Music Listening in the Future: Augmented Music-Understanding Interfaces and Crowd Music Listening

Masataka Goto[1]

[1]*National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, JAPAN*

Correspondence should be addressed to Masataka Goto (m.goto[at]aist.go.jp)

ABSTRACT

In the future, music listening can be more active, more immersive, richer, and deeper by using automatic music-understanding technologies (semantic audio analysis). In the first half of this invited talk, four *Augmented Music-Understanding Interfaces* that facilitate deeper understanding of music are introduced. In our interfaces, visualization of music content and *music touch-up* (customization) play important roles in augmenting people's understanding of music because understanding is deepened through seeing and editing. In the second half, a new style of music listening called *Crowd Music Listening* is discussed. By posting, sharing, and watching time-synchronous comments (semantic information), listeners can enjoy music together with the crowd. Such Internet-based music listening with shared semantic information also helps music understanding because understanding is deepened through communication. Two systems that deal with new trends in music listening — time-synchronous comments and mashup music videos — are finally introduced.

1. INTRODUCTION

One of our research goals is to enrich music listening experiences by deepening each person's understanding of music. Music listening experiences depend on music-understanding abilities. Although the music-composing/performing abilities of musicians are often discussed, the music-understanding abilities of casual listeners have not been well discussed or studied. It is difficult, for example, to define music-understanding abilities and express the results of understanding. In addition, it is difficult to know how others understand music. Music listening is usually an individual experience, and it is impossible to directly observe how others understand elements in music. Similarly, it is common to not notice what one does not understand when listening to music. Because of this, even if listeners want to better understand music or want to improve their ability to understand music, methods to realize those wishes have not yet been established and will have to be discovered.

We have therefore pursued a research approach of building *Augmented Music-Understanding Interfaces* [1] that facilitate deeper understanding of music by using automatic music-understanding technologies based on signal processing. First, visualization of music content plays an important role in augmenting (temporarily supporting) people's understanding of music because *understanding is deepened through seeing*. By recognizing and visualizing elements in music, we let end users (listeners) without expertise understand the existence of elements and the relationships between elements. Second, *music touch-up* (personalization or customization by making small changes to elements in existing music) [2] also helps music understanding because *understanding is deepened through editing*. It lets users naturally observe why music is composed and arranged in a certain way while casually enjoying the content modification (for example, changing the timbre and volume of instrument sounds in music).

Furthermore, Internet-based music listening with shared semantic information could also facilitate deeper understanding of music. It is a new style of music listening that can be called *Crowd Music Listening* (coined by me [3]) in which listeners are not alone anymore while listening to music and can enjoy music together with the crowd (anonymous listeners). By posting, sharing, and watching time-synchronous comments (semantic infor-

mation) about musical pieces or music video clips, it is possible to create a sense of shared listening/watching experience. The most popular web service for the Crowd Music Listening is *NICO NICO DOUGA*, which started in December 2006 and has more than 20 million users in Japan in 2011. Such Crowd Music Listening augments people's understanding of music because *understanding is deepened through communication*. This paper also introduces two systems that we developed by being inspired by the contents and comments on the NICO NICO DOUGA service.

2. AUGMENTED MUSIC-UNDERSTANDING INTERFACES: MUSIC INTERFACES FOR FACILITATING DEEPER UNDERSTANDING OF MUSIC

We have developed four examples of Augmented Music-Understanding Interfaces based on automatic music-understanding technologies, *SmartMusicKIOSK* [4, 5], *LyricSynchronizer* [6, 7, 8], *Drumix* [9], and *INTER* [10]. SmartMusicKIOSK and LyricSynchronizer are music listening interfaces that enable users to browse a musical piece with the help of the visualized music structure and lyrics, respectively. Drumix and INTER are music customization interfaces that enable users to make personal changes to drums and instruments, respectively. Given polyphonic sound mixtures taken from available music recordings, our music-understanding technologies can estimate the music structure (chorus and repeated sections), the lyrics alignment, drum sounds, and other instrument sounds for these interfaces.

2.1. SmartMusicKIOSK (Visualization of Music Structure)

SmartMusicKIOSK [4, 5] is a content-based playback-control interface for within-song browsing or trial listening for popular music. A user can skim rapidly through a musical piece by easily skipping to sections they are interested in while viewing a "music map" shown in the upper window of Figure 1. The music map is a graphical 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. This map helps a user decide where to jump. Clicking directly on a colored section plays that section. There are also buttons for jumping to the next chorus section, and the next or previous 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 survey of audio-based music structure analysis is described in [11].

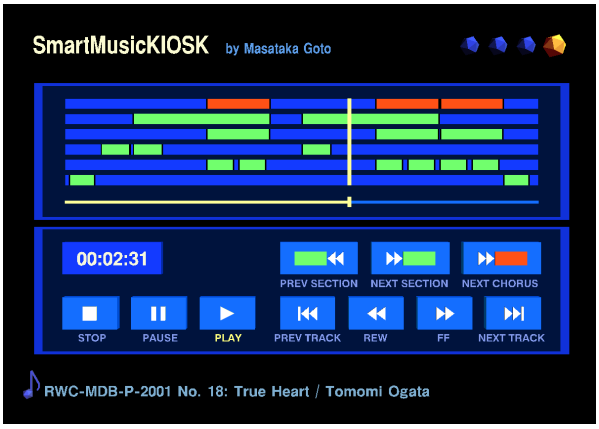

Fig. 1: SmartMusicKIOSK: Music listening station with a chorus-search function [4, 5].

SmartMusicKIOSK facilitates deeper understanding of music structure through music listening that focuses on it and better understanding of musical changes during repetition.

2.2. LyricSynchronizer (Visualization of Lyrics)

LyricSynchronizer [6, 7, 8] is a user interface that displays scrolling lyrics with the phrase currently being sung highlighted during playback of a song as shown in Figure 2. Because the lyrics are automatically synchronized with the song, a user can easily follow the current playback position. 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, LyricSynchronizer first segregates the vocal melody from polyphonic sound mixtures by using our predominant-F0 estimation method, PreFEst [12]. It then detects vocal sections and applies the Viterbi alignment (forced alignment) technique to those sections to locate each phoneme [6, 7, 8]. Because other previous methods [13] lacked the vocal segregation, their performances were limited. Even if the vocal was segregated, preliminary phoneme recognition was difficult [14].

LyricSynchronizer facilitates deeper understanding of lyrics through music listening that focuses on lyrics and better understanding of messages in lyrics.

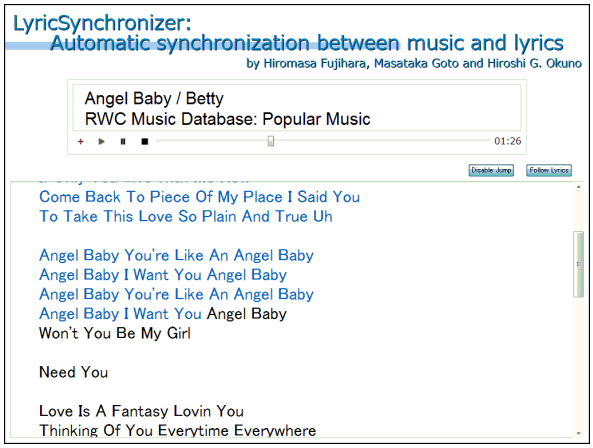

Fig. 2: LyricSynchronizer: Automatic synchronization of lyrics with polyphonic music recordings [6, 7, 8].

2.3. Drumix (Music Touch-up for Drums)

Drumix [9] is a user interface for playing back a musical piece with drums as if another drummer were performing. With a real-time drum-part editing function, a user can not only change the volume or timbre of the sounds of bass and snare drums, (in the upper window of Figure 3), but also rearrange 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.

The onset times of those drums are automatically estimated by our drum-sound recognition method [15], which is based on the adaptation and matching of drum-sound templates. To deal with drum patterns in units of bar (measure), it also uses our beat-tracking method [16]. Other drum-sound recognition methods [17] have also been proposed.

Drumix facilitates deeper understanding of drum sounds through music listening that distinguishes between them and better understanding of how drum sounds and drum patterns can change the feeling of music.

2.4. INTER (Music Touch-up for Instruments)

INTER (Instrument Equalizer) [10, 18] is a user interface for remixing/equalizing multiple audio tracks corresponding to different musical instruments in a polyphonic sound mixture. With a real-time instrument remixing function, a user can change the volume and pan control of each instrument track by using sliders. The

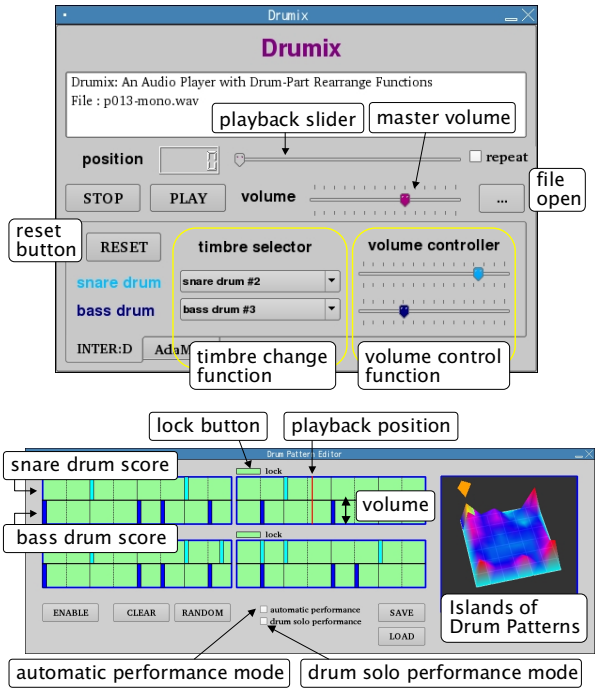

Fig. 3: Drumix: Music player with a real-time drum-part editing function [9].

user can casually change the balance of musical instruments during music playback.

Since it is difficult to extract such tracks from an existing musical piece with high accuracy, we have pursued a research approach called *score-informed sound source separation* [10, 18]. As a musical score, we assume that a standard MIDI file (SMF) synchronized with the input musical piece is available and can be used as prior information in decomposing sound sources. The volume and pan of separated sounds corresponding to each track of SMF can then be controlled. Other methods related to score-informed separation [19, 20] have also been proposed.

INTER facilitates deeper understanding of polyphonic sound mixtures through music listening that focuses on each musical instrument and better understanding of how different remixing balance can change the feeling of music.

2.5. Lessons Learned

Our experiences in designing the above interfaces revealed that the following functions are important:

- Visualization of the content of a musical piece,
- Synchronization of the visualization with music playback, and
- Provision of interactive interfaces for customizing and controlling music playback.

We also found that interfaces able to make users aware of what they tend to not notice are useful. User interaction enhances immersive music listening experiences and promotes a deep, full appreciation of music.

Originally, the above interfaces were developed as *Active Music Listening Interfaces* [2] that enable end users to enjoy music in active ways, but we later found that they provide good examples of Augmented Music-Understanding Interfaces. Active music listening is a way of listening to music through active interactions. Here, the word *active* is not meant to convey that the listeners create new music, but that they take control of their own listening experience.

Various Active Music Listening Interfaces we developed can be classified into three categories, music playback (SmartMusicKIOSK, Cindy, LyricSynchronizer, etc.), music touch-up (Drumix, INTER, etc.), and music discovery (Musicream, MusicRainbow, VocalFinder, etc.) [2]. Every Active Music Listening Interface could potentially be considered an Augmented Music-Understanding Interface, but the level of facilitating deeper understanding of music depends on interfaces. If an interface for music playback has a powerful function of visualizing music content that end users tend to have difficulty understanding, being aware of, or grasping the entire structure of, its interface can be a good candidate of Augmented Music-Understanding Interfaces. SmartMusicKIOSK for the music structure and LyricSynchronizer for the lyrics are thus picked up by this paper. All interfaces for music touch-up, such Drumix and INTER, can usually be Augmented Music-Understanding Interfaces because the casual content modification enables a user to naturally listen to music with a focus on customized elements and understand what happens if those elements are different from the original music. On the other hand, this paper does not pick up an interface for music discovery (retrieval and browsing) as an Augmented Music-Understanding Interface, though it might facilitate deeper understanding of an entire music collection (sets of music pieces). Future work will include development of Augmented Music-Understanding Interfaces for such music collections.

2.6. Future Direction

Since Augmented Music-Understanding Interfaces *temporarily* support people's understanding of music, the next step will be to develop interfaces that *permanently* improve one's ability to understand music. Although there are many ways to improve one's ability to understand a foreign language, such as through language schools and training materials, there are virtually no systematic means to improve music-understanding abilities. Most existing music schools and training materials including music-dictation training are intended for musicians and creators, not for casual listeners. I therefore propose a research approach *"Music-Understanding Ability Training Interfaces"* as an important future direction towards enabling a greater number of people to enjoy music in more depth from more diverse views.

3. CROWD MUSIC LISTENING: INTERNET-BASED MUSIC LISTENING WITH SHARED SEMANTIC INFORMATION

Crowd Music Listening is a shared music listening experience. When the recording of music was not possible in the past, people could only listen to live performances. It was a listening experience shared with other audiences in usual, which was sometimes accompanied by communication. Such communication gave a chance to people to know how others feel or understand in music. When the recording of music became possible, however, people started listening to music alone at anytime, anywhere by using personal music players. It was convenient, but it was not a shared experience anymore.

Crowd Music Listening can revive and reform a shared experience by enabling listeners to communicate shared semantic information or annotation over the Internet. A video communication web service *NICO NICO DOUGA* is the most popular web service for such Crowd Music Listening. It started in December 2006 and has more than 20 million users (including more than 1.1 million paid users) in Japan in February, 2011. It was not developed by ourselves, but developed by *Dwango Co., Ltd.* and managed by *Niwango Co., Ltd.* As with other video sharing services, users can upload, share, and view video clips on this service. In addition, users can post, share, and watch time-synchronous text comments (semantic information) about musical pieces or music video clips.

Since this service is unique and attractive, there are more than 6 million video clips with 31 hundred million user comments. Although this is not just for music, music is one of the major categories of video clips.

The NICO NICO DOUGA service supports novel networked text communication where recent comments by anonymous users are overlaid on the video screen in synchronization with the video playback. Each comment typed in at a specific playback time within a video clip flows from the right to the left over the video screen at that time. Comments can thus be related to various events in a video clip and shared with users who watch its clip currently or in the future. Since 2007 this service also supports live video streaming where user comments are posted and shared in real time: posted comments are immediately overlaid on the video screen of other users watching the same streaming. Although there were relevant web-based video annotation systems [21, 22] where comments can be associated with events in a video clip, those systems did not support the comment overlay.

Since the overlaid comments can create a sense of shared watching/listening experience called *Pseudo-Synchronized Communication* (coined by Satoshi Hamano), users can feel as if they enjoy together with the crowd (anonymous users). In fact, users usually post impressions, interpretations, and feelings about the video/music content. Moreover, we can observe various interesting ways of using comments. Barrage (*DANMAKU* in Japanese) of comments, for example, is a phenomenon where users posted too many comments that almost hide the original video clip at a certain time, which gives an impression that a lot of users are shouting there. ASCII art that is a graphical drawing consisting of text characters also sometimes appears. In music video clips, which are sometimes a sequence of still images with music, original or parody lyrics are often posted in synchronization with the singing voice.

When users watch music video clips, such Internet-based music listening with shared semantic information (comments) enables a greater number of people to enjoy music in more depth from more diverse views. This shared music listening facilitates deeper understanding of music because users can know how others understand music. By reading comments, users could notice what they do not understand in music. The NICO NICO DOUGA service can thus be considered an example of Augmented Music-Understanding Interfaces.

By being inspired by such comments and contents on the NICO NICO DOUGA service, we have developed two systems, *MusicCommentator* [23] and *DanceReProducer* [24]. MusicCommentator is a system that can learn the relationship between music (audio signals) of video clips and their text comments on NICO NICO DOUGA and then automatically generate time-synchronous comments on a new video clip. In other words, it can emulate the commenting behavior of users by using existing clips and users' comments given to those clips. DanceReProducer, on the other hand, is a system that can learn the relationship between music (audio signals) and image sequences (video frames) in dance video clips on NICO NICO DOUGA and then automatically generate a dance video clip appropriate to a given piece of music.

3.1. MusicCommentator: A System for Generating Comments Synchronized with Music Audio Signals

MusicCommentator [23] is a system that generates possible natural-language comments on appropriate temporal positions in a music audio clip (musical audio signals in a video clip). To achieve this, we proposed a joint probabilistic model of audio signals and text comments. As shown in Figure 4, the system consists of a *learning* phase and a *commenting* phase. In the learning phase, the model is trained by using musical audio signals of existing video clips and users' comments given to those clips on the NICO NICO DOUGA service. General language models (uni-, bi-, and tri-grams) are also learned from numerous comments of various clips. In the commenting phase, given a new clip and some of its comments, the joint probabilistic model is utilized to estimate what temporal positions could be commented on and what comments could be added to those positions. It then concatenates possible words by taking language constraints (the general language models) into account.

The joint probabilistic model for multi-modal features is developed by extending a standard hidden Markov model (HMM) as shown in Figure 5. Let $z_t^{(n)}$ be a state representation at frame[1] t in clip n. Acoustic features ($a_t^{(n)}$) and textual features ($w_t^{(n)}, d_t^{(n)}, l_t^{(n)}$) are associated with the same state, and we consider a joint output distribution that indicates how likely four kinds of features $a_t^{(n)}, w_t^{(n)}, d_t^{(n)}, l_t^{(n)}$ jointly occur from each state. As

[1] A frame here is a short duration (256 ms) to be analyzed in an audio clip.

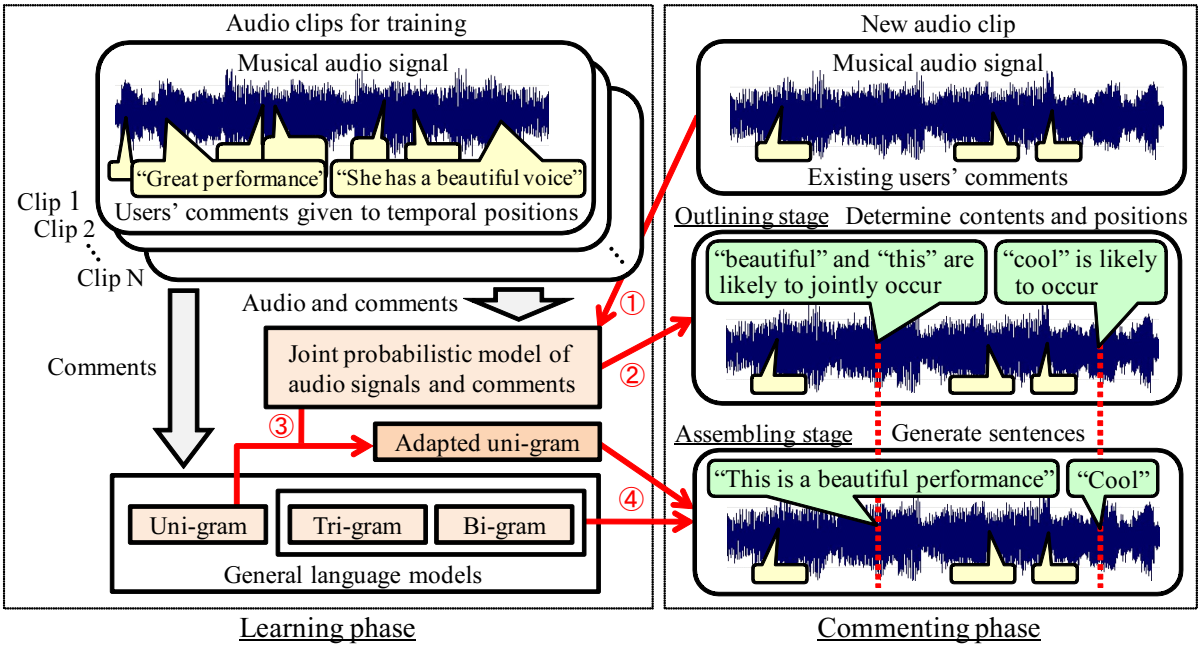

Fig. 4: MusicCommentator: Automatic generation of time-synchronous text comments [23].

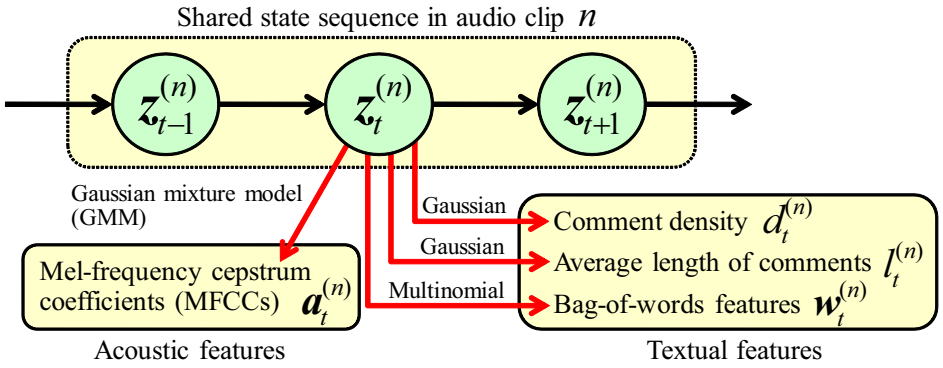

Fig. 5: MusicCommentator: Acoustic and textual features jointly occur from each state in our ergodic hidden Markov model [23].

acoustic feature vectors $a_t^{(n)}$, we use mel-frequency cepstrum coefficients (MFCCs) and their delta components. The content of comments is controlled by a bag-of-words vector $w_t^{(n)} = \{w_{t,1}^{(n)}, \cdots, w_{t,V}^{(n)}\}$, where V is the number of words in a vocabulary generated from comments and $w_{t,v}^{(n)}$ $(1 \leq v \leq V)$ represents the number of occurrences of word v per comment. The comment density $d_t^{(n)}$ indicates the number of comments in each frame and is utilized to learn what temporal positions should be annotated in a target clip. The average length of comments $l_t^{(n)}$ indicates the average number of words in a single comment and is utilized to learn how long comments could be generated.

Our model is an ergodic HMM, which allows any state transition at any time because we cannot assume state sequences in advance. In the learning phase, we

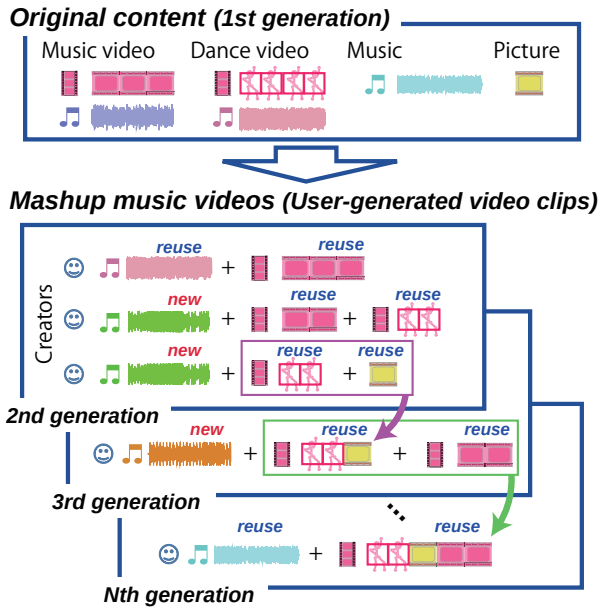

Fig. 6: Generation of user-generated mashup music video clips by reusing existing original content [24].

estimate model parameters by using the Expectation-Maximization (EM) algorithm. In the commenting phase, given a target clip (musical piece), we first determine the most likely sequence of latent states by using the Viterbi algorithm. We can then determine how many comments should be provided in each frame and the most likely bag-of-words features. Comments are finally generated by adapting the general language models into the most likely bag-of-words features and concatenating words based on the adapted models.

Experimental results reported in [23] were promising but revealed that we are still far from the ultimate goal of building a computer system that can describe impressions of music/video clips as natural language as humans do. Because commenting is one of the most sophisticated cognitive functions of humans, it might be too hard to emulate even if we will use more advanced machine learning techniques. We think, however, that this study is an important first challenge. To improve the system, we have to take into account higher-level semantic musical features related to melody, rhythm, and harmony. Other approaches for semantic audio annotation and retrieval [25] could also be integrated. In addition, visual features of music video clips should be dealt with.

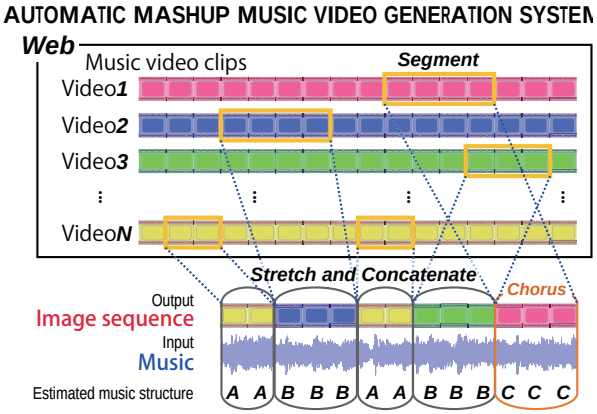

Fig. 7: DanceReProducer: Automatic music video generation system reusing existing video clips [24].

3.2. DanceReProducer: A Mashup Music Video Generation System by Reusing Dance Video Clips

DanceReProducer [24] is a dance video authoring system that can automatically generate a mashup/MAD music video clip[2] for a given piece of music. This system focuses on reusing ever-increasing user-generated video clips on the NICO NICO DOUGA service. A mashup music video clip usually consists of a musical piece (audio signals) and image sequences (video frames) taken from other original video clips. Original video clips are called *1st generation (original) content*, and mashup video clips generated by users can be considered *2nd generation (derivative) content* (Figure 6). Video clips of the 2nd, 3rd, and *N*th generation content on NICO NICO DOUGA sometimes form a massive dependency network [26].

Given a target musical piece, DanceReProducer segments, stretches, and concatenates image sequences of existing dance video clips (Figure 7). In a music video clip, image sequences are often synchronized with or related to music. Such relationships are diverse in different video clips, but were not dealt with by previous methods for automatic music video generation. Our system employs machine learning and beat tracking techniques to model these relationships — i.e., learn the bar-level relationships between music features (MFCCs, spectral

[2]User-generated video clips called *mashup videos* or *MAD movies*, each of which is a derivative (mixture or combination) of original video clips, are gaining popularity on the web and a lot of them have been uploaded and are available on video sharing web services.

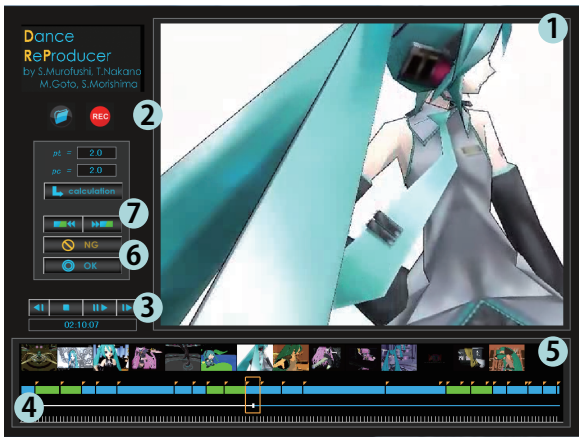

Fig. 8: DanceReProducer: User interface [24].

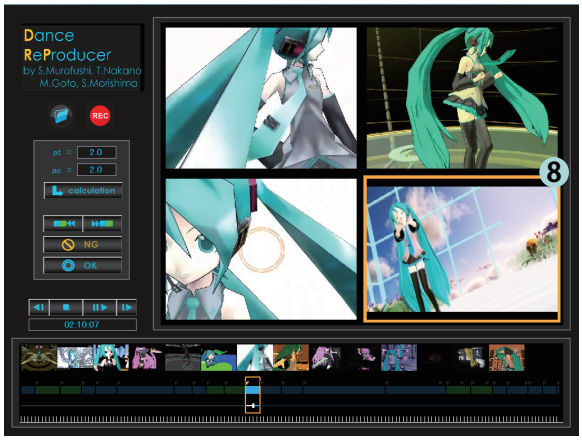

Fig. 9: DanceReProducer: Interactive sequence selection. Four different image sequence candidates are previewed and the lower-right candidate is chosen by a user [24].

flux, zero-crossing rate, etc.) and visual features (optical flow, hue, saturation, brightness, etc.) in user-generated dance video clips on the NICO NICO DOUGA service. Since the view count of a video clip on such a service tends to reflect the content quality, we let video clips having large view counts have larger influence on the machine learning results. To generate new music video clips, short image sequences that have been previously extracted from other music clips are stretched and concatenated so that the emerging image sequence matches the rhythmic structure of the target song.

Besides automatically generating dance video clips, DanceReProducer offers a user interface (Figures 8 and 9) in which a user can interactively change image sequences just by choosing different candidates. This way people with little knowledge or experience in mashup movie generation can interactively create personalized video clips. In more detail, the user can watch the generated image sequence (Figure 8, ①), load a target piece and save the generated video (②), and play back the generated video (③). At the bottom, the playback slider is shown with the music structure estimated by RefraiD [5] (④) and with thumbnail images (⑤). If the automatically generated video clip is satisfactory, the user can just watch it. But if the user does not like generated image sequences, the user can see and compare different candidates by clicking the NG button (⑥) for the current section that can be changed by jump buttons (⑦), and simply choose the preferred one (Figure 9, ⑧).

In our experiences, most video clips generated by DanceReProducer were synchronized regarding rhythm and impression between music and image sequences. But there are still remaining issues since the current implementation is just the first step and has limitations. For example, the use of higher-level semantic musical features will improve the performance. Because we do not use dance-related visual features at all, the current DanceReProducer can be applied to any video clips (even personal home videos [27]), but the use of visual features describing dance motions will also improve the performance.

4. CONCLUSION

We have described that music listening in the future will be enriched by automatic music-understanding technologies. First, Augmented Music-Understanding Interfaces facilitate deeper understanding of music than conventional passive music appreciation by the visualization and customization of music content. SmartMusicKIOSK, LyricSynchronizer, Drumix, and INTER, for example, will respectively facilitate deeper understanding of music structure, messages in lyrics, how drum sounds and patterns can change the feeling of music, and how different remixing balance can change the feeling of music. Second, shared semantic information (comments) by Crowd Music Listening also facilitate deeper understanding of music because people can know how others understand music. Understanding is thus *deepened through seeing, editing, and communication*. Furthermore, MusicCommentator and DanceReProducer are

also based on music-understanding technologies with the focus on machine learning and deal with new trends in music listening — time-synchronous comments and mashup music videos (*N*th generation content).

Considering environmental and energy problems on the earth, music is one of the best earth-friendly entertainment, and this research could contribute to energy-efficient music appreciation/production. To reduce material consumption, it would be better to promote online music without using physical media (tape, record, CD, DVD, etc.). As overwhelming convenience of CDs was the driving force behind the transition from records to CDs, driving forces for the online music are necessary. Music-understanding technologies could contribute to such a driving force by providing attractive convenience or experience such as that achieved on our Augmented Music-Understanding Interfaces and Active Music Listening Interfaces. Furthermore, to reduce energy consumption, the reuse (recycling) of music, such as *N*th generation content and mashup music videos, would be energy efficient and reuse technologies like DanceReProducer would become more important. From the beginning, music production requires less energy/material than other media content such as movies in general. In addition, repeated listening of a musical piece is natural and often increases its attraction. Our Augmented Music-Understanding Interfaces could contribute to enabling users to find attraction of musical pieces, thus resulting in giving those pieces longer life.

Although this paper focuses on music, the concept of Augmented Music-Understanding Interfaces and Crowd Music Listening can be extended to any media content. Future work will therefore include research and development of *Augmented Content-Understanding Interfaces* and *Crowd Content Appreciation*, which could be further extended to technologies for augmenting the understanding of any phenomena in the real world.

5. ACKNOWLEDGMENTS

I thank (in alphabetical order by surname) Hiromasa Fujihara, Katsutoshi Itoyama, Hiroshi G. Okuno, and Kazuyoshi Yoshii, who have worked with me to build the Augmented Music-Understanding Interfaces described in Section 2. I also thank Kazuyoshi Yoshii for MusicCommentator described in Section 3.1 and Tomoyasu Nakano, Sora Murofushi, and Shigeo Morishima for DanceReProducer described in Section 3.2. This work was supported in part by CrestMuse, CREST, JST.

6. REFERENCES

[1] M. Goto, "Augmented music-understanding interfaces," in *Proc. of the 6th Sound and Music Computing Conference (SMC 2009) (Inspirational Session)*, 2009.

[2] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, 2007.

[3] M. Goto, "Keynote talk 'Augmented music-understanding interfaces: Toward music listening in the future'." International Workshop on Advances in Music Information Research 2009 (AdMIRe 2009) of IEEE International Symposium on Multimedia 2009 (ISM 2009), December 2009.

[4] M. Goto, "SmartMusicKIOSK: Music listening station with chorus-search function," in *Proc. of the 16th Annual ACM Symposium on User Interface Software and Technology (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] 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 the IEEE International Symposium on Multimedia 2006 (ISM 2006)*, pp. 257–264, 2006.

[7] H. Fujihara and M. Goto, "Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection," in *Proc. of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, 2008.

[8] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "LyricSynchronizer: Automatic synchronization method between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, 2011. (accepted).

[9] 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," *Trans. of Information Processing Society of Japan*, vol. 48, no. 3, pp. 1229–1239, 2007.

[10] 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 the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pp. I–57–60, 2007.

[11] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pp. 625–636, 2010.

[12] 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.

[13] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 338–349, 2008.

[14] M. Gruhne, K. Schmidt, and C. Dittmar, "Phoneme recognition in popular music," in *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pp. 369–370, 2007.

[15] 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.

[16] 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.

[17] J. Paulus and A. Klapuri, "Drum sound detection in polyphonic music with hidden Markov models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, 2009.

[18] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Parameter estimation for harmonic and inharmonic models by using timbre feature distributions," *Trans. of Information Processing Society of Japan*, vol. 50, no. 7, pp. 1757–1767, 2009.

[19] J. Woodruff, B. Pardo, and R. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pp. 314–319, 2006.

[20] S. Ewert and M. Müller, "Estimating note intensities in music recordings," in *Proc. of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pp. 385–388, 2011.

[21] D. Yamamoto and K. Nagao, "iVAS: Web-based video annotation system and its applications," in *Proc. of the 3rd International Semantic Web Conference (ISWC 2004)*, 2004.

[22] K. Kaji and K. Nagao, "MiXA: A musical annotation system," in *Proc. of the 3rd International Semantic Web Conference (ISWC 2004)*, 2004.

[23] K. Yoshii and M. Goto, "MusicCommentator: Generating comments synchronized with musical audio signals by a joint probabilistic model of acoustic and textual features," in *Proc. of the 8th International Conference on Entertainment Computing (ICEC 2009) (Lecture Notes in Computer Science)*, pp. 85–97, 2009.

[24] T. Nakano, S. Murofushi, M. Goto, and S. Morishima, "DanceReProducer: An automatic mashup music video generation system by reusing dance video clips on the web," in *Proc. of the 8th Sound and Music Computing Conference (SMC 2011)*, 2011.

[25] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[26] M. Hamasaki, H. Takeda, and T. Nishimura, "Network analysis of massively collaborative creation of multimedia contents: Case study of Hatsune Miku videos on Nico Nico Douga," in *Proc. of the 1st International Conference on Designing Interactive User Experiences for TV and Video (uxTV'08)*, pp. 165–168, 2008.

[27] X.-S. Hua, L. Lu, and H.-J. Zhang, "Automatic music video generation based on temporal pattern analysis," in *Proc. of the 12th annual ACM international conference on Multimedia (ACM Multimedia 2004)*, pp. 472–475, 2004.