A SINGING INSTRUMENT FOR REAL-TIME VOCAL-PART ARRANGEMENT OF MUSIC AUDIO SIGNALS

Yuta Ojima[1], Tomoyasu Nakano[2], Satoru Fukayama[2], Jun Kato[2], Masataka Goto[2], Katsutoshi Itoyama[1] and Kazuyoshi Yoshii[1]

[1] Graduate School of Informatics, Kyoto University, Japan

[2] National Institute of Advanced Industrial Science and Technology (AIST)

{ojima, itoyama, yoshii}@sap.ist.i.kyoto-u.ac.jp

{t.nakano, s.fukayama, jun.kato, m.goto}@aist.go.jp

ABSTRACT

This paper presents a live-performance system that enables a user to arrange the vocal part of music audio signals and to add a harmony or troll part by modifying the original vocal part. The pitches and rhythms of the original singing voices can be manipulated in real time while preserving the lyrics. This leads to an interesting experience in which the user can feel as if the original singer is directed to sing a song as the user likes. More specifically, a user is asked to play a MIDI keyboard according to the user's favorite melody. Since the vocal part is divided into short segments corresponding to musical notes, those segments are played back one by one from the beginning such that the pitch and duration of each segment matches a key push by the user. The functions needed for this system are singing voice separation, vocal-part segmentation, key-to-segment association, and real-time pitch modification. We propose three kinds of key-to-segment association to restrict the degree of freedom in manipulation of pitches and durations and to easily generate a harmony or troll part. Subjective experiments showed the potential of the proposed system.

1. INTRODUCTION

We enjoy music not only in a passive way but also in an active way, such as arranging or playing existing songs with a musical instrument. By arranging and playing songs, we can substitute for any part in the music a favorite melody or rhythm pattern. Today, more and more people arrange a part in an existing song, play the arranged part overlaying the original song, and upload the performance to video hosting services such as YouTube.

Arrangement systems for existing songs have already been proposed [1, 2]. These systems help a user arrange an instrumental part in an existing song. Similarly, there exists a situation we hope to arrange a vocal part in existing songs. In fact, many amateur singers uploads vocal covers called "Me singing" on video hosting service [1]. If one does not sing well enough to substitute his or her singing voice for the original singing voice, one needs to make use of a singing voice synthesis system such as *VOCALOID* [3]. Such a system, however, requires lyrics and pitch information in advance, and this information is not easy to get.

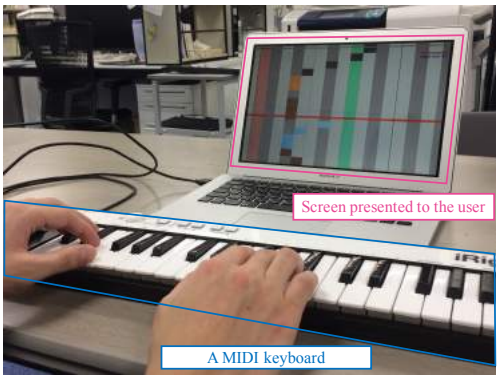

Figure 1. Example of how the proposed system is used. The original vocal part is visualized on a screen, and a user modifies it by using a MIDI keyboard.

This paper proposes a vocal arrangement system that directly manipulates the original singing voice extracted from existing songs in real time (Fig. 1) [2]. There are two advantages in using the original singing voice. The first is that the arrangement is realized without preparing lyrics. The second is that the individuality of the singer is preserved after the arrangements. By preserving the individuality, this system provides an experience as if the original artist sang in a way as a user likes. The arrangements we focus on are pitch modification and changes of onset/offset timing, and we need our system to let users manipulate the pitches, onset times, and durations of notes simultaneously and intuitively. We therefore use a MIDI keyboard as the user interface. This lets users perform singing voice arrangement as if they were playing a piano.

To realize this system, we tackle three problems: estimation and visualization of musical notes in a vocal part, appropriate assignment of a musical note to the key played by a user, and real-time resynthesis of the singing voice after its pitch modified. For the first problem, we first extract an F0 trajectory of the singing voice by using robust principle component analysis (RPCA) [4]. Then, we estimate mu-

[1] More than 98 million "Me singing" videos are uploaded on a popular video sharing service *YouTube*.

[2] A demo video of the proposed system is available online: http://sap.ist.i.kyoto-u.ac.jp/members/ojima/smc2017/index.html

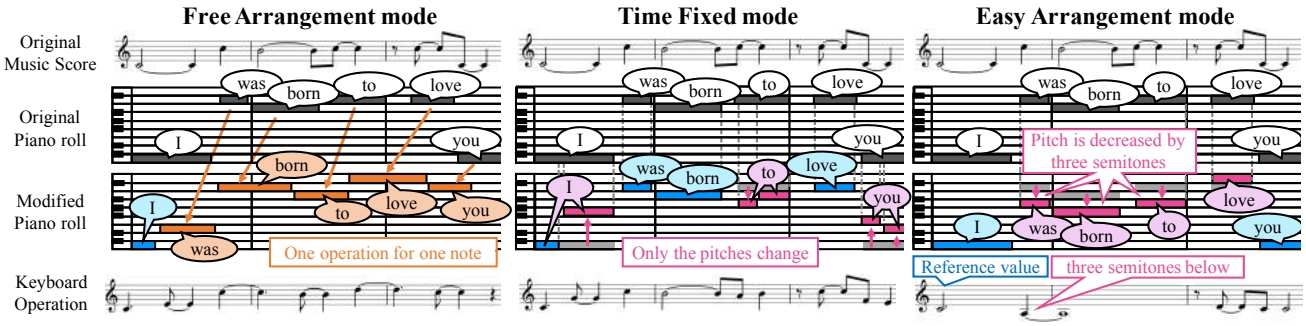

Figure 2. The description of three assignment modes. Each balloon represents the lyrics assigned to a note. In the Free Arrangement mode, the note in the original vocal part is assigned one by one from the beginning for every key operation. In the Time Fixed mode, only the pitches can be changed and the onset/offset timings are not changed. In the Easy Arrangement mode, a user modifies the pitch by specifying its interval relative to the original melody.

sical notes from the F0 trajectory to visualize pitch information by using a Hidden Markov Model (HMM) [5]. For the second problem, we prepare three assignment modes to make it easy to reflect users' various intentions for arrangement (Fig. 2). Since these assignment modes can be switched during the performance, a user can perform as they likes. For the third problem, we use the pitch synchronous overlap and add (PSOLA) algorithm because the algorithm is simple enough to work in real time. Since the arrangement is performed in real time using a MIDI keyboard, this system can be used for a live performance as a DJ plays a song.

2. RELATED WORK

In this section, we consider work related to different aspects of our vocal arrangement system. We first review various music applications and then we focus on arrangement systems because our system aims to arrange an existing music by using signal processing methods. At the end of this section, we review the signal processing methods that are deeply relevant to our system.

2.1 Music Applications

Our system supports active music listening by visualizing the vocal part and by replacing the melody with the one a user likes. Regarding active music listening, several studies have been made. Goto *et al.* [6] proposed a Web service called *Songle* that enriches music listening experience by visualizing the music structure, beat structure, melody line, and chords. Mauch *et al.* [7] proposed an interface called *Song Prompter* that displays the structure of a song, chords, and lyrics without preparing musical scores or manually aligning the lyrics and chords with the audio. Nakra *et al.* [8] designed an interactive conducting system that enables a user to change the tempo and dynamics of an orchestral recording by *Wii Remote*.

On the other hand, another aspect of our system is that it is a music generation system. Many music generation support systems have been proposed. Simon *et al.* [9] proposed one that generates a chord progression to accompany the given vocal melody. Using this system, users who are not familiar with musicography can easily try composing a musical piece. McVicar *et al.* [10] proposed a system that composes a guitar tablature when a chord progression is given as an input. Since this system needs tablature and chords given in advance for training, the generated guitar tablature may vary according to the data used in training. This contributes to the generation of guitar tablature that reflects individuality underlying in the training data. Methods that generate the harmony part have also been proposed. Yi *et al.* [11] proposed a method to generating four-part harmony automatically when a melody is given. Dannenberg *et al.* [12] focused on the music performance generated by cooperating with computers. They aim to enable musicians to incorporate computers into their performances by creating computer performers that play music along with music.

Our system is also regarded as an automatic accompaniment system for a vocal part generated by a user. Several such accompaniment systems have been proposed. Originally, Dannenberg [13] proposed an algorithm for finding the best match between an input music score and a solo performance with allowing some mistakes. Using that algorithm, he designed a system that produced an accompaniment in real time. Cont [14] also proposed an automatic accompaniment system. He used hidden hybrid Markov/semi-Markov models to estimate the current score position and dynamics of a tempo from input audio and a symbolic music score.

2.2 Arrangement Systems for Existing Songs

Many systems that can arrange existing songs have been proposed [1, 2, 15]. Yasuraoka *et al.* [2] proposed a music manipulation system that can change the timbre and phrase of a certain musical instrument in polyphonic music when the corresponding musical score is given. They define a musical-tone model to extract the target musical instrument. In this system, a user is asked to show his or her desirable arrangement in advance in the form of a music score. Tsuzuki *et al.* [16] proposed a system that extracts multiple vocal parts from existing songs and uses them to make a mashup. That system extracts multiple vocal parts of different singers singing the same song. The extracted voices are then overlapped on one shared accompaniment.

Some systems that assist the manipulation of an original instrumental part in real time have also been proposed. Yoshii *et al.* [1] proposed an arrangement system for a drum part. It enables a user to control the volume and

timbre of a drum part and to edit drum patterns in real time by using a drum sound recognition method and beat-tracking. Yamamoto *et al.* [15] propose a user interface for synthesizing a singing voice for a live improvisational performance. This interface enables real-time singing synthesis when lyrics information is given in advance.

2.3 Signal Processing Methods

We review relevant studies on vocal-and-accompaniment source separation (VASS), musical note estimation methods from F0 trajectories, and pitch modification in this section.

2.3.1 Source Separation

Methods for vocal-and-accompaniment source separation (VASS) have intensively been studied [4, 17, 18]. Rafii *et al.* [17] proposed a repeated pattern extraction technique (REPET). They focus on repetitive structures in music to extract a vocal part since accompaniments are repeated while a vocal part is not repeated. Huang *et al.* [18] realized singing voice separation by using robust principle component analysis (RPCA). They model an accompaniment part with repetitive structures as a low-rank component and model a vocal part as a sparse component. Ikemiya *et al.* [4] combined RPCA and pitch estimation of singing voices to realize VASS. They focus on the mutual dependency of singing voice separation and F0 estimation, and they improve singing voice separation in a unified framework.

2.3.2 Musical Note Estimation from F0 Trajectories

There have been several attempts to estimate musical notes when an F0 trajectory is given. In note estimation, discrete values with the interval of semitones are estimated from a continuous F0 trajectory. In *Songle* [6] these discrete values are estimated by using a majority-vote method for every time unit (*e.g.*, half beat). Laaksonen *et al.* [19] proposed a method of creating a symbolic melody when a chord progression and audio data are given. Ryynänen *et al.* [20] considered some states within one musical note (*e.g.*, vibrato and overshooting) and represented the inter-state transitions by using a hidden Markov model (HMM). Nishikimi *et al.* [5] represented the generative process of an F0 trajectory from underlying musical notes by using a Bayesian HMM.

2.3.3 Pitch Modification Algorithms

Pitch modification is a major requirement in music arrangement and has been examined in many studies [21]. In the time domain, pitch modification that preserves the original duration is achieved by time stretching and compression/expansion of the waveform. Many algorithms for time stretching have been proposed, and the simplest is that of Roucos *et al.* [22]. In their method, the waveform is divided into overlapping segments of a fixed length, and then they are shifted and added to realize time stretching. Hamon *et al.* [23] expanded this algorithm to make use of pitch information. In their method, the length of divided segments is determined according to the original pitch. Some of these segments are repeated or discarded to realize time stretching or time compression.

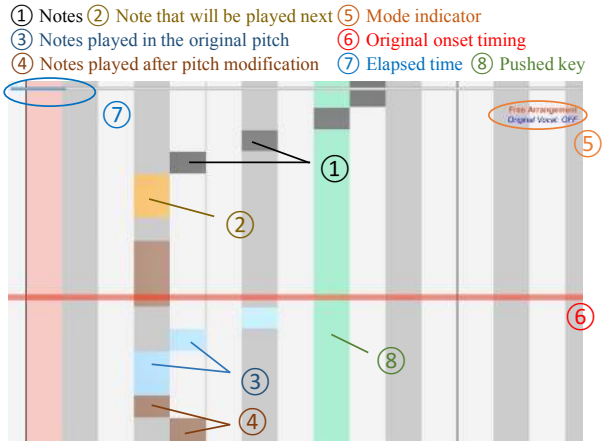

Figure 3. A screenshot of the proposed system. Each note comes down from above and is colored corresponding to its status. The current playing mode is indicated in the upper right corner. The pushed key is highlighted green.

3. USER INTERFACE

This section gives detailed descriptions of the proposed system and its user interface (Fig. 3). Section 3.1 explains the user experience which this system provides. Section 3.2 describes the user interface in detail.

3.1 Overview of the Proposed System

We propose a vocal arrangement system that enables users to manipulate a vocal part as if they were playing an instrument by using a MIDI keyboard. We focus on the following functions:

- Replacing the melody of the original vocal part with a melody as the user likes
- Adding a harmonic part
- Adding a troll part (*i.e.* an auxiliary part that follows a main part)

On the occasion of these arrangements, users change the features of the vocal part (*i.e.* the pitches, onset times, and durations) simultaneously. To treat these manipulations intuitively, we use a MIDI keyboard as the user interface. The target users are therefore the persons who can play the piano.

In this system, the vocal part is divided at the point where the pitch changes and is arranged. We henceforth refer to each divided segment as a *note*. When a key of a MIDI keyboard is pushed, the system first determines a note that is a target of his or her key operation. We henceforth refer to this note as an *target note*. Then the pitch and the duration of the target note are modified corresponding to user's key operation. We prepare three arrangement modes to determine the target note and to specify its pitch: *Free Arrangement mode*, *Time Fixed mode* and *Easy Arrangement mode* (Fig. 2).

Free Arrangement mode The note which has the earliest onset time in the notes that have not played by a user yet is estimated as the target note. The pitch of the target

note is shifted to the pitch of a key played by a user. While key operation is needed for every musical note, a user can change the onset time and the pitches no matter how far they are from the original ones.

Time Fixed mode The note being played in the original vocal part at the moment a user pushes the key is estimated as the target note. While the onset times and the durations are unchangeable, a user can change the pitch at the middle point of a note in this mode. The pitch of the target note is shifted to the pitch of a key played by a user as in the Free Arrangement mode. Moreover, a note is automatically assigned to music one by one even though a user keeps a key pushed for multiple notes. A user therefore does not have to push a key again if they wants to play multiple continuous notes in the same pitch.

Easy Arrangement mode The target note is determined in the same way as in the Time Fixed mode. In this mode, how the system modifies the pitch of the target note is specified by the number of semitones between a key pushed by a user and C4 (*e.g.*, if a user pushes E4, the pitch of the target note is increased by four semitones).

These three arrangement modes are prepared to cope with several different motivations for arrangement. When a user wants to manipulate all the vocal part as they likes, the Free Arrangement mode is suitable. On the other hand, when a user wants to manipulate a portion of the vocal part and leaves the rest as it is, the Time Fixed mode or the Easy Arrangement mode is suitable since with it the user does not have to make keyboard manipulations for notes that they wants to leave unchanged. When a user wants to add a harmonic part, the Easy Arrangement mode may be helpful because typical harmonic parts are played in a certain pitch relative to the original melody line (*e.g.*, third below) and the onset times and durations in harmonic parts are the same as ones in the original melody. Since a user can switch these modes at a keyboard during the performance, thereby more easily accessing the functions they need to accomplish the arrangement, the system can handle the different kinds of arrangement in a unified framework.

Furthermore, since some users want to eliminate the original vocal part while others just want to add a harmonic or troll part, leaving the original vocal part unchanged, the interface lets a user toggle the original singing voice on and off. This operation is also controllable during the performance. A user can also pause or resume the song by using a MIDI keyboard. Since a user needs to keep his or her eyes on the screen and cannot spare the attention needed to operate any equipments other than a MIDI keyboard, all the operations during the performance are controlled using only a MIDI keyboard.

3.2 Design of the Screen

The requirements for the screen presented to the user are as follows:

- The pitches and durations of musical notes in the original vocal part are shown in the way that they are intuitively and immediately evident to the user.

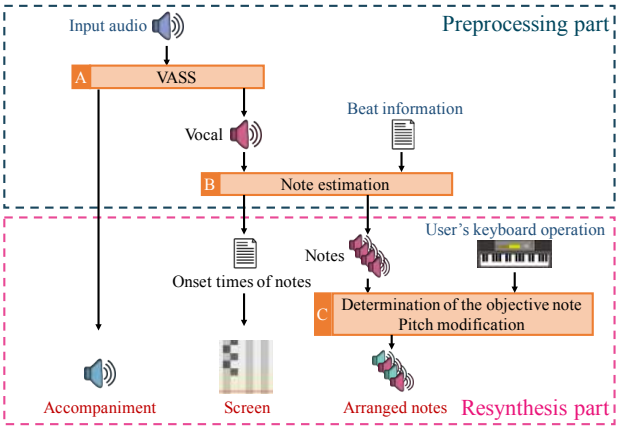

Figure 4. Overview of the proposed system: Inputs are indicated by blue letters and outputs are indicated by red letters.

- The response to a user's operation is displayed immediately.
- Users can predict the effect of their operation before performing the operation.

The first requirement is imposed because users sometimes refer to the original melody during their arrangement. The second requirement is imposed to make arrangement smooth and stress-free for users. The last requirement is imposed to let users get the modified performances they expect.

We designed the screen to meet these requirements (Fig.3). To meet the first requirement, we design the screen to have the horizontal direction indicating the pitches of the original vocal part. Moreover, we colored each line white or gray in order to represent white and black keys on a MIDI keyboard. This layout is like the one of a MIDI keyboard. Moreover, the line that indicates C4 (MIDI note number 60) is shown in red to make the octave of keys explicit. Musical notes in the original vocal part are visualized as rectangles coming down from above. This design is similar to the one of existing musical games, so users easily understand the presented screen.

To meet the second requirement, the color of the line corresponding to the key played by a user is changed to light green and the colors of notes that have already been played by user are changed into light blue or brown. By these changes in colors, the interface lets users understand the effect of their key operation and the pushed key without looking at their hands.

To meet the last requirement, the target note at the moment is shown in orange. This helps prevent unexpected manipulations.

Since the number of pitches that appear in the original vocal part varies according to the song, the number of lines shown in the screen changes according to the highest pitch and the lowest pitch in the song. The elapsed time is visualized at the top of the window.

4. SYSTEM IMPLEMENTATION

This section describes the signal processing techniques used in the system. As shown in Fig.4, the system is roughly divided into two parts: a *preprocessing part* and a

resynthesis part. In the preprocessing part, a singing voice is extracted from the song. This is enabled by vocal-and-accompaniment source separation (VASS). Then the timings that the pitch switches are estimated and the singing voice is divided at those timings. This is enabled by note estimation from an F0 trajectory. In the resynthesis part, a vocal part is manipulated and resynthesized according to the user's operations. The pitch of the note is modified to the one specified by the user, and then a modified singing voice is re-synthesized.

4.1 Vocal-and-Accompaniment Source Separation

We used the VASS method proposed by Ikemiya *et al.* [4] to extract a singing voice ("A" in Fig. 4) because it not only extracts a singing voice but also estimates an F0 trajectory that can be used to estimate musical notes. This method first represents the given music spectrogram as the sum of a low-rank matrix and a sparse matrix by applying RPCA. Since the singing voice is predominant in the sparse matrix, we extract the sparse matrix by applying a binary-mask and treat the extracted sparse matrix as a *rough* vocal spectrogram. Then the most likely F0 trajectory of a rough vocal spectrogram is estimated by subharmonic summation. Using this F0 trajectory, a harmonic mask that passes only the harmonic partials of estimated F0s is made. Finally it is integrated with the binary-mask mentioned above to extract a vocal spectrogram and an accompaniment spectrogram from the music spectrogram. The singing voice and the accompaniment are then synthesized using these spectrograms.

4.2 Note Estimation from an F0 Trajectory

We used an HMM-based method [5] for musical note estimation ("B" in Fig. 4). In this method, the process generating an F0 trajectory from underlying musical notes is modeled by using a Bayesian HMM. When beat information is given in advance, the pitches with the interval of semitones are estimated at a sixteenth-note level. Using this information, the notes of the vocal part are displayed to a user and the extracted singing voice is divided by note durations.

4.3 Pitch Modification

After the assignment of the note to the key, the pitch of the note is modified to the pitch of the key ("C" in Fig. 4). As noted in Section 2.3.3, several pitch modification algorithms have been proposed. We used PSOLA algorithm (Fig. 5) because it is simple enough to work in real time. Since it uses pitch information, the sound quality is also improved compared to the simplest SOLA algorithm. In this algorithm, first the waveform is divided by twice the length of its estimated wavelength, which is calculated from the estimated pitch. Then the divided segments are rearranged for time stretching, in which some segments are repeated or discarded if needed. Each segment is overlapped with the neighboring segments. After the rearrangement, overlapped segments are weighted by a window function and added. Finally, the rearranged segments are compressed or expanded to the original duration.

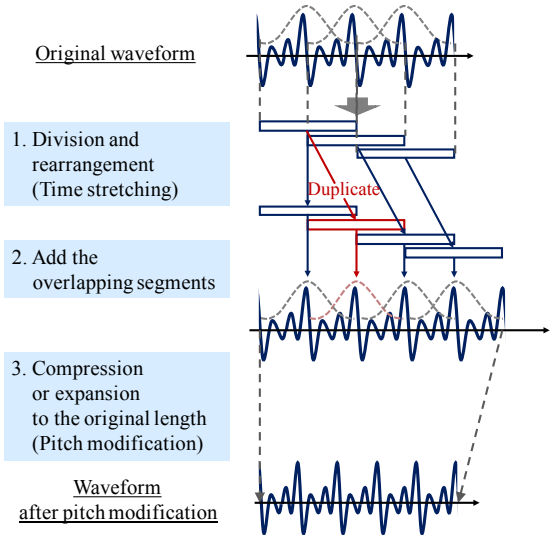

Figure 5. PSOLA algorithm: First the waveform is divided by twice the length of its wavelength and rearranged to realize time stretching, and then it is compressed or expanded to realize pitch modification.

5. EXPERIMENT

This section reports the subjective experiments conducted to evaluate the effectiveness of this system.

5.1 Experimental Conditions

We asked four subjects to use this system on RWC-MDB-P-2001 No.7 in the RWC music database (popular) [24]. In the experiment, we first showed the subjects the performance made by one of the authors as a tutorial. Then, we observed how each subject used the system and asked for his or her opinions in terms of the functionality of the system, the user interface, and the signal processing. We also asked the subject to evaluate the system in terms of the evaluation items shown below in a 5-point Likert scale (*i.e.*, strongly disagree, disagree, neutral, agree, strongly agree).

1. The Free Arrangement mode was easy to use.
2. The Time Fixed mode was easy to use.
3. The Easy Arrangement mode was easy to use.
4. The displayed information was understandable.
5. The lyrics were distinguishable after the pitch modification.
6. The pitches were modified as intended.
7. The individuality of the singer was preserved after the pitch modification.

In this experiment, we used as the notes shown on the screen the correct notes in the experiment. Finally, we used as the notes shown on the screen the notes estimated from an F0 trajectory to examine whether or not the precision of notes shown in the screen affects the ease of performance.

The subjects were four university students (three males and one female). All the subjects had played the piano for more than five years. Before the experiment, we asked the subject to listen to the original song.

	Strongly disagree	Disagree	Nautral	Agree	Strongly agree
Free Arrangement mode was easy to use	1	2	1		
Time Fixed mode was easy to use		1		3	
Easy Arrangement mode was easy to use	1			1	2
Displayed information was understandable			2	2	
Lyrics were understandable		1		2	1
Pitches were modified as intended			1	2	1
Individuality was preserved	2		1	1	

Figure 6. Evaluation of the proposed system. Each number represents the number of subjects who selected the corresponding item.

5.2 Experimental Results

In the experiment, the subjects tried multiple modes implemented on the system and the Time Fixed mode was mostly used. Most subjects used the Easy Arrangement mode to add a harmonic part to the original vocal part. In the Free Arrangement mode, sometimes it seemed hard for the subjects to comprehend the note they were playing. The result of the system evaluation is shown in Fig. 6. According to the result, the Time Fixed mode was popular among all the users. Moreover, the Easy Arrangement mode was mostly highly evaluated. On the other hand, the Free Arrangement mode was lowly evaluated due to its difficulties. As for the signal processing, the pitches were modified as the users intended, and the lyrics were recognizable after the pitch modification. On the other hand, this result indicates that the individuality of the singer was lost after the pitch modification. The remarkable opinions we obtained after the experiment are shown below.

The following opinions were obtained on the function:

- The system enabled me to get the sense of manipulating the vocal as if I were playing an instrument.
- It was fun to manipulate the vocal part as if the singer hit a wrong note.
- The assignment mode could be switched seamlessly.
- This system may be useful for practicing piano.
- *The easy mode* was useful for adding the expressions such as glissando or vibrato.

The obtained opinions show the effectiveness of the system as a kind of musical instruments for a singing voice.

The following opinions were obtained on the user interface:

- Practice is needed to arrange the original song because the manipulation speed required is too fast for arrangement at sight.
- The system should be unified into one device such as a tablet because it is strange that the screen and the keyboard are separated.

The difficulty on playing at sight is also seen in an ordinary piano performance.

The following opinions were obtained on the signal processing:

- Pitch modification sometimes significantly degrades the sound quality.
- The gap between the correct pitch and the estimated pitch makes me feel unpleasant when the estimated notes were shown on the screen.

Considering these opinions, we assume that the former problem occurs when the F0s of the singing voices contain some vibratos. PSOLA algorithm do not work well in such situations because it assumes that the F0s are invariant within one note. On the other hand, the latter problem occurs when extracted F0s contain some accompaniments and therefore some unnatural notes are generated.

5.3 Discussion

The opinions obtained encourage us to tackle following three problems: (1) making it easier to master the system, (2) improving sound quality after pitch modification, and (3) adding more functions. First, we plan to make the system easier to use by implementing it on a tablet that has a touchscreen as an input device to unify the device that displays information with the device that a user operates. By unifying the device, a user is free from paying attention to two different devices (*i.e.* the screen and the keyboard) at once and able to concentrate on their hand. Second, we try to use short-time Fourier transform and handling the signal in the time-frequency domain (*e.g.*, a phase vocoder). The sound quality after pitch modification is low due to the vibratos in singing voices. In the time-frequency domain, on the other hand, the envelope of the voice spectrum can be estimated and therefore it may contribute to preserving the individuality of the original singer. This improvement may make the arranged voice clear enough for a listener to recognize the original lyrics. Third, we are considering the feasibility of adding functions to change the volume and to assign two or more notes at the same time, which were desired in the subjective experiment.

6. CONCLUSION

This paper proposed a system for singing voice arrangement in real time. The subjective experimental results indicated that the proposed system provided users with the experience that they were arranging a vocal part as if playing an instrument. On the other hand, we found that the system is hard to use at first because the screen and the keyboard are separate devices. We therefore plan to implement this system in tablet that has a touchscreen as an input device. We also plan to improve the sound quality by using other signal processing methods such as a phase vocoder. Moreover, we plan to extend the function for arrangement of an accompaniment part and a drum part as well as a vocal part. This extension would provide a new user experience, the rearrangement of the component parts in songs. While this kind of arrangement is already partially realized by electronic organs, our goal is different from them because we aim to use the original parts contained in the original song.

Acknowledgments

This study was partially supported by JST CREST Project, JST ACCEL Project, JSPS KAKENHI 24220006, 26700020, 26280089, 16H01744. In this study, We used RWC music database (popular).

7. REFERENCES

[1] K. Yoshii *et al.*, "Drumix: An audio player with real-time drum-part rearrangement functions for active mu-

sic listening," *Trans. of IPSJ*, vol. 48, no. 3, pp. 1229–1239, 2007.

[2] N. Yasuraoka *et al.*, "Changing timbre and phrase in existing musical performances as you like: manipulations of single part using harmonic and inharmonic models," in *Proc. of the 17th ACM Int. Conf. on Multimedia*, 2009, pp. 203–212.

[3] H. Kenmochi *et al.*, "Vocaloid–commercial singing synthesizer based on sample concatenation," in *Proc. of the Interspeech Conf.*, 2007, pp. 4009–4010.

[4] Y. Ikemiya *et al.*, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *Proc. of ICASSP*, 2015, pp. 574–578.

[5] R. Nishikimi *et al.*, "Musical note estimation for F0 trajectories of singing voices based on a Bayesian semi-beat-synchronous HMM," in *Proc. of ISMIR*, 2016, pp. 461–467.

[6] M. Goto *et al.*, "Songle: A web service for active music listening improved by user contributions," in *Proc. of ISMIR*, 2011, pp. 311–316.

[7] M. Mauch *et al.*, "Song prompter: An accompaniment system based on the automatic alignment of lyrics and chords to audio," in *Late-breaking session at the 10th ISMIR*, 2010.

[8] T. M. Nakra *et al.*, "The ubs virtual maestro: an interactive conducting system," in *NIME*, 2009, pp. 250–255.

[9] I. Simon *et al.*, "Mysong: automatic accompaniment generation for vocal melodies," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2008, pp. 725–734.

[10] M. McVicar *et al.*, "Autorhythmguitar: Computer-aided composition for rhythm guitar in the tab space," in *Proc. of Joint ICMC and SMC Conf.*, 2014.

[11] L. Yi *et al.*, "Automatic generation of four-part harmony," in *Proc. of UAI Applications Workshop*, 2007.

[12] R. B. Dannenberg *et al.*, "Human-computer music performance: From synchronized accompaniment to musical partner," in *Proc. of SMC*, 2013.

[13] R. B. Dannenberg, "An on-line algorithm for real-time accompaniment," in *ICMC*, 1984, pp. 193–198.

[14] A. Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Trans. on PAMI*, vol. 32, no. 6, pp. 974–987, 2010.

[15] K. Yamamoto *et al.*, "Livo: Sing a song with a vowel keyboard," *JIP*, vol. 24, no. 3, pp. 460–468, 2016.

[16] K. Tsuzuki *et al.*, "Unisoner: An interactive interface for derivative chorus creation from various singing voices on the web," in *Proc. of ICMC*, 2014, pp. 790–797.

[17] Z. Rafii *et al.*, "Music/voice separation using the similarity matrix," in *Proc. of ISMIR*, 2012, pp. 583–588.

[18] P.-S. Huang *et al.*, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. of ICASSP*, 2012, pp. 57–60.

[19] A. Laaksonen, "Automatic melody transcription based on chord transcription," in *Proc. of ISMIR*, 2014, pp. 119–124.

[20] M. P. Ryynänen *et al.*, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[21] Z. Udo *et al.*, *DAFX - Digital Audio Effects*. Wiley, 2002.

[22] S. Roucos *et al.*, "High quality time-scale modification for speech," in *Proc. of ICASSP*, 1985, pp. 493–496.

[23] C. Hamon *et al.*, "A diphone synthesis system based on time-domain prosodic modifications of speech," in *Proc. of ICASSP*, 1989, pp. 238–241.

[24] M. Goto *et al.*, "RWC music database: popular, classic, and jazz music databases," in *Proc. of ISMIR*, 2002, pp. 287–288.