A MUSIC PERFORMANCE ASSISTANCE SYSTEM BASED ON VOCAL, HARMONIC, AND PERCUSSIVE SOURCE SEPARATION AND CONTENT VISUALIZATION FOR MUSIC AUDIO SIGNALS

Ayaka Dobashi **Yukara Ikemiya** **Katsutoshi Itoyama** **Kazuyoshi Yoshii**
Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University, Japan
{dobashi, ikemiya, itoyama, yoshii}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper presents a music performance assistance system that enables a user to sing, play a musical instrument producing harmonic sounds (*e.g.*, guitar), or play drums while playing back a karaoke or minus-one version of an existing music audio signal from which the sounds of the user part (singing voices, harmonic instrument sounds, or drum sounds) have been removed. The beat times, chords, and vocal F0 contour of the original music signal are visualized and are automatically scrolled from right to left in synchronization with the music play-back. To help a user practice singing effectively, the F0 contour of the user's singing voice is estimated and visualized in real time. The core functions of the proposed system are vocal, harmonic, and percussive source separation and content visualization for music audio signals. To provide the first function, vocal-and-accompaniment source separation based on RPCA and harmonic-and-percussive source separation based on median filtering are performed in a cascading manner. To provide the second function, content annotations (estimated automatically and partially corrected by users) are collected from a Web service called Songle. Subjective experimental results showed the effectiveness of the proposed system.

1. INTRODUCTION

In our daily lives, we often enjoy music in an active way, *e.g.*, sing a song or play a musical instrument. Although only a limited number of commercial music CDs include accompaniment (karaoke) tracks, karaoke companies provide those tracks for most major songs. To attain this, every time a new CD is released, a music expert is asked to manually transcribe the music (make a MIDI file). One of the main issues of this labor-intensive approach is that the sound quality of accompaniment tracks generated by MIDI synthesizers is often far below that of the original tracks generated by real musical instruments. In addition, the karaoke tracks that are originally available are usually completely instrumental and do not include chorus voices. The situation is much worse for people who want to sing minor songs or play musical instruments because they cannot use any karaoke or minus-one recordings (music recordings without particular instrument parts).

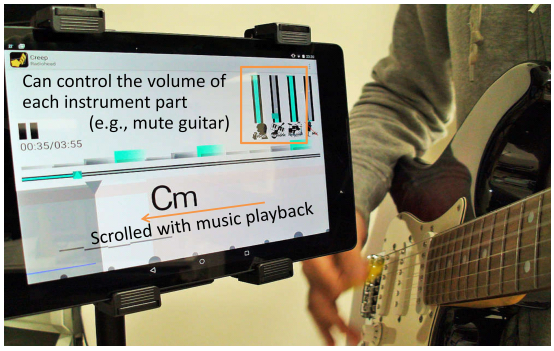

Figure 1. Example of how the proposed system is used: A user is playing a guitar with the playback of the other instrument parts (singing and drums) during seeing beat times and chord progressions displayed on a tablet.

In this paper we describe a music performance assistance system that enables a user to sing a song, play a harmonic musical instrument (*e.g.*, guitar), or play drums while playing back a minus-one version of an existing music recording (Fig. 1). When a user wants to sing a song, for example, the system plays back the accompaniment sounds by removing only predominant singing voices from the original recording (karaoke mode). An advantage of the proposed system is that accompaniment sounds it provides include chorus voices included in the original recording. Since the F0 of the user's singing voice is estimated and recorded in real time, the user can easily review his or her singing by comparing the F0 contour of the user's singing voice with that of the professional singing voice. When a user wants to play a harmonic instrument or drums, the system works similarly as it does in the karaoke mode. To further help a user, the beat times and chord progressions are displayed and automatically scrolled with the music playback. Since this system is implemented on a tablet computer, users can enjoy music in an active way anywhere.

To implement the system, we tackle two problems: vocal, harmonic, and percussive source separation (VHPSS) and content visualization for music audio signals. For the first, we propose a new method that combines vocal-and-accompaniment source separation (VASS) based on robust principal component analysis (RPCA) [1] and harmonic-and-percussive source separation (HPSS) based on median filtering [2] in a cascading manner. For the second, we col-

lect music-content annotations on music recordings from a Web service called Songle [3]. This service can automatically analyze four kinds of musical elements (beat times, chord progressions, vocal F0s, and musical structure) for arbitrary music audio signals available on the Web and visualize the analysis results on the user's browser. A key feature of Songle is that users can, as in Wikipedia, correct the analysis results if they find errors. Using this crowdsourcing Web service, the proposed system can keep the music content shown to users up-to-date.

2. RELATED WORK

This section reviews several studies related to our system in terms of three aspects: automatic accompaniment, active music listening, and sound source separation.

2.1 Automatic accompaniment

Automatic accompaniment systems using score information (MIDI data) of accompaniment parts have been developed for the two decades [4, 5]. Tekin *et al.* [6] and Pardo *et al.* [7], for example, proposed score following systems that can play back accompaniment sounds in synchronization with the performance of a user including tempo fluctuations and repeats of particular regions. Nakamura *et al.* [8] developed an improved system called Eurydice that can deal with repeats of arbitrary regions. Although some studies have tried to synchronize high-quality audio signals of accompaniment parts with user performances [9, 10], it is generally difficult to follow user performances played by polyphonic instruments (*e.g.*, piano). Mauch *et al.* [11] proposed a system called SongPrompter that can generate accompaniment sounds (drums and bass guitar) for any music audio signals without using the score information. To achieve this, the beat times and the F0s of bass lines are automatically estimated from music audio signals. The lyrics and chords given by a user are automatically synchronized with those signals and a display of the lyrics and chords is automatically scrolled in time to the music.

2.2 Active music listening

Active music listening [12] has recently been considered to be a promising research direction. "Active" means any active experience to enjoy listening to music (*e.g.*, touching-up music while playing it). Improved end-user's computing environments and music analysis techniques are making interaction with music more active. Goto *et al.* [3], for example, developed a Web service called Songle that helps a user better understand the content of a musical piece (repeated sections, beat times, chords, and vocal F0 contour) while listening to music by automatically estimating and visualizing the musical content. Yoshii *et al.* [13] proposed an audio player called Drumix that enables a user to intuitively customize drum parts included in the audio signal of a popular song without affecting the other sounds. Itoyama *et al.* [14] proposed a system that allows a user to control the volumes of individual instrument parts in real time by using a method of score-informed source separation. Yasuraoka *et al.* [15] proposed a method that enables a user to

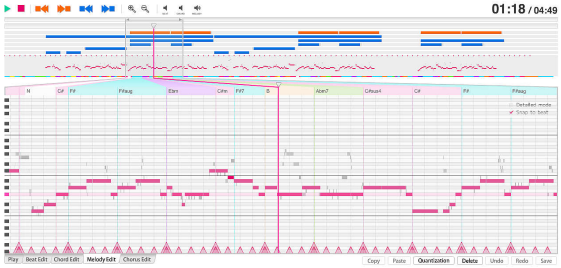

Figure 2. A screenshot of the web service called Songle: The repeated sections, beat times, chords, and vocal F0s of music audio signals are visualized on the browser.

freely edit a phrase of a particular instrument part in music audio signals while preserving the original timbre of the instrument. Fukayama and Goto [16] proposed a system that allows a user to mix the characteristics of chord progressions used in different music audio signals. Giraldo and Ramirez [17] proposed a system that changes the emotion of music in real time according to brain activity data detected by a brain-computer interface. Mancini *et al.* [18] proposed a system that, by analyzing user's motion, allows a user with mobile devices and environmental sensors to physically navigate in a physical or virtual orchestra space in real time. Chandra *et al.* [19] proposed a system that allows a group of participants with little or no musical training to play together in a band-like setting by sensing their motion with mobile devices. Tsuzaki *et al.* [20] proposed a system that assists a user to create derivative chorus music by mashing up multiple cover songs.

2.3 Source separation

A lot of effort has recently been devoted to vocal-and-accompaniment source separation (VASS) for music audio signals. Rafii and Pardo [21], for example, proposed a method called REPET that separates each short segment of a target music spectrogram into vocal components that significantly differ from those of the adjacent segments and accompaniment components that repeatedly appear in the adjacent segments. Liutkus *et al.* [22] generalized the concept of REPET in terms of kernel-based modeling by assuming that a source component at a time-frequency bin can be estimated by referring to other particular bins that are defined according to a source-specific proximity kernel. Huang *et al.* [23, 24] pioneered to use robust principal component analysis (RPCA) or deep neural networks for singing voice separation in an unsupervised or supervised manner. To improve the performance of VASS, Rafii *et al.* proposed a method that combines singing voice separation based on REPET with vocal F0 estimation. A similar method was proposed by Ikemiya *et al.* [1]. A key feature of this method is that only the singing voices corresponding to a predominant F0 contour are extracted and the other singing voices (*e.g.*, chorus voices) are separated as accompaniment sounds.

Several attempts have also been made to harmonic-and-percussive sound separation (HPSS). Yoshii *et al.* [13, 25] proposed a method that detects the onset times of drums by using a template adaptation-and-matching method and

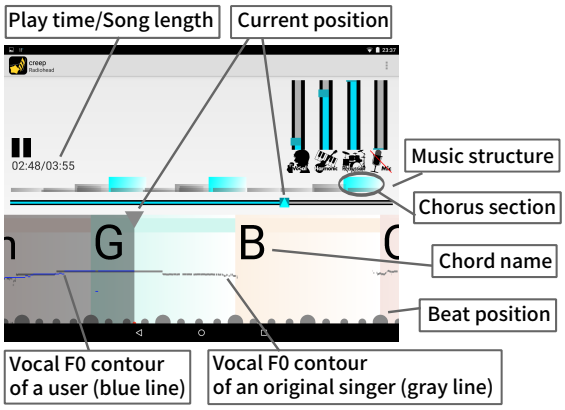

Figure 3. A screenshot of the user interface.

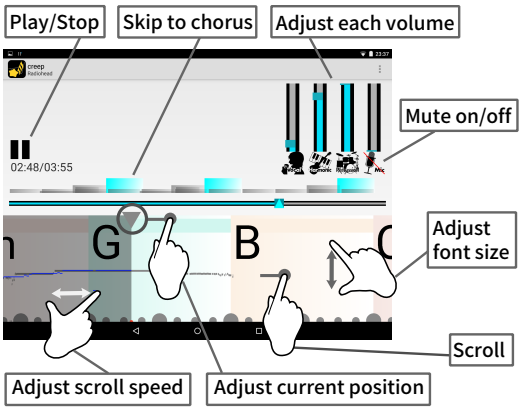

Figure 4. How to use the proposed system.

subtracts the drum sounds. Gillet and Richard [26] proposed a method that estimates a time-frequency subspace mask and then uses Wiener filtering. Rigaud *et al.* [27] proposed a method to extract drum sounds from polyphonic music by using a parametric model of the evolution of the short-time Fourier transform (STFT) magnitude. Miyamoto *et al.* [28] focused on the difference of isotropic characteristics between harmonic and percussive components and separated those components by minimizing a cost function. Fitzgerald *et al.* [2] also focused on the anisotropy, but used median filtering instead of a cost function.

3. USER INTERFACE

This section describes the GUI of the proposed system implemented on an Android tablet (HTC Nexus9). Figure 3 shows the components of the interface and Figure 4 shows how it is used. This interface provide two main functions: instrument-based volume control and music-content visualization. Although the proposed system is originally intended for music performance assistance, it is also useful for active music listening. Using the volume control function, users can listen to music while focusing on a particular instrument part. In addition, users can enjoy the music content visualized in real time as in the web service called Songle [3]. This helps a user better understand music and play a musical instrument in a musically meaningful and expressive manner.

3.1 Instrument-based volume control

The system allows a user to independently adjust the volumes of main vocals, harmonic instruments (including chorus vocals), and drums. In the upper right of the interface, three volume sliders corresponding to the different parts are provided. Another rightmost slider is used for controlling the volume of the microphone input in the karaoke mode (Figure 3). When the volume of a part the user sings or plays is turned down, the system plays back a karaoke or minus-one version of the original music audio signal. This function is useful for the practice purpose. When a vocalist of a typical rock band wants to practice a singing skill, for example, the volume of singing voices can be turned down. If the vocalist wants to sing a song and play a guitar simultaneously, the volumes of both singing voices and harmonic accompaniment sounds can be turned down. If all members of the rock band cannot meet together, it would be useful to play back only musical instrument parts corresponding to absent members. Because the system is implemented on a tablet computer and can be easily carried, it allows users to get sounds by the whole band whenever and wherever.

In the current implementation, vocal, harmonic, and percussive source separation (VHPSS) should be performed on a standard desktop computer in advance of using the volume control function. Since the computational power of recent embedded CPUs has rapidly grown, stand-alone VHPSS for any music audio signal stored on the tablet could be achieved in the near future.

3.2 Music content visualization

A display of chords and beat times is automatically scrolled in synchronization with the playback of the music. Since the chords, beat times, and vocal F0s are displayed at the bottom (Figure 3), a guitarist can play a guitar while watching the chords, and a vocalist can sing while checking his or her own F0s. The gray and blue lines on the chord pane are the vocal F0 contour of the original singing voices and that of the user's singing voices. The system allows a user to adjust the playback position by swiping horizontally on the chord pane (playback position can also be adjusted with the seek bar at the center). The system allows a user to adjust the display range of chord progressions by using a horizontal pinch in/out. A narrow display range allows a user to read each cord name clearly, while a broad one allows a user to read the following chords earlier. The playback speed is not changed by this operation. The system allows a user to adjust the font size of chord names by using a vertical pinch in/out. The triangle at the top of the chord pane indicates the current playback position. The system allows a user to adjust the location of the triangle by making a long press and horizontal swipe. Moving it to right allows a user to easily read the current chord and check the user's vocal F0s in real time, while moving it to the left allows a user to read the following chords easily.

A lot of overlapping rectangles over the central seek bar shows a hierarchical structure of a target music audio signal. The triangular mark on the seek bar shows where the current position on the structure is. The rectangles having

the same height indicate repeated sections. The light blue rectangles indicate chorus sections. When one of the blue rectangles is tapped on the screen, the playback position directly jumps to the start of the corresponding chorus section. This function helps a user practice playing the same section repeatedly.

4. SYSTEM IMPLEMENTATION

This section explains the technical implementation of the proposed system. The two main functions of the user interface described in Section 3 call for the development of vocal, harmonic, and percussive source separation (VHPSS) and automatic content analysis for music audio signals.

4.1 Source separation of music audio signals

We aim to separate music audio signals into singing voices, harmonic accompaniment sounds, and percussive sounds. To do this, the audio signals are first separated into singing voices and the other accompaniment sounds, which are further separated into harmonic sounds and percussive sounds. Figure 5 shows an overview of our approach.

4.1.1 Vocal and accompaniment source separation

We use the state-of-the-art method of singing voice separation [1] because it achieved the best performance in the singing voice separation track of MIREX 2014. As shown in Figure 6, robust principal component analysis (RPCA) is used for separating the amplitude spectrogram of a target music audio signal into a sparse matrix corresponding to singing voices and a low-rank matrix corresponsing to accompanying sounds. After a binary mask is made by comparing the two matrices in an element-wise manner, the vocal spectrogram is roughly obtained by applying the mask to the input mixture spectrogram.

The vocal F0 contour is then estimated by an extension of subharmonic summation (SHS) [1]. This method yields more accurate and smooth F0 contours than Songle. The following salience function $H(t, s)$ on a logarithmic scale is used [29]:

$$H(t,s) = \sum_{n=1}^{N} h_n P(t, s + 1200 \log_2 n), \tag{1}$$

where t is a frame index, s is a log-frequency [cents], $P(t, s)$ is the amplitude at time t and frequency s, N is the number of harmonic partials considered, and h_n is a partial weight. The A-weighting function considering the nonlinearity of the human auditory system is applied to the vocal spectrogram before computing $H(t, s)$, and the vocal F0 contour $\hat{S}$ is estimated by using the Viterbi algorithm as follows:

$$\hat{S} = \underset{S_1,\ldots,S_T}{\arg\max} \sum_{t=1}^{T-1} \{\log a_t H(t, s_t) + \log T(s_t, s_{t+1})\}, \tag{2}$$

where $T(s_t, s_{t+1})$ is a transition probability from the current F0 s_t to the next F0 s_{t+1} and a_t is a normalization factor. The basic SHS method without temporal continuity is also used for estimating the F0 contour of the user's singing voice in real time. A harmonic mask that passes only harmonic partials of given F0s is made on the assumption that the energy of vocal spectra is localized on harmonic partials. After the RPCA and harmonic masks are integrated, the vocal spectrogram is finally obtained by applying the integrated mask to the input spectrogram.

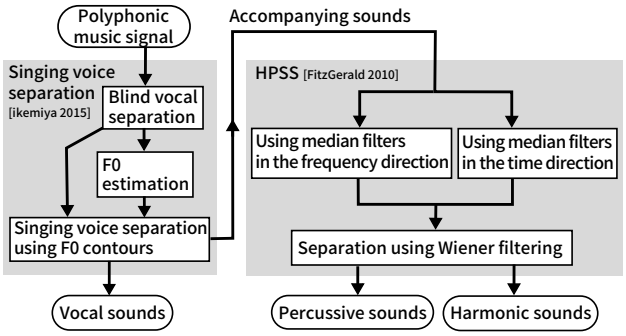

Figure 5. Source separation

4.1.2 Harmonic and percussive source separation

We use a method of harmonic and percussive source separation based on median filtering [2] because of its high performance, easy implementation and low computation cost.

The elements $P_{t,h}$, $H_{t,h}$ and $W_{t,h}$ of the percussive amplitude spectrogram $\mathbf{P}$, the harmonic spectrogram $\mathbf{H}$, and the given spectrogram $\mathbf{W}$ satisfy the following conditions:

1. $P_{t,h} \geq 0$;
2. $H_{t,h} \geq 0$;
3. $P_{t,h} + H_{t,h} = W_{t,h}$;

where t is a frame index and h is a frequency. As Figure 7 shows, this method focuses on the following observations:

1. Harmonic instrument sounds in a spectrogram are stable in the time-axis direction;
2. Percussive sounds in a spectrogram are stable in the frequency-axis direction;

Therefore it is possible to obtain $H_{t,h}$ and $P_{t,h}$ by removing the steep parts with median filters. Soft masks based on Wiener filtering are obtained by

$$M_{H_{t,h}} = \frac{H_{t,h}^p}{(H_{t,h}^p + P_{t,h}^p)}, \tag{3}$$

$$M_{P_{t,h}} = \frac{P_{t,h}^p}{(H_{t,h}^p + P_{t,h}^p)}, \tag{4}$$

where p is the power to which each individual element of the spectrograms is raised. Output spectrograms $\hat{\mathbf{H}}$ and $\hat{\mathbf{P}}$ are defined as follows:

$$\hat{\mathbf{H}} = \hat{\mathbf{S}} \otimes \mathbf{M_H} \tag{5}$$

$$\hat{\mathbf{P}} = \hat{\mathbf{S}} \otimes \mathbf{M_P} \tag{6}$$

where $\otimes$ represents element-wise multiplication and $\hat{\mathbf{S}}$ is the input mixture spectrogram.

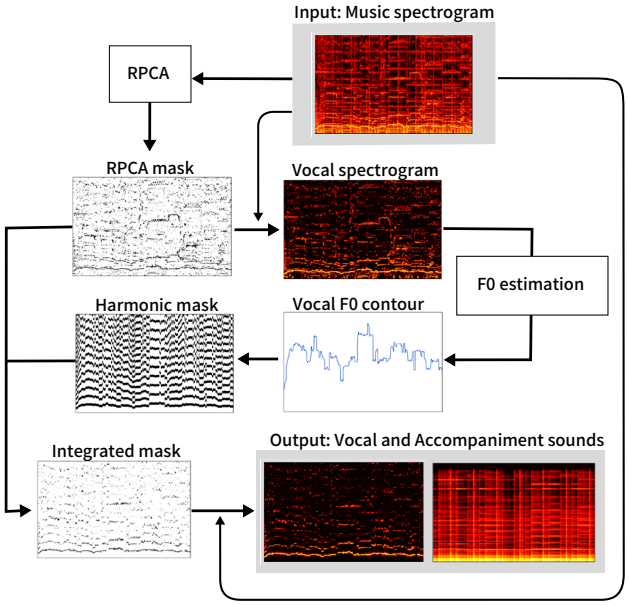

Figure 6. Vocal-and-accompaniment source separation

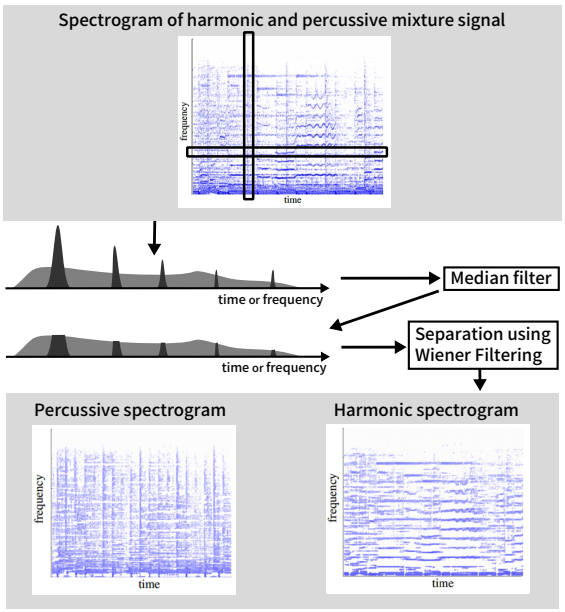

Figure 7. Harmonic-and-percussive source separation

4.2 Content analysis of music audio signals

The information for playing is obtained from Songle, which is a web service for active music listening [3]. It displays, for any music on the Web, automatic analysis results for various aspects, such as repeating structure, beat time, chords and vocal F0s. The user can correct errors by making annotations, so its accuracy gradually increases. By using this annotation mechanism, it is possible to sequentially update the data on a tablet.

5. EXPERIMENT

This section reports a subjective experiment conducted for evaluating the effectiveness of the proposed system.

5.1 Experimental conditions

A subject was asked to play a guitar according to the play-back of a Japanese popular song while using the proposed system. The subject was a 23-year-old university student who had played guitar for eight years. The effects on playing were examined with regard to three differences. More specifically, whether the system displayed information of music content or did not, whether the music-content information was correct or not, and whether or not the subject was allowed to adjust the volume of the guitar were investigated by observation and in interviews after the experiment. The detailed instructions were as follows:

1. Listen to the song;
2. Play guitar without performance support;
3. Play guitar with performance support.

The subject did as instructed under each of the following three conditions:

A) Chord and beat information were displayed or not (artist: Spitz, song: Cherry),

B) Automatic analysis results were corrected or not (artist: Perfume, song: polyrhythm),

C) Accompaniment volume was adjusted or not (artist: Aiko, song: Atashi no mukou).

Note that since music-content information is downloaded from Songle, estimation errors of music content are often included in an actual use if no users have corrected them.

5.2 Experimental results

The proposed system worked well as we expected and the perceptual quality of accompaniment sounds generated by the instrument-based volume control function reached a practical level. The result of the condition A showed that the visualization of chord information facilitated the music performance of the user. The result of the condition B indicated that although the visualization of automatic chord recognition results has some support effects, recognition errors often make it difficult to play the guitar in a comfortable way. This indicates the effectiveness of using Songle for keeping the music content shown to the user up-to-date.

Several kinds of improvements were suggested in terms of system usability. First, it would be better to show chord diagrams because unfamiliar chords often appear. Second, showing the highlights of a song would be helpful for planning a performance. A key transpose function would often be useful for making the performance easier [1] .

6. CONCLUSION

This paper presented a music performance assistance system based on vocal, harmonic, and percussive source separation of music audio signals. The beat times, chords, and vocal F0 contour are collected from Songle and are automatically scrolled from right to left in synchronization with the music play-back. To help a user practice singing effectively, the F0 contour of the user's singing voice is estimated and visualized in real time. The subjective experi-

[1] A demo video is available on http://winnie.kuis.kyoto-u.ac.jp/members/dobashi/smc2015/

mental results indicated that the system actually facilitates playing and increases a sense of play.

We plan to develop an intelligent function that follows the performance of a user including tempo fluctuations. In addition, we will tackle the implementation of all separation and analysis algorithms on a tablet computer.

Acknowledgments

This paper was partially supported by JST OngaCREST Project and KAKENHI No. 24220006, No. 26700020, and No. 26280089.

7. REFERENCES

[1] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *ICASSP*, 2015.

[2] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *DAFX*, 2010.

[3] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and N. Tomoyasu, "Songle: An active music listening service enabling users to contribute by correcting errors," *Interaction*, pp. 1363–1372, 2012.

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

[5] B. Vercoe, "The synthetic performer in the context of live performance," in *ICMC*, 1984, pp. 199–200.

[6] M. E. Tekin, C. Anagnostopoulou, and Y. Tomita, "Towards an intelligent score following system: Handling of mistakes and jumps encountered during piano practicing," in *CMMR*, 2005, pp. 211–219.

[7] B. Pardo and W. Birmingham, "Modeling form for on-line following of musical performances," in *AAAI*, 2005, pp. 1018–1023.

[8] E. Nakamura, H. Takeda, R. Yamamoto, Y. Saito, S. Sako, and S. Sagayama, "Score following handling performances with arbitrary repeats and skips and automatic accompaniment," *IPSJ Journal*, pp. 1338–1349, 2013.

[9] C. Raphael, "Music plus one: A system for flexible and expressive musical accompaniment," in *ICMC*, 2001, pp. 159–162.

[10] A. Cont, "ANTESCOFO: Anticipatory synchronization and control of interactive parameters in computer music," in *ICMC*, 2008, pp. 33–40.

[11] M. Mauch, H. Fujihara, and M. Goto, "SongPrompter: An accompaniment system based on the automatic alignment of lyrics and chords to audio," in *ISMIR*, 2010.

[12] M. Goto, "Active music listening interfaces based on signal processing," in *ICASSP*, 2007, pp. 1441–1444.

[13] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Drumix: An audio player with real-time drum-part rearrangement functions for active music listening," *Information and Media Technologies*, pp. 601–611, 2007.

[14] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models." in *ISMIR*, 2008, pp. 133–138.

[15] N. Yasuraoka, T. Abe, K. Itoyama, T. Takahashi, T. Ogata, and H. G. Okuno, "Changing timbre and phrase in existing musical performances as you like: manipulations of single part using harmonic and inharmonic models," in *ACM Multimedia*, 2009, pp. 203–212.

[16] S. Fukayama and M. Goto, "Harmonymixer: Mixing the character of chords among polyphonic audio," *ICMC-SMC*, pp. 1503–1510, 2014.

[17] S. Giraldo and R. Ramirez, "Brain-activity-driven real-time music emotive control," in *ICME*, 2013.

[18] M. Mancini, A. Camurri, and G. Volpe, "A system for mobile music authoring and active listening," *Entertainment Computing*, pp. 205–212, 2013.

[19] A. Chandra, K. Nymoen, A. Voldsund, A. R. Jensenius, K. H. Glette, and J. Tørresen, "Enabling participants to play rhythmic solos within a group via auctions," in *CMMR*, 2012, pp. 674–689.

[20] K. Tsuzuki, T. Nakano, M. Goto, T. Yamada, and S. Makino, "Unisoner: An interactive interface for derivative chorus creation from various singing voices on the web," *SMC*, 2014.

[21] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *ISMIR*, 2012, pp. 583–588.

[22] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," in *IEEE Trans. on Signal Processing*, 2014.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," *ISMIR*, 2014.

[24] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *ICASSP*, 2012, pp. 57–60.

[25] K. Yoshii, M. Goto, and H. G. Okuno, "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 333–345, 2007.

[26] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 529–540, 2008.

[27] F. Rigaud, M. Lagrange, A. Robel, and G. Peeters, "Drum extraction from polyphonic music based on a spectro-temporal model of percussive sounds," in *ICASSP*, 2011, pp. 381–384.

[28] K. Miyamoto, H. Kameoka, N. Ono, and S. Sagayama, "Separation of harmonic and non-harmonic sounds based on anisotropy in spectrogram," in *Acoustical Society of Japan Autumn conference*, 2008, pp. 903–904.

[29] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the acoustical society of America*, pp. 257–264, 1988.