Regular Paper

HamoKara: A System that Enables Amateur Singers to Practice Backing Vocals for Karaoke

Mina Shiraishi[1,a)] Kozue Ogasawara[1,b)] Tetsuro Kitahara[1,c)]

Received: January 24, 2019, Accepted: September 11, 2019

Abstract: Creating harmony in karaoke by a lead vocalist and a backing vocalist is enjoyable, but backing vocals are not easy for non-advanced karaoke users. First, it is difficult to find musically appropriate submelodies (melodies for backing vocals). Second, the backing vocalist has to practice backing vocals in advance in order to play backing vocals accurately, because singing submelodies is often influenced by the singing of the main melody. In this paper, we propose a backing vocals practice system called *HamoKara*. This system automatically generates a submelody with a rule-based or probabilistic-model-based method, and provides users with an environment for practicing backing vocals. Users can check whether their pitch is correct through audio and visual feedback. Experimental results show that the generated submelodies are musically appropriate to some degree, and the system helped users to learn to sing submelodies to some extent.

Keywords: back vocal, submelody generation, probabilistic model, karaoke

1. Introduction

Karaoke is one of the most familiar forms of entertainment related to music. At a bar or a designated room, a person who wants to sing orders the karaoke machine to play back the accompaniment of his/her favorite songs. During the playback of the accompaniment, he/she enjoys singing. Because karaoke is enjoyable for many people, techniques enhancing the entertainability of karaoke have been developed [2], [3], [4].

One way to enjoy karaoke is to create harmony with two singers. Typically, one person plays *lead vocals* (singing the main melody) and the other person plays *backing vocals* (singing a different melody). The melody sung by the backing vocalist (we call this a *submelody* here) typically has the same rhythm as the main melody but has different pitches (for example, a third above or below from the main melody). If the lead vocalist and backing vocalist are able to sing in harmony with each other, the sound will be pleasant. However, this is not easy in practice. First, it is difficult to find a musically appropriate submelody. Particularly for songs that do not contain backing vocals in the original CD recordings, the backing vocalist has to create an appropriate submelody him/herself, which requires musical knowledge and skill. Second, accurately playing backing vocals in pitch requires advanced singing skill. If the two singers do not have sufficient singing skill, the pitch of the backing vocals may be influenced by the lead vocals. To avoid this, the backing vocalist needs to learn to sing accurately in pitch by practicing the backing vocals in advance.

In this paper, we propose a system called *HamoKara*, which enables a user to practice backing vocals. This system has two functions. The first function is automatic submelody generation. For popular music songs, the system generates submelodies and indicates them with both a piano-roll display and a guide tone. The second function is support of backing vocals practice. While the user is singing the indicated submelody, the system shows the pitch (fundamental frequency, F0) of the singing voice on the piano-roll display. Because we adopted almost the same interface design as used in the converntional karaoke machine, the user can easily use our system. The user can find out how close the pitch of his/her singing voice comes to the correct pitch.

The rest of the paper is organized as follows: in Section 2, we present a brief survey on related work from the perspectives of harmonization and singing training. In Section 3, we describe the overview of our system. In Section 4, we present the details of our automatic submelody generation methods. In Section 5, we report results of evaluations conducted to confirm the effectiveness of our system. Finally, we conclude the paper in Section 6.

2. Related Work

2.1 Harmonization

Harmonization has two general approaches. The output of the first approach is a sequence of chord symbols, such as C-F-G7-C, for a given melody [5], [6], [7]. The second approach assigns concrete notes to voices other than the melody voice. The typical form in the latter approach of harmonization is a four-part harmony, which consists of soprano, alto, tenor, and bass voices. Four-part harmonization is a traditional part of

[1] College of Humanities and Sciences, Nihon University, Setagaya, Tokyo 156–8550, Japan
[a)] shiraishi@kthrlab.jp
[b)] ogasawara@kthrlab.jp
[c)] kitahara@kthrlab.jp

The content of this paper has been presented at the 15th Sound and Music Computing Conference [1]. A video related to this paper is available at: https://www.youtube.com/watch?v=C2_P7-jMwC8 the theoretical education of Western classical musicians, and so numerous researchers have attempted to generate it automatically [8], [9], [10], [11], [12], [13], [14], [15].

The feature of our task is the creation of a two-voice harmony under the existence of the accompaniment. This is because, in karaoke, one sings together with the accompaniment played back by the karaoke machine. Submelodies should therefore be consonant with the accompaniment, typically given as MIDI-equivalent sequence data, as well as with the main melody.

A very similar problem was dealt with by Kiribuchi et al. [16]. To compose duet pieces automatically, they developed a method for yielding submelodies to a piece composed by Orpheus [17]. Because Orpheus generates accompaniment tracks, the submelody generator should consider the consonance of the submelody with the accompaniment. Although the details are not presented in Ref. [16], we suppose that the information of conditions for composition, such as chord progression, is also available in the submelody generator. On the other hand, our input is MIDI tracks, in which the chord transcription is not assumed to exist. We, therefore, have to extract harmonic features from the given MIDI tracks.

2.2 Singing Practice Support

As commonly known, almost all commercial karaoke machines have a singing scoring function. For example, in US Patent 5719344 [18], a technique has been proposed that is based on comparing the karaoke user's voice and a referential voice extracted from the original CD recording. In addition, Tsai et al. [19] also proposed an automatic evaluation method of karaoke singing.

Computer-assisted singing learning systems have been developepd, such as Singad, Albert,Sino & See, andWinSINGAD (See [20] for details). These systems aim at self-monitoring the user's singing with visual feedback of the pitch, spectra, and/or other acoustic/musical features.

In addition, Hirai et al. [21] developed the Voice Shooting Game to improve poor-pitch singing. In this game, the vertical position of the user's character on the screen is synchronized with the pitch of the user's singing voice. The user has to control his/her pitch accurately to control the vertical position of the character and to catch target items.

Nakano et al. [22] developed a singing-skill visualization system called *MiruSinger*. This system extracts and visualizes acoustic features, particularly the pitch and vibrato sections, from the user's singing voice.

Lin et al. [23] developed a real-time interactive application that runs on a smartphone, which enables the user to learn the intonation and timing while singing.

These studies and systems focus on solo singing, not on situations where two persons sing together in harmony. Fukumoto et al. [24] developed a system, RelaPitch, that enables the user to train relative pitch. Listening to a standard tone, the user tries to sing the pitch that has a specified interval from the standard tone. This system focused on training the creation of harmony by singing, but does not focus on creating harmony in actual songs.

3. System Overview

3.1 Basic System Design

The goal of our system is to enable the user to practice backing vocals for a song from popular music (especially Japanese popular music, or J-pop). To achieve this, the system has to support the following functions:

(1) Submelody generation
Some songs do not have submelodies in the original CD recordings. For such songs, the system has to generate submelodies automatically.

(2) Practice support through audio and visual feedback
The submelody for a specified song is indicated through the piano-roll display and the guide tone during the playback of the accompaniment. It is not easy for most users to recognize whether the singing voice has the correct pitch (i.e., the same pitch as the guide tone). The system should therefore show the user how accurate his/her singing pitch is through audio and visual feedback.

Our system assumes that music data (main melodies, accompaniments, and lyrics of songs) are given in the MIDI format, because most modern karaoke machines use MIDI-equivalent data.

Given the MIDI file of the target song, a submelody is generated for the entire main melody (from beginning to end). The submelody is assumed to have the same rhythm as the main melody; hence, each note in the submelody completely corresponds to each note in the main melody one-for-one. Submelodies are often sung only at limited sections (such as chorus sections) in actual music scenes, but our system generates a submelody for the entire main melody. This is so that users can simply ignore the submelody indication except in the sections where they want to sing harmony.

It is possible to generate submelodies that have higher or lower pitches than the main melody. The user can select an *upper submelody* or a *lower submelody* at the start screen.

A typical phenomenon occurs when one person plays lead vocals and another person plays backing vocals: here, their singing pitches are influenced by each other. In particular, backing vocals may be greatly influenced by lead vocals because the backing vocalist has to sing an unfamiliar melody. One solution to this problem is to enable the backing vocalist to check his/her pitch accuracy during the practice phase. Our system enables the backing vocalist to control the volume balance between his/her own voice and the guide tone in order to easily check his/her voice. In addition, he/she can find the pitch accuracy through visual feedback. The pitch of his/her voice and the correct pitch are displayed on the piano-roll screen in real time.

Because karaoke is a well-established form of entertainment, we thought the interface design should be as close as possible to the conventional karaoke. Therefore, we designed the user interface (particularly the screen) based on commonly used karaoke machines. In the lower half of the screen, the lyrics of the song being played are displayed. The color of the lyrics changes with the progression of the song's playback. In the upper half of the screen, the piano roll is displayed to show the user the pitch of his/her singing voice and the correct pitch. This display was made

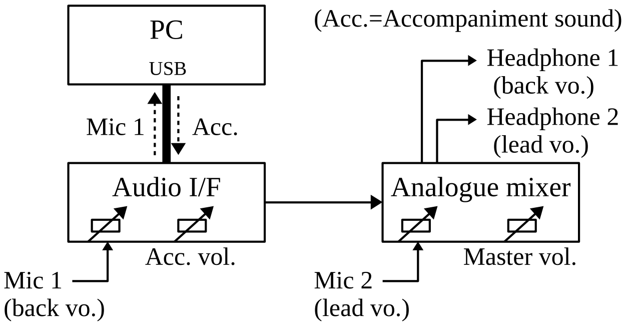


Fig. 1 Block diagram of our system. Our system consists of a PC, an audio interface, and an analogue mixer as well as microphones and headphones. The audio interface has separate volume controls for the backing vocals and accompaniment sound.

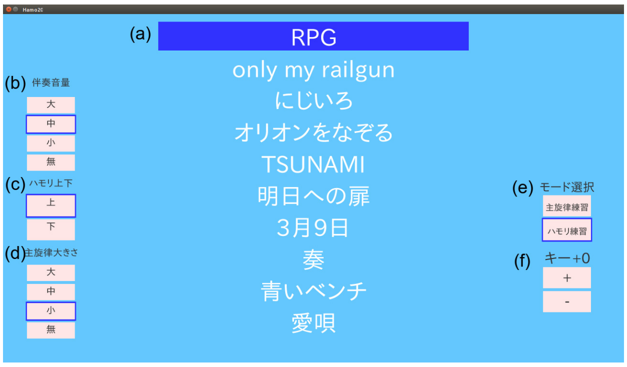


Fig. 2 Start screen. (a) Song selection, (b) accompaniment volume, (c) upper submelody or lower submelody, (d) guide tone volume, (e) practice of the main melody or submelody, (f) key transposition.

as large as possible because it is the most important indication for singing practice. Although moving pictures are an important factor in enhancing the entertainment of karaoke, we omitted this because it is not necessarily needed for singing practice. The details of the screen design are described in Section 3.4.

3.2 System Configuration

The block diagram of our system is shown in **Fig. 1**. Loading and playing back MIDI data, generating submelodies, and displaying lyrics and submelodies on the screen are performed by the PC. The microphone for the backing vocals is connected to the USB audio interface communicating with the PC. The audio signal of the backing vocals is sent from the audio interface to the PC, and the accompaniment sound is sent from the PC to the audio interface. The volume of the backing vocals and the accompaniment can be adjusted independently on the audio interface.

In contrast, the microphone for lead vocals is connected to the analogue mixer, not to the audio interface. This is in order to not send the audio signal of the lead vocals to the PC. To make it possible to listen to both lead vocals and backing vocals, the headphones for both vocalists are connected to the analogue mixer. If only a backing vocalist uses this system without any lead vocalist, the analogue mixer is not necessary.

3.3 Start Screen

Once the system is launched, the screen shown in **Fig. 2** appears. The user then selects a song to be played back. Then, the user selects an *upper submelody* or a *lower submelody*, how many

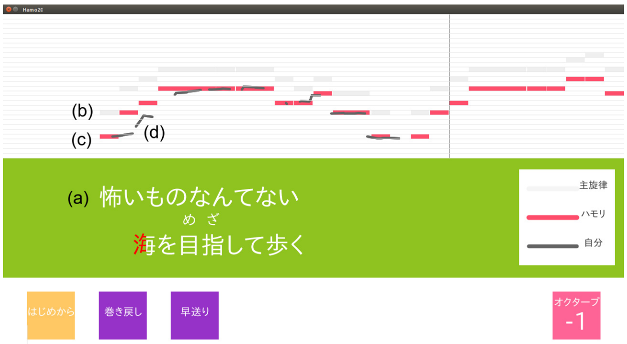


Fig. 3 Main screen. (a) Lryics, (b) the piano roll of the main melody, (c) the piano roll of the generated submelody, (d) the pitch of the user's singing voice.

semitones the key is transposed by (the degree of transposition (if necessary)) and adjusts the volume of the guide tone of the main melody and the accompaniment sound.

3.4 Main Screen

After the user selects a song and other settings, a submelody is generated for the song. Then, the playback of the accompaniment starts. During the playback, the lyrics are displayed in a typical karaoke display style (**Fig. 3**). In the upper half of the screen, the main melody and the submelody are displayed on the piano-roll screen.

The user's singing voice is analyzed in real time. Every 5 ms, the pitch (fundamental frequency, F0) of the user's singing voice is estimated with the DIO F0 estimator [25] and displayed on the piano-roll screen.

Every four measures, the ratio is calculated of the frames in which the difference between the singing pitch and correct pitch is smaller than 200 cent. In this calculation, the frames are excluded when this difference is greater than 800 cent, along with silent frames, because it could be caused by a double-pitch or half-pitch estimation error. If this ratio is lower than 50%, the message "Listen to your voice more carefully" is displayed on the bottom of the screen. This message is prepared to urge the user to turn the volume of his/her voice up.

4. Submelody Generation Methods

The most simple method for generating a submelody is to generate the note of the major or minor 3rd below each note of the main melody (in the case of a *lower* submelody). However, which of the 3rds below is suitable depends on the harmony of the accompaniment. Sometimes the perfect 4th below is more suitable because both the major and the minor 3rd below may cause dissonance with the accompaniment. To select an appropriate note interval between the main melody and submelody, we propose two methods: a probabilistic-model-based (shortened to *prob-based* below) method and a rule-based method.

4.1 Prob-based Method

Let $n_1, \cdots, n_m$ and $h_1 \cdots, h_m$ be sequences of notes of the main melody and submelody, respectively. Each n_i and each h_i take an integer between 0 and 11, which corresponds to a pitch class (C, C♯, $\cdots$, B). The relative degree of appearance of each pitch class

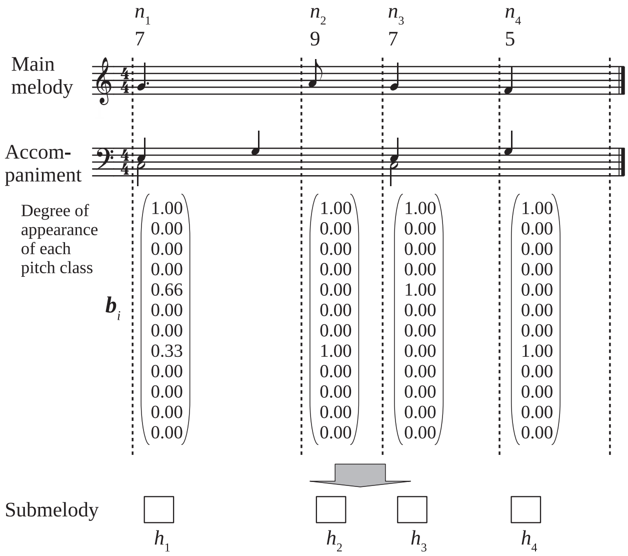


Fig. 4 Pitch class of the main melody $\{n_i\}$, pitch class of a submelody $\{h_i\}$, and the relative degree of each pitch class in the accompaniment $\{\boldsymbol{b}_i\}$. This example has only one part in the accompaniment for simplicity, but actually the accompaniment consists of more-than-one parts.

in the accompaniment from the onset to the offset of each note n_i is calculated and is denoted as a 12-dimensional vector $\boldsymbol{b}_i$; each element of this vector takes a value between 0.0 and 1.0 (**Fig. 4**). Let k be the key of the target song [*1].

What to calculate here is $h_1, \cdots, h_m$ maximizing $g(h_1, \cdots, h_m)$ given by

$$g(h_1, \cdots, h_m) = P(h_1, \cdots, h_m \mid n_1, \cdots, n_m, \boldsymbol{b}_1, \cdots, \boldsymbol{b}_m, k).$$

Using Bayes' theorem, this conditional probability can be written as follows:

$$\begin{aligned} &g(h_1, \cdots, h_m) \\ &= \frac{P(n_1, \cdots, n_m, \boldsymbol{b}_1, \cdots, \boldsymbol{b}_m \mid h_1, \cdots, h_m, k)\ P(h_1, \cdots, h_m \mid k)\ P(k)}{P(n_1, \cdots, n_m, \boldsymbol{b}_1, \cdots, \boldsymbol{b}_m, k)}. \end{aligned}$$

Considering that we optimize the conditional probability with only respect to $(h_1, \cdots, h_m)$, $g(h_1, \cdots, h_m)$ satisfies:

$$\begin{aligned} &g(h_1, \cdots, h_m) \\ &\propto P(n_1, \cdots, n_m, \boldsymbol{b}_1, \cdots, \boldsymbol{b}_m \mid h_1, \cdots, h_m, k)\ P(h_1, \cdots, h_m \mid k). \end{aligned}$$

We assume $h_1, \cdots, h_m$ and $n_1, \cdots, n_m$ to have Markov property. Leaving out the dependency between observable variables (n_i and b_i), this can be written as

$$\begin{aligned} &g(h_1, \cdots, h_m) \\ &\propto P(h_1|k)P(n_1|h_1)P(\boldsymbol{b}_1|h_1) \prod_{i=2}^{m} P(h_i|h_{i-1}, k)P(n_i|h_i)P(\boldsymbol{b}_i|h_i). \end{aligned}$$

Then, we decompose $P(h_i|h_{i-1}, k)$ into $P(h_i|h_{i-1})$ and $P(h_i|k)$ with log-linear interpolation as follows:

$$P(h_i|h_{i-1}, k) = \frac{1}{w}\ P(h_i|h_{i-1})\ P(h_i|k),$$

where w is a coefficient for letting the summation of the combined distribution be 1. Thus, the formula to be maximized can be described as

$$\begin{aligned} &g'(h_1, \cdots, h_m) \\ &= P(h_1|k)P(n_1|h_1)P(\boldsymbol{b}_1|h_1) \prod_{i=2}^{m} P(h_i|k)P(h_i|h_{i-1})P(n_i|h_i)P(\boldsymbol{b}_i|h_i). \end{aligned}$$

We can maximize $g(h_1, \cdots, h_m)$ by maximizing this $g'(h_1, \cdots, h_m)$ using dynamic programming.

Below, we describe the details of each factor of this probability.

- $P(h_i|k)$ is a probability for the note name of the submelody's i-th note given key k. This probability is defined as

$$P(h_i|k) = \begin{cases} 0.15 & (h_i \text{ is a tonic, mediant, or dominant note for key } k) \\ 0.10 & (h_i \text{ is one of the other diatonic notes for key } k) \\ 0.03 & (h_i \text{ is not a diatonic note for key } k) \end{cases}$$

- $P(h_i|h_{i-1})$ is a probability for note transitions in the submelody. We assume the tendency of note transitions in the submelody to be close enough to that of note transitions in the main melody. We then calculate this probability from the main melody of the target song.
- $P(n_i|h_i)$ represents the relationship between the main melody and submelody. Both melodies tend to have an interval of the major 3rd, minor 3rd, or perfect 4th. We then calculate this probability based on the following equation:

$$P(n_i|h_i) = \begin{cases} \alpha & (|n_i - h_i| = 0 \text{ (perf 1st)}) \\ 2\alpha & (|n_i - h_i| = 1 \text{ (min 2nd)}) \\ 4.5\alpha & (|n_i - h_i| = 2 \text{ (maj 2nd)}) \\ 15\alpha & (|n_i - h_i| = 3 \text{ (min 3rd)}) \\ 20\alpha & (|n_i - h_i| = 4 \text{ (maj 3rd)}) \\ 15\alpha & (|n_i - h_i| = 5 \text{ (perf 4th)}) \\ 0 & (|n_i - h_i| = 6 \text{ (arg 4th)}) \\ 5\alpha & (|n_i - h_i| = 7 \text{ (perf 5th)}) \\ 0 & (|n_i - h_i| \geq 8) \end{cases}$$

  where α is a coefficient to make the summation of the probabilities equal to 1.
- $P(\boldsymbol{b}_i|h_i)$ represents the relationship between the submelody and the accompaniment. As described above, the 12-dimensional vector $\boldsymbol{b}_i = (b_{i,0}, \cdots, b_{i,11})$ represents the relative degree of appearance of each pitch class in the accompaniment from the onset to the offset of the main melody's i-th note n_i. By assuming each element in $\boldsymbol{b}_i$ to be independent of each other, the above probability can be reduced as follows:

$$P(\boldsymbol{b}_i|h_i) = \prod_{j=0}^{11} P(b_{i,j}|h_i),$$

  where $P(b_{i,j}|h_i)$ is an emission probability for how apparent the pitch class j is in the accompaniment (i.e., an observation) given the submelody's pitch class h_i (i.e., a hidden state). When $|i - j| = 1$ (minor 2nd) or 2 (major 2nd),

[*1] For simplicity, our modeling does not consider modulation. For a song involving modulation, we treat sections before and after modulation as separate songs.

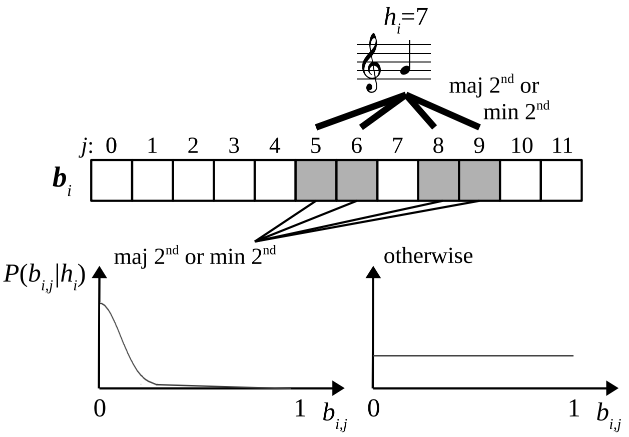


Fig. 5 Probability distribution of $P(\boldsymbol{b}_i|h_i)$.

the appearance of the pitch class j should not be high in the accompaniment because it will cause dissonance. We therefore define $P(b_{i,j}|h_i)$ as following a normal distribution $\mathcal{N}(b_{i,j}; 0.0, \sigma_1^2)$ (σ_1^2 is experimentally determined). Otherwise, we consider $P(b_{i,j}|h_i)$ to follow a continuous uniform distribution [*2] (**Fig. 5**).

4.2 Rule-based Method

Like the prob-based method, $n_1, \cdots, n_m$ and $h_1, \cdots, h_m$ represent sequences of notes of the main melody and submelody, respectively, and each variable takes an integer between 0 and 11, corresponding to a pitch class. From the accompaniment, a sequence of 12-dimensional vectors $\{\boldsymbol{b}_i\}$ is extracted in the same way as the prob-based method.

For each i, the highest four values from the 12 elements of $\boldsymbol{b}_i = (b_{i,0}, \cdots, b_{i,11})$ are extracted and the pitch classes corresponding to them are obtained. The set of the obtained pitch classes are denoted here by C_i. In the case of lower submelodies, the pitch class of i-th note of the submelody, h_i, is determined by the following rules:

(1) If C_i includes the pitch class of the major 3rd below n_i, h_i is set to this pitch class ($h_i = n_i - 4$).

(2) If C_i includes the pitch class of the minor 3rd below n_i, h_i is set to this pitch class ($h_i = n_i - 3$).

(3) If C_i includes the pitch class of the perfect 4th below n_i, h_i is set to this pitch class ($h_i = n_i - 5$).

(4) Otherwise, h_i is determined so that it has a parallel motion with the main melody from the previous note ($h_i = h_{i-1} + (n_i - n_{i-1})$).

Rules for upper submelodies are similarly defined.

5. Experiments

5.1 Evaluation of Harmonization

5.1.1 Conditions

Table 1 Rates of concordance (ROCs) between generated submelodies and ground truth [%].

	Prob-based			Rule-based		
	Ave	Max	Min	Ave	Max	Min
Upper	52.3	64.5	24.3	55.5	75.0	33.6
Lower	53.6	71.0	30.5	64.2	87.9	37.6

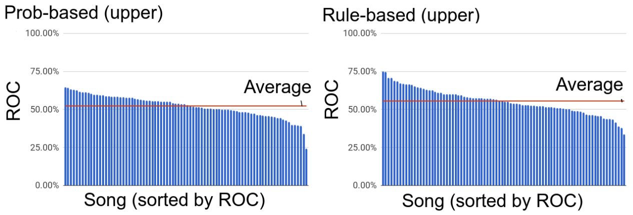


Fig. 6 ROC for each song (upper submelodies).

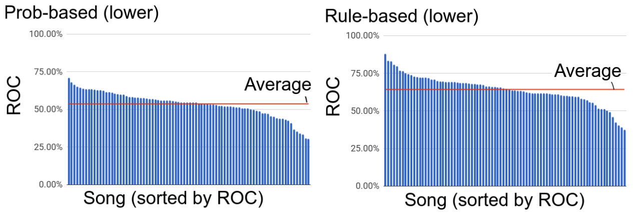


Fig. 7 ROC for each song (lower submelodies).

We asked a graduate student at Japan's top-level music university to add submelodies (both upper and lower submelodies) to the MIDI data of 85 J-pop songs. The MIDI data were purchased at Yamaha Music Data Shop (https://yamahamusicdata.jp/). The MIDI data sold here consist of the main melody (in Track 1), the accompaniment (in other tracks), and lyrics text data (as MIDI Meta Events).

We calculated the rates of concordance (ROCs) between these ground truth submelodies and the generated submelodies. Note that generated submelodies are not necessarily musically inappropriate even if they are not concordant with the ground truth because there can be more-than-one musically appropriate submelodies for the same song even though we have only one ground truth for each song.

5.1.2 Results

The ROCs are listed in **Table 1**. The average rates are 52.3% for upper submelodies and 53.6% for lower submelodies with the prob-based method; while these are 55.5% for upper submelodies and 64.2% for lower submelodies with the rule-based method. **Fig. 6** and **Fig. 7** show the ROC for each song. We can see that the ROCs vary among songs in all conditions.

5.1.3 Discussions

We consider two reasons why the prob-based method is inferior to the rule-based method. The first reason is frequent appearance of unison (the same note is selected for both main melody and submelody; that is, $n_i = h_i$). In fact, 4.78% of the generated submelodies are in unison. In the submelodies created by the expert (the ground truth), only 0.20% notes are in unison (unison is not generated by the rule-based method) . When $n_i = h_i$, $P(n_i|h_i)$ is set to α ($\approx$ 0.016), which is sufficiently smaller than $P(n_i|h_i)$ given that $|n_i - h_i| = 3, 4, 5$. A possible reason why unison appeared despite such a low emission probability could be transition probabilities were not appropriately estimated.

The second reason of the rule-based method's superiority is that the same pitch class is consecutively repeated with the prob-based method. The numbers of appearances of the repetition

[*2] In the actual implementation, we use a normal distribution $\mathcal{N}(b_{i,j}; 0.5, \sigma_2^2)$ with a sufficiently large constant σ_2 as an approximation of a continous uniform distribution because the programming library we used supports only normal distributions.

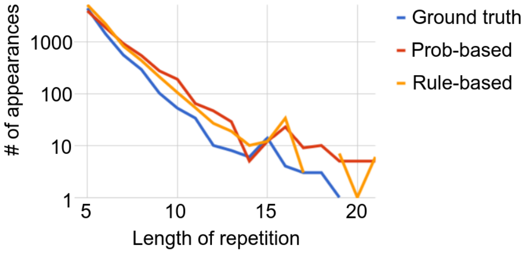


Fig. 8 The total number of appearances of repetitions of the same pitch class in 85 upper submelodies.

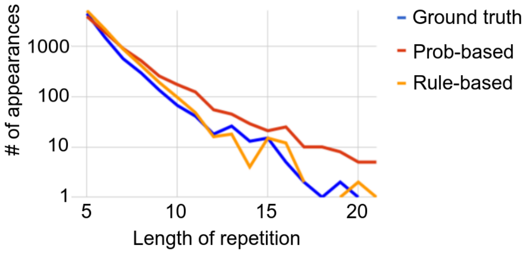


Fig. 9 The total number of appearances of repetitions of the same pitch class in 85 lower submelodies.

of the same pitch class are shown in **Fig. 8** and **Fig. 9**. From these figures, we can see that, for example, a sequence of ten consecutive notes with the same pitch class appear at about 200 points in the 85 upper submelodies generated with the prob-based method. This occurs because the probability of self-transition $P(h_i = x \mid h_{i-1} = x)$ is higher than $P(h_i = y \mid h_{i-1} = x)$ (for all $y \neq x$). This could becaused by data sparseness: the transition probabilities were calculated from the main melody of the target song only. In 30 songs for which the generated submelody include 10 or more consecutive notes with the same pitch class, 10 or more pitch classes have the highest transition probability for self-transition. In the other 25 songs, 5 or 6 pitch classes have a higher transition probability for non-self-transition than that for self-transition.

5.1.4 Subjective Evaluation by an Expert

We discussed submelody generation performance based on the ROCs above, but low ROCs do not necessarily mean musical inappropriateness. We therefore asked an expert (an associate professor at a university of music) to subjectively evaluate the generated submelodies. Specifically, she checked musically inappropriate notes by listening to the submelodies with the main melodies and the accompaniments. To avoid burdening her, only the five-lowest ROC songs were evaluated for each method.

The results are listed in **Table 2**. The rates of musically appropriate notes fall between 67% and 97%; hence, it was confirmed that our submelody generation methods are appropriate, at least to some extent.

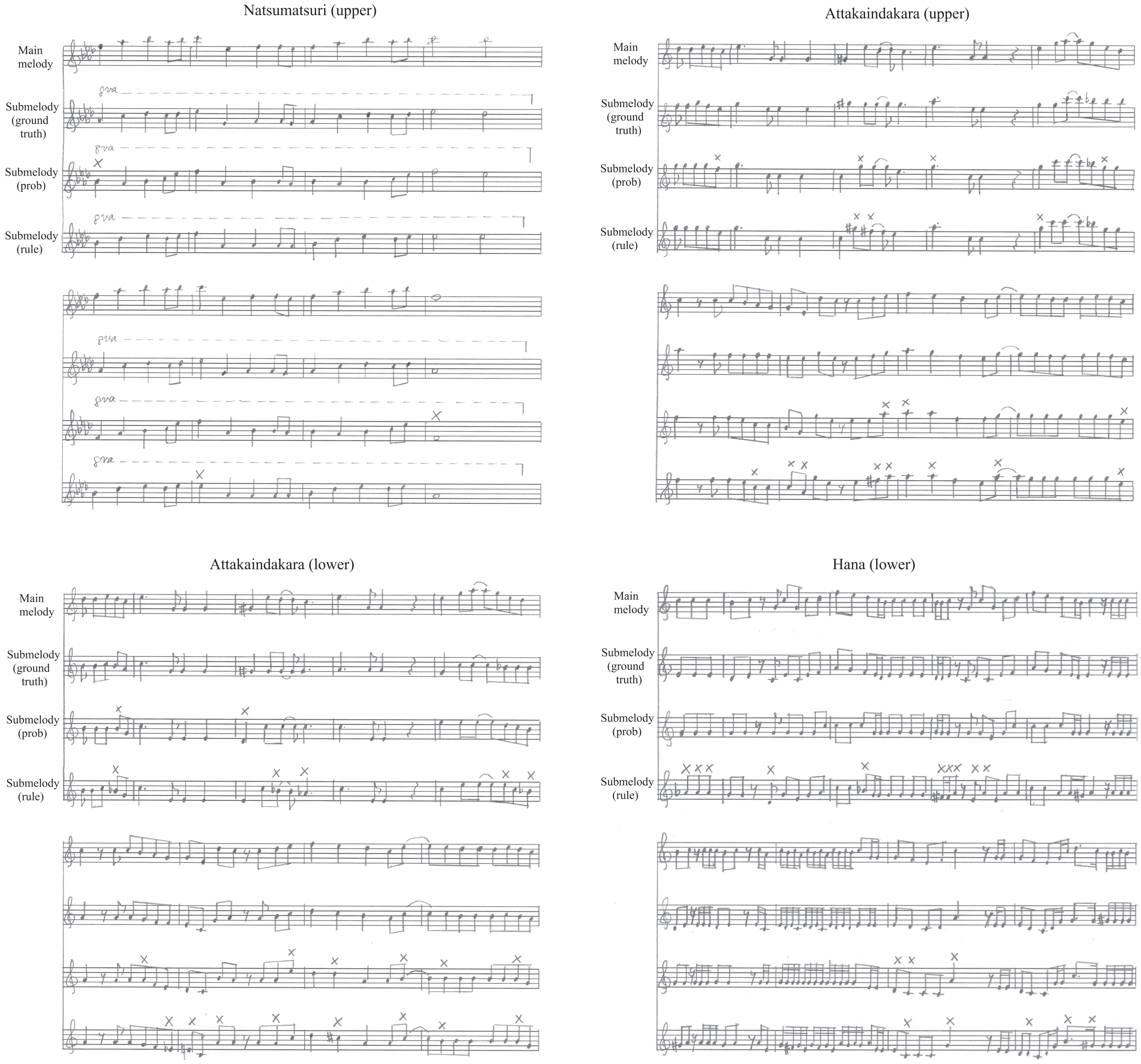
Table 2 Results of subjective evaluation by a musical expert. Rate 1: ROC to the ground truth; Rate 2: the rate of appropriate notes judged by the expert.

Method	Song	Rate 1	Rate 2
Prob	Natsumatsuri (upper)	24.36%	81.58%
	Attakaindakara (lower)	30.51%	83.24%
	Osaka Lover (lower)	30.98%	73.80%
	Hana (lower)	33.45%	97.46%
	365 Nichi (lower)	34.09%	79.81%
Rule	Natsumatsuri (lower)	33.64%	85.53%
	Hana (lower)	37.56%	90.30%
	Aoi Benchi (upper)	37.77%	85.33%
	Attakaindakara (upper)	38.98%	67.03%
	Attakaindakara (lower)	39.23%	78.92%

5.1.5 Example-based Qualitative Discussion

We quantitatively discuss differences between generated submelodies and ground truth based on some examples. Four examples are shown in **Fig. 10**. Our observations are summarized as follows:

- *Natsumatsuri* (*upper*)
  The ground truth of the upper submelody for this song consists of only the note of the major or minor 3rd above each note in the main melody. In this sense, it is a simple submelody, but it is characterized by the fact that most notes included in the submelody are not chord notes but 7th notes. Use of 7th notes for submelodies enriches the harmony, so this approach is commonly used. On the other hand, the generated submelodies use many notes of the perfect 4th above the main melody (e.g., B♭ for F, E♭ for B♭, and F for C) because notes for submelodies are preferentially selected from chord notes. To achieve the approach using 7th notes like the ground truth, we need to model, for example, the richness of a harmony, not only avoiding dissonance.
- *Attakaindakara* (*upper*)
  The ground truth of the submelody mainly consists of notes of the 3rd above the main melody but also includes notes of the 6th above. In addition, the submelody sometimes uses oblique motions (in the 2nd, 4th, and 5th measures). On the other hand, all notes except one note in the prob-based and rule-based submelodies are notes of the major 3rd, minor 3rd, or perfect 4th above the main melody. This could be one reason why some musically unnatural notes were generated. In particular, the rule-based method tends to select the note of the major 3rd above preferentially, so it sometimes selects such notes (e.g., F♯ in the 7th measure) even though they are musically unnatural.
- *Attakaindakara* (*lower*)
  Similarly to the upper submelody for the same song, the ground truth includes both parallel motions with 3rd intervals and oblique motions (in the 2nd and 4th measures). The prob-based submelody uses chord notes where passing notes are better (e.g., the last B in the auftakt, the last A in the 5th measure, and the last B in the 8th measure). Indeed, these notes were judged to be musically inappropriate in the subjective evaluation (Section 5.1.4). In the rule-based submelody, the problem of musically unnatural notes of the major 3rd from the main melody is more severe than the lower submelody (B♭ in the auftakt, B♭ and A♭ in the 2nd measure, and E♭ in the 5th measure).
- *Hana* (*lower*)
  Almost all notes in the ground truth of the submelody are 6th below the main melody. They are just one octave below the upper submelody. The prob-based submelody mainly consists of notes of 4th below the main melody and the G note is repeated. This submelody does not cause strong dissonance, but it is slightly monotonous because it almost consists of only G, C, and E. The rule-based submelody for this


Natsumatsuri (upper)

Attakaindakara (upper)

Attakaindakara (lower)

Hana (lower)

Fig. 10 Examples of generated submelodies (cross symbols "×" are attached to notes considered musically inappropriate by the human evaluator in Section 5.1.4). The melodies of these songs were written by Hashi Jinta (*Natsumatsuri*), Kumamushi (*Attakaindakara*), and Orange Range (*Hana*).

song also suffered from the problem of musically unnatural notes of the major 3rd from the main melody (e.g., A♭ in the auftakt and B♭ in the 2nd measure).

To summarize, at least in these four examples, the prob-based method rarely generates dissonant submelodies, but tends to fail to select passing notes. The rule-based method sometimes selects notes of the major 3rd from the main melody that cause dissonance.

5.2 Evaluation of Effects of Audio/Visual Feedback

We conducted an experiment to confirm how the audio and visual feedback functions are effective in the practice of backing vocals.

5.2.1 Used Systems

The used systems are listed in **Table 3**. Because the purpose of the experiment is to confirm the effects of audio and visual feedback functions, we compared our system with systems in which either of the feedback functions was disabled. To reduce the participants' burden, we omitted the comparison with the system in which both feedback functions were disabled.

Table 3 Systems used in the experiments (*proposed).

	Audio feedback	Visual feedback
System A*	Enabled	Enabled
System B	Enabled	Disabled
System C	Disabled	Enabled

5.2.2 Participants

The participants were 12 university students (male: 5, female: 7). Two of the participants had more than five years of piano experience, but the others had no particular musical experience. No one had special singing experience.

5.2.3 Procedure

Every participant was asked to use the three systems listed in

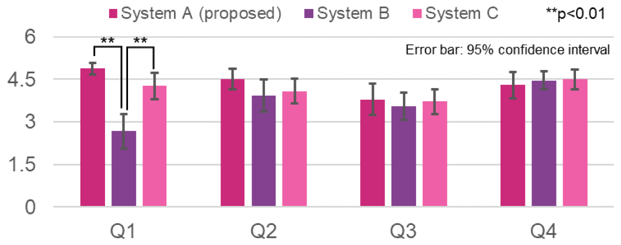
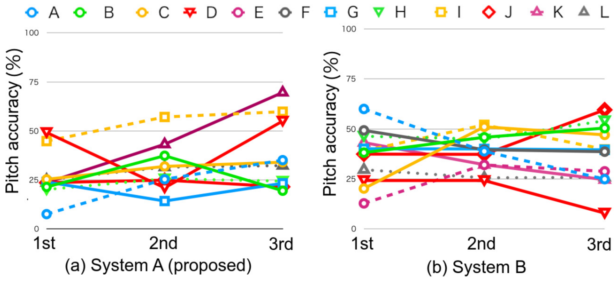


Fig. 11 Results of questionnaire in Section 5.2.



(a) System A (proposed)

(b) System B

Fig. 12 Pitch accuracy in the experiment in Section 5.2.

Table 3. For each system, participants did the following:

(1) Repeat the following three times
   (a) Practice Song 1 with the given system for two minutes
   (b) Sing Song 1 as if really singing in karaoke (non-practice singing)
(2) Answer a questionnaire
(3) Repeat the following three times
   (a) Practice Song 2 with the given system for two minutes
   (b) Sing Song 2 as if really singing in karaoke (non-practice singing)
(4) Answer the same questionnaire

To prevent the effect of getting used to the songs and the systems, we used different songs for different systems and counterbalanced the order of the systems among the participants. In non-practice singing, the experimenter (the second author) sang the main melody together with the participant's submelody singing; the visual feedback function was disabled (the audio feedback was enabled) regardless of the given system for the practice phase. To obtain sufficient accuracy of pitch estimation, the participant and the experimenter used headphones in all cases.

In the questionnaire, we asked each participant the following four questions:

Q1 Could you grasp the pitch of your voice accurately?

Q2 Do you think that the practice phase was useful for the non-practice singing?

Q3 Could you practice the submelody singing and prevent it from being influenced by the main melody?

Q4 Was your voice influenced by the main melody's voice in the non-practice singing?

The participants answered these questions on a scale of 0 to 6. Lower ratings are better only for Question 4. The ratings were tested using the Wilcoxon rank-sum test. The Bonferroni correction was used for multiple comparisons.

5.2.4 Used Songs

We used six famous J-pop songs: "Sakuranbo" (Ai Ohtsuka), "Natsu No Hi No 1993" (Class), "Glamorous Sky" (Mika Nakashima), "Asu No Tobira" (I WiSH), and "RPG" (Sekai No Owari). To minimize the burden on participants, we used only the chorus sections of these songs. To prevent the effect of the difference in lyrics, we asked the participant to sing them by "la-la-la."

5.2.5 Results —Questionnaire—

The results (**Fig. 11**) can be summarized as follows:

Q1 The ratings for System A are significantly superior to those for System B with a significance level of 1%. This means that the visual feedback of the pitch of users' singing is to some extent effective in helping them grasp the pitch accuracy of their voices.

Q2 No significant differences among the three systems are shown.

Q3 Also for this question, no significant differences among the systems are shown. This occurs because our system plays the main melody with a guide tone, not a singing voice. Therefore, our system could not create a situation in which singing the submelody is influenced bythe main melody in the practice phase. This could be improved by introducing a singing synthesis engine, such as Vocaloid.

Q4 This question also shows no significant difference among the systems. However, four participants commented, in an interview after the experiment, that they were less influenced by the main melody when they practiced with System A than with the other systems. This implies a possibility that our system is effective in preventing the user's singing from being influenced by the main melody.

Although we did not conduct a systematic evaluation for the interface design, all participants were able to use our system without confusion.

5.2.6 Results —Pitch Accuracy—

We analyzed how the pitch is improved by comparing pitch accuracy in the first, second, and third non-practice singing phases of the same song. The same method as the pitch estimator was used for the audio feedback. Specifically, the pitch of the singing voice is esimated every 5 ms, and the ratio of the frames in which the difference from the correct pitch is smaller than 200 cent is calculated. The silent frames and the frames in which the difference is greater than 800 ms are excluded in calculating this ratio. Only for System A, two participants were excluded from this analysis because more than half of the frames of their singing voices were excluded.

The results are shown in **Fig. 12**. By comparing the pitch accuracy at the first and third singing, we can see that six of the 10 participants (60%) improved the pitch accuracy by more than 5% with System A. On the other hand, five of the 12 participants (42%) improved with System B.

5.3 Evaluation of Key Transposition

When the lead vocalist's pitch range does not match the range of the main melody of the target song, he/she may transpose the key to match those pitch ranges. If this is the case, the backing vocalist also has to sing the submelody in the transposed key. However, it is difficult to sing the submelody in a different key from the key in the practice phase. Our system therefore enables

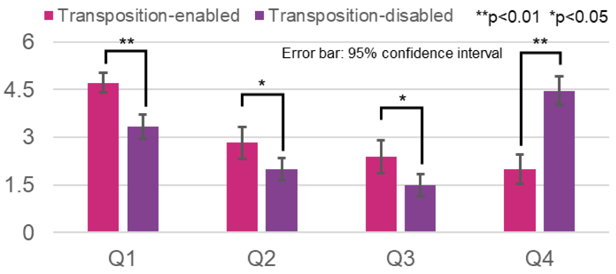


Fig. 13 Results of questionnaire in Section 5.3.

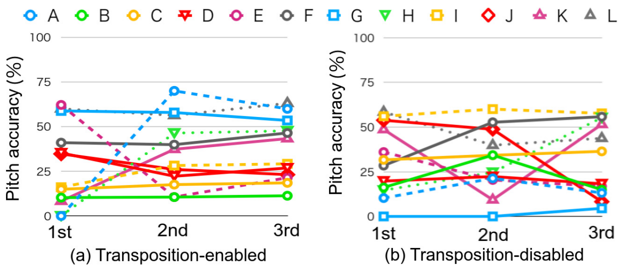


Fig. 14 Pitch accuracy in the experiment in Section 5.3.

the user to practice backing vocals in any key by transposing the key. In this procedure, we confirm the effects of this transposition function.

5.3.1 Participants

The participants are 13 university students (male: 6, female: 7) who did not participate in the procedure in Section 5.2. Six participants have more than five years ofpiano experience, and one participant has more than five years ofbass experience.

5.3.2 Procedure

On each of the *transpostion-enabled* and *transposition-disabled* conditions, every participant did the same procedure as Section 5.2 (however, here they sang three songs).

During the non-practice singing, the key was transposed up a major 2nd in both conditions. During the practice phase, the key was set to the original key in the transposition-disabled condition, while the participants freely selected the keyin the transposition-enabled condition. Because they were informed in advance that they had to sing in the major-2nd-above key at the non-practice singing phase, all participants selected this key.

The questionnaire consists of the following questions:

Q1 Do you think that the practice phase was useful for the non-practice singing?

Q2 Do you think that you sang well in the practice phase?

Q3 Was it easy to sing submelodies in the non-practice singing?

Q4 Did you experience difficulty singing the submelody in a non-original key (a key different from the original CD recording) in the non-practice singing?

The participant answered these questions on a scale of 0 to 6. Lower ratings are better only for Question 4. The ratings were tested using the Wilcoxon rank-sum test.

5.3.3 Used Songs

We used the same songs as Section 5.2 in the same way.

5.3.4 Results —Questionnaire—

The results (**Fig. 13**) can be summarized as follows:

Q1–Q3 The ratings in the transposition-enabled condition were on average at least 1.3 higher than in the transposition-disabled condition, and the significance of the differences were confirmed with a significance level of 5% (for Q1, even with 1%). This shows that practice is more effective when the same key used in the non-practice singing phase is selected.

Q4 The rating in the transposition-enabled condition was on average at least 2.0 lower than in the transition-disabled condition, and the significant difference was shown with a significance level of 1%. This implies that it is not easy to accurately sing submelodies in a non-original key after practicing the submelodies in the original key.

5.3.5 Results —Pitch Accuracy—

We analyzed the improvement of pitch accuracy during the non-practice singing in the same way as Section 5.2.5. One participant (the same participant in both conditions) was excluded, because more than half of the frames of this participant's singing were excluded.

The results are shown in **Fig. 14**. When we compare the first and third singing, four of the 12 participants markedly improved their pitch accuracy (more than 10%) in the transposition-enabled condition, while two participants improved in the transcription-disabled condition.

6. Conclusions

In this paper, we proposed a system to practice backing vocals for karaoke. To enable karaoke users to learn backing vocals, we had to resolve two issues: automatic generation of submelodies, and support of backing vocals practice through audio and visual feedback. Through experiments, we confirmed that our system resolved these issues to some extent.

However, we still have a number of issues. There is still room for improvement on the parameter tuning of our model. We also should try data-driven learning of the model parameters. In addition, the experiments were short due to considering the participants' burden; longer experiments are necessary to investigate how the backing vocal skill is being improved.

Acknowledgments This research was partly supported by JSPS KAKENHI Grant Nos. JP16K16180, JP16H1744, JP16KT0136, and JP17H00749.

References

[1] Shiraishi, M., Ogasawara, K. and Kitahara, T.: HamoKara: A System for Practice of Backing Vocals for Karaoke, *Proc. 15th Sound and Music Computing Conference* (*SMC 2018*), pp.511–518 (2018).

[2] Cano, P., Loscos, A., Bonada, J., de Boer, M. and Serra, X.: Voice Morphing System for Impersonating in Karaoke Applications, *Proc. ICMC* (2000).

[3] Daido, R., Hahm, S.-J., Ito, M., Makino, S. and Ito, A.: A System for Evaluating Singing Enthusiasm for Karaoke, *Proc. ISMIR* (2011).

[4] Kurihara, T., Kinoshita, N., Yamaguchi, R. and Kitahara, T.: A Tambourine Support System to Improve the Atmosphere of Karaoke, *Proc. SMC* (2015).

[5] Hu, J., Guan, Y. and Zhou, C.: A hierarchical approach to simulation of the melodic harmonization process, *IEEE International Conference on Intelligent Computing and Intelligent Systems*, Vol.2, pp.780–784 (2010).

[6] Kawakami, T., Nakai, M., Shimodaira, H. and Sagayama, S.: Hidden Markov Model Applied to Automatic Harmonization of Given Melodies, *IPSJ SIG Notes*, 99-MUS-34, pp.59–66 (2000). (in Japanese).

[7] Miura, M., Kurokawa, S., Aoi, A. and Yanagida, Y.: Yielding harmony to given melodies, *Proc. 18th International Congress on Acoustics*,

pp.3417–3420 (2004).

[8] Ebciolu, K.: An expert system for harmonizing chorales in the style of J.S. Bach, *Journal of Logic Programming*, Vol.8, pp.145–185 (1990).

[9] Hild, H., Feulner, J. and Menzel, W.: HARMONET: A neural net for harmonizing chorales in the style of J.S. Bach, *Advances in Neural Information Processing*, Vol.4, pp.267–274 (1991).

[10] Pachet, F. and Roy, P.: Formulating constraint satisfaction problems on part-whole relations: The case of automatic harmonisation, *Workshop at European Conference on Artificial Intelligence* (1998).

[11] Allan, M. and Williams, C.K.I.: Harmonising chorales by probabilistic inference, Vol.17, pp.25–32 (2005).

[12] Phon-Amnuaisuk, S., Smaill, A. and Wiggins, G.: Chorale harmonization: A view from a search control perspective, *Jounal of New Music Research*, Vol.35, pp.279–305 (2006).

[13] Yi, L. and Goldsmith, J.: Automatic generation of four-part harmony. Conference on Uncertainty in Artificial Intelligence-Applications Workshop, Vol.268 (2007).

[14] Buys, J. and van der Merwe, B.: Chorale harmonization with weighted finite-state transducers, *23rd Annual Symposium of the Pattern Recognition Association of South Africa*, pp.95–101 (2012).

[15] Suzuki, S. and Kitahara, T.: Four-part Harmonization Using Bayesian Networks: Pros and Cons of Introducing Chord Nodes, *Journal of New Music Research*, Vol.43, No.3, pp.331–353 (2014).

[16] Kiribuchi, D., Fukayama, S., Saito, D. and Sagayama, S.: Automatic Duet Composition from Japanese Lyrics, *Proc. 75th National Cnvention of IPSJ*, Vol.2, pp.303–304 (2013). (in Japanese).

[17] Fukayama, S., Nakatsuma, K., Sako, S., Nishimoto, T. and Sagayama, S.: Automatic Song Composition from the Lyrics exploiting Prosody of the Japanese Language, *Proc. Sound and Music Computing Conference*, pp.299–302 (2010).

[18] Pawate, B.: Method and System for Karaoke Scoring (1995), US Patent 5719344.

[19] Tsai, W.-H. and Lee, H.-C.: Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features, *IEEE Trans. Audio, Speech, Lang., Procss.*, Vol.20, No.4, pp.1233–1243 (2012).

[20] Hoppe, D., Sadakata, M. and Desain, P.: Development of real-time visual feedback assistance in singing training: A review, *Journal of Computer Assisted Learning*, Vol.22, pp.308–316 (2006).

[21] Hirai, S., Katayose, H. and Inokuchi, S.: Clinical Support System for Poor Pitch Singers, *IEICE Trans. Inf. & Syst.*, Vol.J84-D-II, No.9, pp.1933–1941 (2001). (in Japanese).

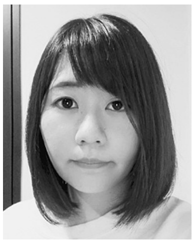
[22] Nakano, T., Goto, M. and Hiraga, Y.: MiruSinger: A Singing Skill Visualization Interface Using Real-Time Feedback and Music CD Recordings as Referential Data, *Proc. IEEE International Symposium on Multimedia*, pp.75–76 (2007).

[23] Lin, K.W.E., Anderson, H., Hamzeen, M. and Lui, S.: Implementation and Evaluation of Real-Time Interactive User Interface Design in Self-learning Singing Pitch Training Apps, *Proc. ICMC-SMC*, pp.1693–1697 (2014).

[24] Fukumoto, A., Hashida, M. and Katayose, H.: RelaPitch: A Relative Pitch Training System With Singing, *Proc. Entertainment Computing Symposium*, pp.332–333 (2016).

[25] Morise, M., Kawahara, H. and Katayose, H.: Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech, *AES 35th International Conference* (2009).

Mina Shiraishi received her B.S. degree from Nihon University in 2018. When she was at the university, she was engaged in music computing research.



Kozue Ogasawara received her B.S. degree from Nihon University in 2018. When she was at the university, she was engaged in music computing research.

Tetsuro Kitahara received his B.S. degree from Tokyo University of Science in 2002 and his M.S. and Ph.D. degrees from Kyoto University in 2004 and 2007, respectively. From 2005 to 2007, he was a JSPS Research Fellow (DC2). From 2007 to 2010, he was a PostDoc researcher at Kwansei Gakuin University. Since 2010, he has been a faculty member at Nihon University, where he is currently an Associate Professor. His research interests include sound and music computing. He received several awards including the 2nd Kyoto University President's Award. He is a member of IPSJ, IEICE, and ASJ.