ICASSP 2007 (The 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing) Proceedings, pp.IV-1441-1444, April 2007. (Invited Paper)

ACTIVE MUSIC LISTENING INTERFACES BASED ON SIGNAL PROCESSING

Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, JAPAN ＜m.goto [at] aist.go.jp＞

ABSTRACT

This paper introduces our research aimed at building *"active music listening interfaces"*. This research approach is intended to enrich end-users' music listening experiences by applying music-understanding technologies based on signal processing. Active music listening is a way of listening to music through active interactions. We have developed seven interfaces for active music listening, such as interfaces for skipping sections of no interest within a musical piece while viewing a graphical overview of the entire song structure, for displaying virtual dancers or song lyrics synchronized with the music, for changing the timbre of instrument sounds in compact-disc recordings, and for browsing a large music collection to encounter interesting musical pieces or artists. These interfaces demonstrate the importance of music-understanding technologies and the benefit they offer to end users. Our hope is that this work will help change music listening into a more active, immersive experience.

Index Terms— Music, user interfaces, interactive systems, information retrieval, music understanding.

1. INTRODUCTION

People who can *actively* interact with music have traditionally been considered musicians, whether composers, singers, or instrument performers, in the sense that "actively" implies the creation of music. On the other hand, ordinary people have been just listeners who could interact with music only *passively* by, for example, appreciating music, hearing background music, or listening to a music program on radio or TV. In the past, before it became possible to record the audio signals of music, non-musicians could only listen to live performances. Then, when the recording of music to audio storage media became a reality, some people started interacting with music in more active ways. For example, they could specify the playback order of musical pieces on an audio cassette tape or recordable compact disc, pursue high audio-playback quality (high fidelity), or adjust frequency characteristics by using graphic equalizers or simple tone controls (boost or cut) for bass and treble.

Recent advances in computer hardware, the spread of the Internet, and the availability of inexpensive audio input/output devices have now made it easy to rip music from CDs or download music from online music stores, and to handle collections of many musical pieces on a personal computer. It has also become possible to load a large number of pieces onto a portable music player enabling anyone to carry their personal collection of music anywhere and to listen to it at anytime. In addition to these changes in end-user environments, music-understanding technologies based on signal processing [1, 2, 3] have progressed to the level of dealing with CD recordings, which contain a mixture of polyphonic sounds from multiple instruments. These technological advances have greatly affected how ordinary people (non-musicians, listeners, or end users) interact with music. Even if such people cannot create music, current technologies enable them to interact with music in various active ways.

We have therefore pursued a research approach of building *"active music listening interfaces"* based on music-understanding technologies. In this approach, "active" does not mean the creation of new music, but any active experience that is part of enjoying music, such as finding new ways of music playback, touch-up (making small changes), retrieval, and browsing. Active music listening interfaces, for example, enable a user to skim rapidly through a musical piece by easily skipping sections of no interest, control the timbre of an instrument performance in CD recordings by replacing it with another instrument in real time, and browse a large music collection to find unfamiliar but interesting musical pieces.

The following sections show how music understanding and signal processing can contribute to end-user interfaces and change music listening into a more active experience.

2. ACTIVE MUSIC LISTENING

Active music listening can be discussed with regard to various aspects of music experience by end users. For example, we can think of the following scenarios and potential forms of active music listening.

2.1. Music playback

During music playback, the easiest *active* interaction by a user is to skip unappealing songs by pressing the next-track button on a music player. When music recording was not possible in the past, such skipping was impossible, but it is now easy with CD players or computer-based music players. Skipping sections of no interest within a song, however, is a bit more difficult with such players. If the music structure of a song could be automatically detected from CD recordings, it could be used to guide a user interacting with the song by freely skipping unappealing sections within the song.

As is apparent from music (promotion) videos such as those shown on Music Television (MTV), music playback accompanied by visual images enables end users to immerse themselves in the music or simply enjoy music more. In fact, recent computer-based music players often support a music visualization function that shows music-synchronized animation in the form of geometrical drawings moving synchronously with waveforms and frequency spectrums. A user can *actively* interact with the animation by pressing buttons to change drawing patterns when a visualization function supports such real-time interaction. If the visualization could be more closely related to the musical content, it would provide end users with unique experiences. For example, if musical beats could be automatically understood in CD recordings, the animation could be rigidly synchronized with the beats.

Lyrics often play an important role in songs and are sometimes appreciated by end users with special interests. A user sometimes refers to the printed or displayed lyrics during music playback to see words that the user cannot catch from the song or to truly understand messages the lyrics contain. In a more *active* way, a user may enjoy singing the lyrics along with music playback. These users, however, should themselves keep track of the current playback position in the lyrics. If the lyrics could be automatically synchronized with CD recordings, though, users could enjoy music playback while seeing the lyrics with the phrase being sung highlighted, much like on a *karaoke* machine. Even on a small screen like that of a cel-

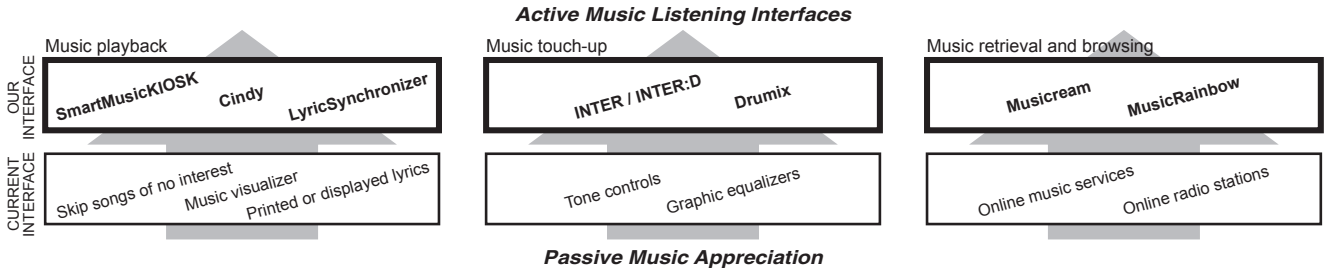


Fig. 1. Overview of our research approach with seven active music listening interfaces based on signal processing.

lular phone, the phrase from the lyrics currently being played can be automatically displayed and scrolled. Conversely, lyrics synchronization enables the user to click on a word in the lyrics shown on a screen to automatically change the current playback position to that word. Furthermore, if the lyrics could be automatically translated to another language in the future, a user could listen to a foreign song while seeing the synchronized lyrics in the user's native language through automatic translation from the original foreign lyrics.

2.2. Music touch-up

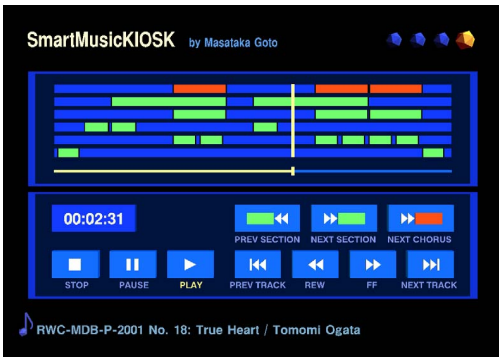
While listening to music, end users might wish that a song be sung by another singer, an instrument be played by another performer, or a particular part be performed in a different way. Realizing such changes is not possible with CD recordings, unless special individual tracks (separate recordings before mixing) corresponding to different instruments are available. If individual tracks like those of the vocal melody and drums could be automatically estimated and segregated from CD recordings, a user could *actively* change the volume or timbre of each instrument's track. In this paper, we name such individual changes (i.e., personalization or customization) of music by end users *"music touch-up"*. The touch-up supported by conventional music players is limited to the adjustment of frequency characteristics: only graphic equalizers or tone controls for bass and treble are available.

2.3. Music retrieval and browsing

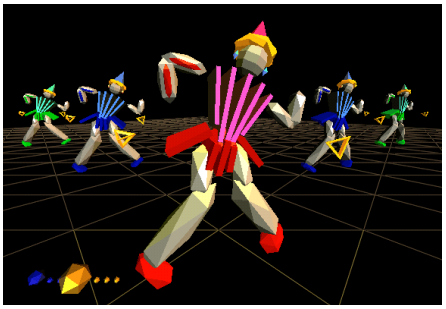
End users might know all the musical pieces in their personal music collections, but can never know all of the millions of musical pieces now available from online music stores or flat-rate, all-you-can-hear music subscription services. It is therefore necessary to help users retrieve music or browse a large music collection to *actively* find and listen to musical pieces of certain types. In fact, there are various commercial online services in which a user can retrieve musical pieces by specifying bibliographic metadata in text form (title, artist name, genre, etc.), browse lists of similar artists, or listen to personalized Internet radio stations. Most of these services are based on text metadata, manual annotation, or data of human (purchase) behavior, none of which rely on music understanding. If the contents of musical pieces could be automatically understood, the similarity between newly released or little known pieces, playlists, and artists could be measured. This content-based analysis is not biased by popularity and can be complementary to the approach taken by existing services to enable a user to actively interact with a music collection on the basis of content-based similarity.

3. ACTIVE MUSIC LISTENING INTERFACES BASED ON MUSIC-UNDERSTANDING TECHNOLOGIES

Music-understanding technologies based on signal processing are important to realize interfaces for active music listening as described in Section 2. State-of-the-art technologies make it possible to automatically estimate various *music scene descriptions* [1, 2, 3] in CD recordings, such as melody, bass, drums, musical beats, and chorus and phrase repetition. To prove that these technologies are indeed useful for end users, we have built seven active-music-listening interfaces, as shown in Figure 1. The following subsections briefly explain how our interfaces enable end users to enjoy music in more active ways.



(a) SmartMusicKIOSK: A user can actively listen to various parts of a song while moving back and forth as desired on the visualized song structure (the "music map" in the upper window).



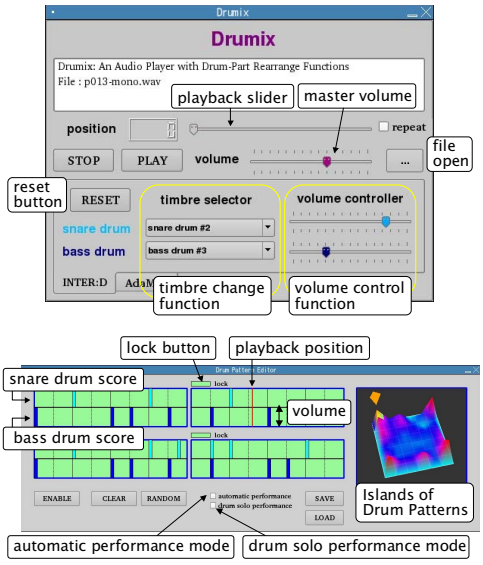
(b) Cindy: A user can actively select a dance sequence of music-synchronized virtual dancers during music playback.

Fig. 2. Active music listening interfaces for music playback.

3.1. SmartMusicKIOSK: Music listening station with a chorus-search function

SmartMusicKIOSK [4, 5] is a content-based playback-control interface for within-song browsing or trial listening for popular music. With this interface, a user can jump and listen to the chorus with just a push of the next-chorus button in the lower window of Figure 2(a). A user can also skip sections of no interest by interactively changing the playback position while viewing a "music map" as shown in the upper window of Figure 2(a). The music map is a visual representation of the entire song structure consisting of chorus sections (the top row) and repeated sections (the five lower rows). On each row, colored sections indicate similar (repeated) sections.

The chorus sections and various repeated sections are automatically estimated by our chorus-section detection method, RefraiD [5]. RefraiD tries to detect all repeated chorus sections appearing in a song with a focus on popular music.



(a) Drumix: A user can actively change drum sounds and drum patterns during music playback.

Fig. 3. Active music listening interface for music touch-up.

3.2. Cindy: Virtual dancers driven by beat tracking

Cindy [6] is a computer-graphics system that displays virtual dancers (Figure 2(b)) whose motions and positions change in time to musical beats in real time. This system has several dance sequences, each for a different dance motion mood. A user can select dance sequences one after another by pressing buttons during music playback. Since the timing of each motion and sequence change is automatically synchronized by beats, the user can readily enjoy the active interaction of switching sequences.

The musical beats and measures (bar lines) are automatically estimated through our beat-tracking method [6], which can deal with popular music with or without drum-sounds. Such real-time beat-driven dancer animation was first achieved in 1994 [7].

3.3. LyricSynchronizer: Automatic synchronization of lyrics with CD recordings

LyricSynchronizer is a system that displays scrolling lyrics with the phrase currently being sung highlighted during playback of a song. Because the lyrics are automatically synchronized with the song, a user can easily follow the current playback position even on a small screen. Moreover, a user can click on a word in the lyrics shown on a screen to jump to and listen from that word.

For this synchronization, we first segregate the vocal melody from polyphonic sound mixtures by using our predominant-F0 estimation method, PreFEst [8]. We then detect vocal sections and apply to those sections the Viterbi alignment (forced alignment) technique to locate each phoneme (vowel) [9].

3.4. INTER: Instrument equalizer for CD recordings

INTER [10] is a user interface for remixing/equalizing multiple audio tracks corresponding to different instruments in CD recordings by changing the volume of each track. Since it is very difficult to extract (i.e., demix) such tracks from CD recordings, we need further research such as [11] to fully realize INTER as proposed in [10]. As the first step, though, we developed a drum-sound equalizer called *INTER:D*. A user can actively change the volume or timbre of the sounds of bass and snare drums during music playback.

The onset times of those drums are automatically estimated by our drum-sound recognition method [12], which is based on the adaptation and matching of drum-sound templates.

3.5. Drumix: Audio player with a real-time drum-part editing function

Drumix [13] is a user interface for playing back a musical piece with drums as if another drummer were performing. With this interface, a user not only changes the volume or timbre of the sounds of bass and snare drums (in the upper window of Figure 3(a)), but also rearranges drum patterns of bass and snare drums (in the lower window). The user can casually switch drum sounds and drum patterns as the urge arises during music playback in real time.

This is an advanced version of INTER:D and uses the same drum-sound recognition method [12]. To deal with drum patterns in units of measure (bar), it also uses our beat-tracking method [6].

3.6. Musicream: Integrated music-listening environment for active, flexible, and unexpected encounters with musical pieces

Musicream [14] is a user interface for discovering and managing musical pieces. The idea behind Musicream is to see if we can break free from stereotyped thinking of how music playback interfaces must be based on lists of song titles and artist names. To satisfy a desire like "I want to hear something", it allows a user to unexpectedly come across various pieces similar to other pieces the user likes. As shown on the right side in Figure 4(a), disc icons representing pieces flow one after another from top to bottom, and a user can select a disc and listen to it. By dragging a disc in the flow, the user can easily pick out other similar pieces (attach similar discs). To satisfy a desire like "I want to hear something my way", Musicream gives a user greater freedom of editing playlists by generating a playlist of playlists. Since all operations are automatically recorded, the user can also visit and retrieve a past state as if using a time machine.

While any similarity measure between two musical pieces can be used for Musicream, the current implementation uses the content-based similarity between 30-dimensional feature vectors computed by Tzanetakis's MARSYAS [15]. Other interesting interfaces using content-based analysis have also been reported, such as "GenreSpace" and "GenreGram" [16], and "Islands of Music" [17].

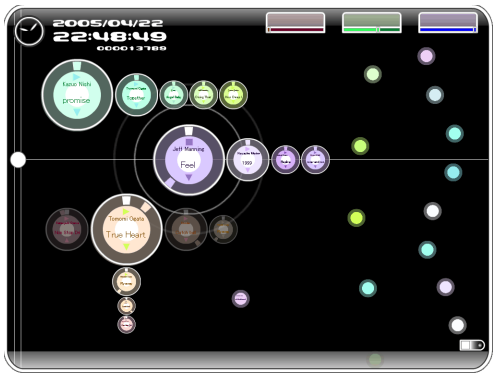
3.7. MusicRainbow: Artist discovery interface using audio-based similarity and web-based labeling

MusicRainbow [18] is a user interface for discovering unknown artists. As shown in Figure 4(b), all artists in a music collection are mapped on a circular rainbow where colors represent different styles of music. Similar artists are automatically mapped near each other and summarized with word labels at three different hierarchical levels. A user can rotate the rainbow by turning a knob and find an interesting artist by referring to the word labels. By pushing the knob, the user can select and listen to the artist highlighted at the midpoint on the right side.

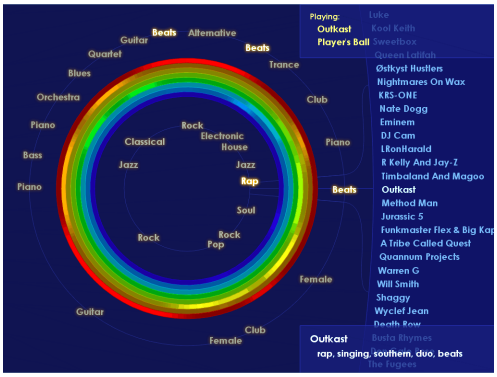
MusicRainbow is based on the content-based similarity between artists, which is computed from the similarity between musical pieces on the basis of low-level audio statistics similar to Musicream. The artists are then summarized with word labels extracted from web pages related to the artists.

4. CONCLUSION

We have described our research aimed at building active music listening interfaces that will enable non-musician users to enjoy music in more active ways than conventional passive music consumption. The interfaces we have built demonstrate the potential of music-understanding technologies based on signal processing. The current forms of automatic music understanding, however, are still limited compared to human music understanding. For example, the results of understanding always include errors, and we have to design interfaces that are more error resistant. At the same time, further research



(a) Musicream: A user can actively browse a music collection to discover musical pieces.



(b) MusicRainbow: A user can actively browse a music collection to discover artists.

Fig. 4. Active music listening interfaces for music retrieval and browsing.

on music understanding is needed to open new possibilities for various interfaces in terms of quality and quantity.

This paper has focused on the signal processing of musical audio signals, but the signal processing referred to in the paper's title can be any type of signal processing, such as the signal processing that lies behind human behavior or brain activity. If a future system could understand how good or comfortable a user feels during playback of a musical piece by monitoring and analyzing body or brain activities [19], for example, similar pieces could be recommended that would gratify latent user desires. Since people might worry about privacy issues in such monitoring, however, we have to be careful how it is applied. If it is used for purely private purposes, though, such a combination of body/brain signal processing with music signal processing will be promising for both user modeling and interface development.

We believe *active music listening* is beneficial not only because it allows a user to better enjoy music, but also because it promotes a better understanding and greater appreciation of music. The importance of *"activeness"* has already been recognized in different areas: *active listening* was proposed to enable better communication by paraphrasing the speaker's words, and *active reading* was proposed to enable better understanding of documents by underlining, highlighting, and commenting on the text. While an active reading interface with digital ink annotation [20] has been developed, its approach of supporting existing practices is different from our approach of creating novel listening experiences that cannot be achieved without new technologies. We believe active music listening helps users move from superficial consumption to a deep, full appreciation of music and enhances immersive music listening experiences.

Some of the scenarios envisioned in Section 2 have not yet become reality. We plan to work on the realization of these scenarios, such as a lyrics playback interface that displays music-synchronized native-language lyrics automatically translated from the original foreign-language lyrics and a music touch-up interface that replaces the original singing voice in CD recordings with the voice of another singer. We hope that various active music listening interfaces will be developed by many other researchers, and that such interfaces will become popular, thus changing everyday music listening into a more active experience.

ACKNOWLEDGMENTS: I thank (in alphabetical order by surname) Hiromasa Fujihara, Takayuki Goto, Hiroshi G. Okuno, Elias Pampalk, and Kazuyoshi Yoshii, who have been working with me to build the active music listening interfaces described in Section 3. This research was supported by CrestMuse, CREST, JST.

5. REFERENCES

[1] A. Klapuri and M. Davy, eds., *Signal Processing Methods for Music Transcription*. Springer, 2006.

[2] M. Goto, "Music scene description project: Toward audio-based real-time music understanding," in *Proc. of ISMIR 2003*, pp. 231–232, 2003.

[3] M. Goto and K. Hirata, "Invited review: Recent studies on music information processing," *Acoustical Science and Technology (edited by the Acoustical Society of Japan)*, vol. 25, no. 6, pp. 419–425, 2004.

[4] M. Goto, "SmartMusicKIOSK: Music listening station with chorus-search function," in *Proc. of UIST 2003*, pp. 31–40, 2003.

[5] M. Goto, "A chorus-section detection method for musical audio signals and its application to a music listening station," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.

[6] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[7] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," in *Proc. of ACM Multimedia 94*, pp. 365–372, 1994.

[8] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[9] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *Proc. of ISM 2006*, pp. 257–264, 2006.

[10] K. Yoshii, M. Goto, and H. G. Okuno, "INTER:D: A drum sound equalizer for controlling volume and timbre of drums," in *Proc. of EWIMT 2005*, pp. 205–212, 2005.

[11] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals," in *Proc. of ICASSP 2007*, 2007.

[12] 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*, vol. 15, no. 1, pp. 333–345, 2007.

[13] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Drumix: An audio player with functions of realtime drum-part rearrangement for active music listening," *Trans. of Information Processing Society of Japan*, vol. 48, no. 3, 2007.

[14] M. Goto and T. Goto, "Musicream: New music playback interface for streaming, sticking, sorting, and recalling musical pieces," in *Proc. of ISMIR 2005*, pp. 404–411, 2005.

[15] G. Tzanetakis and P. Cook, "MARSYAS: A framework for audio analysis," *Organized Sound*, vol. 4, no. 3, pp. 169–175, 2000.

[16] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," in *Proc. of ISMIR 2001*, pp. 205–210, 2001.

[17] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *Proc. of ISMIR 2003*, pp. 201–208, 2003.

[18] E. Pampalk and M. Goto, "MusicRainbow: A new user interface to discover artists using audio-based similarity and web-based labeling," in *Proc. of ISMIR 2006*, pp. 367–370, 2006.

[19] H. Katayose, N. Nagata, and K. Kazai, "Investigation of brain activation while listening to and playing music using fNIRS," in *Proc. of ICMPC 2006*, 2006.

[20] B. N. Schilit, G. Golovchinsky, and M. N. Price, "Beyond paper: supporting active reading with free form digital ink annotations," in *Proc. of CHI '98*, pp. 249–256, 1998.