

1) Initial centered pitch curve: According to the score information and phone segmentation, each phone is assigned one of the three categories: staying, transition, or floating. Using only staying segments, each mean for the pitch in the score (a) is calculated as μ_a. In this step, segments of unvoiced consonants are not used because the pitch estimation may not be accurate.

Then, pitch contour is centered by subtracting μ_a from the pitch contour y. The centered pitch contour y_c is fitted by polynomial regression of degree d. In this study, by preliminary experience, d is determined to be from four to six according to the number of phones Here, let the fitted curve be p. Figure 4 shows the centered contour and the regression curve for the contour of Figure 3. In this phrase, $d = 4$. Figure 4 is plotted in linear scale for frequency.

The pitch for all its segments is not staying, and thus, its mean cannot be calculated. For such pitches, the mean is calculated using the value of the nearest calculated pitch by subtracting the default scale difference value. The scale value is correspondent with 100 cents. For example, in Table I, if μ_{10} was not calculated, μ_{10} is calculated as $\mu_{10} = \mu_6 * 2^{(10-6)/12}$.

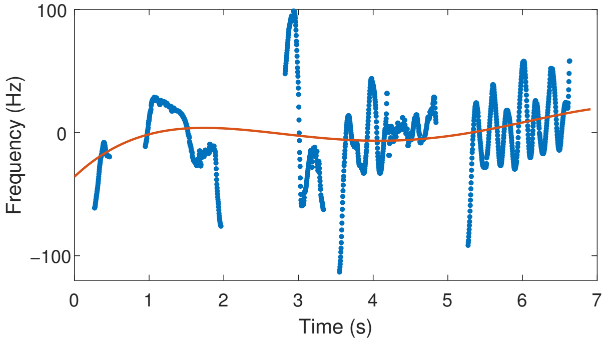

Figure 4. An initial centered pitch contour (dot plot) and an initial centered curve (solid line)

2) Intra-note processing: In Noh singing, note transition time is not strictly determined; hence, if the score directs the pitch to ascend at the i-th note n_i, the pitch ascent delays at the $i + 1$-th note n_{i+1} in the execution [7]. Therefore, in the following process, if necessary, the search range can be extended.

The pitch transition of Noh singing does not start at the beginning of syllables [1]. Adjusting to this knowledge, frames before the beginning of transition are treated the same as staying notes.

a) Floating: Floating is observed as a huge peak in pitch contour. Let the current note be n_i. The initial search

ISBN: 978-1-61208-464-0

range is the vowel region of n_i. For the reason mentioned above, let the search range extend to the next note as $[n_i, n_{i+1}]$. To determine such a peak, first, we subtract the curve corresponding to the pitch of the note in the score a_i from the pitch contour y. Let the result be d. Hence, $d = y - p_i$, $p_i = p + \mu_{a_i}$

Then, let the highest and the nearest of the border between n_i and n_{i+1} peak be a floating transition. Let the intersections between the peak of the pitch contour and p_i be c_1, c_2. The frames before c_1 and after c_2 in the range are treated as staying notes. If there are no intersections for the peak, let the first two nearest local minima of the peak be c_1, c_2. If the pitch contour vanished, let the end point of the contour be c_1, c_2.

b) Transient: Let the current note be n_i. The initial search range is the vowel region of n_i. First, calculate the polynomial curve corresponding to the initial pitch p_{i1} and the final pitch p_{i2}, the same as above p_i. Let the intersections between y and p_{i1} be $s = s_1, \ldots, s_K$, and the intersections between y and p_{i1} be $e = e_1, \ldots, e_L$. If necessary, y is interpolated in search range. Let the first e_m, which is greater than s_k, be the final point. Then, $y(s_k, e_m)$ is fitted by polynomial regression of degree d_t. In this study, d_t is determined as 2. The frames before s_k and after e_m in the range are treated as staying notes.

If there is no s, the search range is extended to the preceding consonant region of n_i. If there is no e, the search range is extended to the next note. Then, s and/or e is calculated again. In addition, if there is also no s, let the first point of the search range be the initial point. Also, if there is no e, let the last point of the search range be the final point.

F. Re-estimation of centered pitch curve and intra note processing

Consequent to intra-note processing, the regions of staying notes are changed. Using this changed data, re-estimate the centered pitch curve. In this step, outliers are ignored using the threshold of three σ because the staying notes' region may include unprescribed ornaments. Then, intra-note processing is executed again.

V. Results and Discussion

Figure 5 shows an example of melody transcription of the pitch contour of Figure 2 using manually labeled phone segmentation. The bullets indicate the onsets of the note. The transcription can be considered as a smooth fit of the graphic notation in Figure 1 to the pitch contour. From this transcription, we can observe the difference between the score and the execution. For example, in the score, the first float is in the first note; however, in the transcription, the float is in the second note. The second and third floats cross the border of the note.

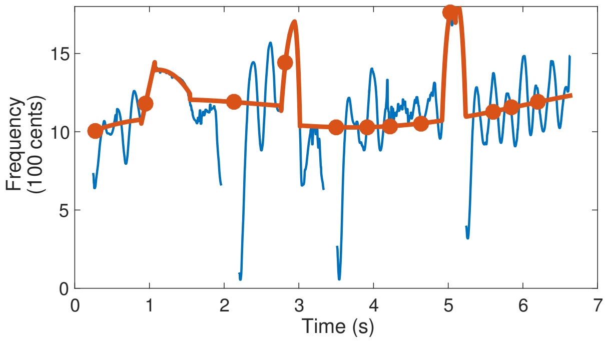

Figure 5. A melody transcription. The blue line is f0 contour. The red line is the melody transcribed with the onset timing using the bullets.

Figure 6 shows the executions of the same phrase by different performers. Both transcriptions share the similar shape but the onsets are different. This sample shows a merit of the proposed method, because such different onsets correspond to different scores in the discrete notation, similar to a score of Western music. In addition, by observing the pitch contours residual to the transcriptions, the vibrato types differ. In the first plot, vibrato is asymmetric to the melodic contour. However, in the second plot, the vibrato is more symmetrical than that in the first. In addition, there is a kind of peak similar to float after the dips at approximately 6 s in both transcriptions. Asymmetric vibrato is one of the most specific characteristics of Noh singing, and is considered as an unprescribed ornament. This suggests that new ornaments will be discovered using this method.

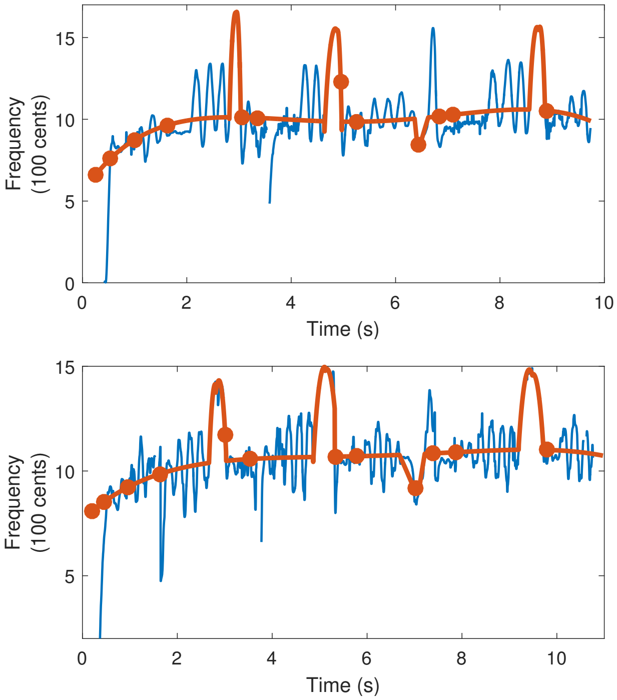

Figure 6. Comparison of the transcriptions of different singers

In Figures 5 and 6, at the beginning of the phrase, pitch contours ascend from much a lower pitch to the target pitch. This ornament is suggested in another commentary [14]. By gaining the degree of the polynomial, transcription better fits such an ornament; however, the gained degree causes an outfit at the other end, there is a lot of literary commentary, which is informal, subjective and ambiguous. Such ornaments are required in quantification, and this transcription will be an aid to that process.

We evaluate the melody transcription accuracy using 17 phrases (from 6 different pieces) in the commentary [7]. In total, we tested 23 audio phrases by four singers. As a general HMM, we used a monophone model trained using read speech of 98 h duration collected from 361 speakers [12]. If a single singer sings an entire target piece, all the data of the target piece can be used as the speaker adaptation data. If multiple singers sing a target piece, two cases will exist. If there is any other pieces that a target singer sings entirely, we can use the adapted model by the data from the piece. If there are no other pieces that a target singer can sing completely, we can use an adapted model of another piece sung by another singer.

Regarding the phone segmentation accuracy, the absolute error compared with the hand-labeled onset was 0.37 s by the general HMM and 0.089 s by the adapted HMMs. Figure 7 shows an example of melody transcription using highly erroneous phone segmentation from the upper pitch contour in Figure 6. The average absolute onset error was 0.23 s. Comparing the data in Figure 6, the first float was missing and the transient just after 6 s was missing. For both cases, successive vowels caused segmentation errors, for example the error of the vowel in the first float was 1.56 s and the error near the transient was 0.39 s, due to the phonation difference in modern Japanese speech. Such vowel differences were reduced by MLLR adaptation.

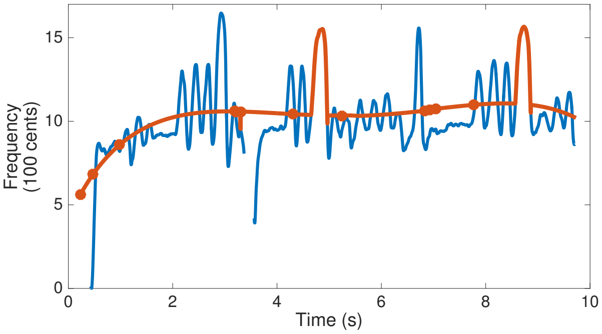

Figure 7. A melody transcription using erroneous segmentation

We evaluated this type of transcription error using the average absolute pitch error and compared it to the transcription estimated using manually-labeled phone segmentation in the cent domain. The average absolute pitch error of the transcription in Figure 7 was 50 cents. The average absolute pitch error was 24 cents by the general HMM segmentation and 13 cents by the adapted HMM segmentation. Figure 8 shows the relationship between the phone segmentation error and the transcription estimation error. The adapted model did not estimate erroneous transcription. For modern or Western music, the onset detection accuracy is evaluated within 50 ms window [10], but for Noh singing more broader window, e.g., 200 ms, might be appropriate for evaluation, because Noh singing is very slow, e.g., the average phone duration was approximately 400 ms in the evaluation data, and the onset is flexible and not as important as it is in Western music.

VI. Conclusion

Here, the first melody transcription framework to reflect acoustics for Noh singing is proposed. It was achieved by comparative observations among score information, commentary, and acoustic signals of its multiple executions. The calculation is based on global polynomial regression and modification within the note using score information. According to visual judgment of the plots, the resultant transcription fitted the pitch contour well, and the proposed method's continuous notation was suitable for the flexible nature of Noh singing. Moreover, the potential of discovering new unprescribed ornaments by melody transcription is suggested. For automatic transcription estimation, the speaker adaptation technique using Noh singing audio data as adaptation data was effective.

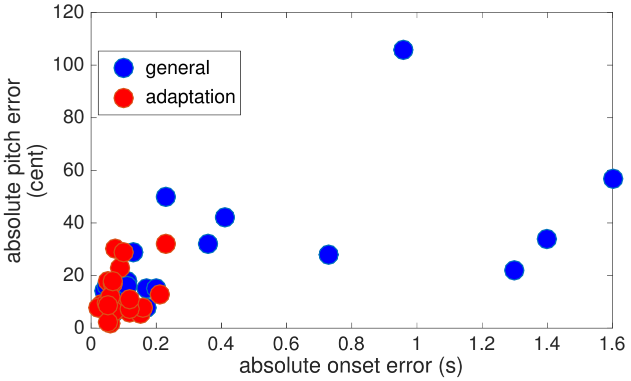

Figure 8. The relationship between the segmentation error and the transcription error

One of the most important remaining issues is evaluation. We are planning to interview professional Noh performers about the concept of melody transcription and the resultant transcriptions.

References

[1] I. Nakayama, "Comparison of vocal expressions between Japanese traditional and western classical-style singing, using a common verse," The Journal of the Acoustical Society of Japan, vol. 56, no. 5, 2000, pp.343–348. (in Japanese)

[2] O. Fujimura, et. al., "Noh voice quality," Logopedics Phoniatrics Vocology, vol. 34, 2009, pp. 157–170.

[3] I. Yoshinaga and J. Kong, "Laryngeal Vibratory Behavior in Traditional Noh Singing," Tsinghua Science and Technology, vol. 17, no. 1, 2012, pp. 94–103.

[4] T. Minagawa, "Japanese "Noh" music," Journal of the American Musicological Society, vol. 10, no. 3, 1957, pp. 181–200.

[5] I. Takakuwa, "*Noh/Kyogen utai no hensen*," Hinoki Shoten, Tokyo, Japan, 2015. (in Japanese)

[6] Z. Serper, "*Noh no kotoba no yokuyo*," Engeki Kenkyu, vol. 36, 2013, pp. 51–80. (in Japanese)

[7] K. Miyake, "*Fushi no seikai*(new revised edition)," Hinoki Shoten, Tokyo, Japan, 2012. (in Japanese)

[8] J. Salamon and E. Gómes, "Melody extraction from polyphonic music signals using pitch contour characteristics," IEEE Transactions on Audio, Speech, & Language Processing, vol. 20, no. 6, 2012, pp. 1759–1770.

[9] D. Robinson, "Equal loudness filter," 2015, URL: http://replaygain.hydrogenaud.io/proposal/equal_loudness.html [accessed: 2015-11-25].

[10] J. P. Bello, et. al., "A tutorial on onset detection in music signals," IEEE Transactions on Audio, Speech, & Language Processing, vol. 13, no. 5, 2005, pp. 1035–1047.

[11] N. H. Adams, M. A. Bartsch, J. B. Shifrin, and G. H. Wakefield, "Time series alignment for music information retrieval," Proceedings of ISMIR-04, 2004, pp. 303–310

[12] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," Proceedings of ICSLP2004, 2004, pp. 3069–3072

[13] C. L. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol. 9, 1995, pp. 171–185.

[14] M. Yokomichi, “*Nogaku kogi note (utai hen)*,” Hinoki Shoten, Tokyo, Japan, 2013. (in Japanese)

APPENDIX

CD/WMA(WINDOWS MEDIA AUDIO) FILE LIST

- WMF files (All titles are distributed by *Hinoki-Shoten*
  - *Hagoromo*, KANZE Motomasa (solo)
  - *Kiyotsune*, KANZE Motoaki with others
  - *Tamura*, KANZE Motomasa with others
  - *Yashima*, KANZE Motoaki with others
  - *Momijigari*, KANZE Motomasa with others
  - *Hashibenkei*, KANZE Motomasa (solo)
  - *Touboku*, KANZE Motomasa with others
- CDs (*Kanze-ryu Utai Nyumon* CD series (All titles are distributed by kanze.com, sung by KANZE Yoshimasa (solo))
  - *Hagoromo*
  - *Tamura*
  - *Momijigari*
  - *Tsurukame*
  - *Hashibenkei*
  - *Touboku*
- CDs
  - *Hagoromo*, OOE Matasaburo with others, (*Nohgaku Meibankai*)
  - *Kiyotsune*, UMEWAKA Naoyoshi (solo), (*Nohgaku Meibankai*)
  - *Tamura*, KANZE Kiyokazu with others, (*Toei Sound Family*)
  - *Tamura*, FUJINAMI Shigemitsu with others, (Columbia)

ISBN: 978-1-61208-464-0