Songrium RelayPlay: A Web-based Listening Interface for Continuously Playing User-generated Music Videos of the Same Song with Different Singers

Masahiro Hamasaki
National Institute of Advanced Industrial
Science and Technology (AIST), Japan
masahiro.hamasaki@aist.go.jp

Keisuke Ishida
National Institute of Advanced Industrial
Science and Technology (AIST), Japan
ksuke-ishida@aist.go.jp

Tomoyasu Nakano
National Institute of Advanced Industrial
Science and Technology (AIST), Japan
t.nakano@aist.go.jp

Masataka Goto
National Institute of Advanced Industrial
Science and Technology (AIST), Japan
m.goto@aist.go.jp

ABSTRACT

This paper describes "Songrium RelayPlay," a Web-based user interface for continuously and seamlessly playing back music videos that contain voices of various vocalists singing the same song. Since famous songs often have cover (Me Singing) videos sung by various vocalists on video-sharing services, our interface automatically aligns those videos to their original song to provide a new experience of interactively switching vocalists while listening to the song. Our backend system collects a number of instances of such videos from the Web by means of a Web-mining technique and then our listening interface plays them in relays using signal processing technologies. Even if users listen to a song only once, they can enjoy various singing voices by switching vocalists phrase by phrase (relay-playing). We implemented and publicly launched Songrium RelayPlay where users can enjoy over 18,000 songs having 0.4 million derivative singing videos.

1. INTRODUCTION

With the spread of social media, many popular songs have not only the original version sung by the original vocalist but also derivative (cover) versions sung by various other vocalists [1, 2]. Such derivative versions are often shared as user-generated content (UGC) on video-sharing services and people can enjoy multiple versions of the same song. For example, you can find over 2 million derivative singing videos of *"Let It Go"* [1] on YouTube. Vocalists range from professional to amateur, with some videos being viewed millions of times.

In this paper, we propose a novel Web-based music listening interface called *Songrium RelayPlay* for intelligent playback of a song by concatenating different parts of singing voices sung by various vocalists. We call this style of playing *"relay-playing."* When a user relay-plays cover songs of an original song X, a cover song video A is played for the first 30 seconds, then a cover song video B for the next 30 seconds, then a cover song video C for the next 30 seconds, and so on. Since each video is switched seamlessly without stopping the song, the user can continue listening to a song with multiple singing voices as shown in Figure 1. This style of singing is actually not unique: for example, *"We Are The World"* [2] is a famous song sung by multiple singers. However, "We Are The World" was originally created as an international relay-played song. In contrast, Songrium RelayPlay can relay-play many cover songs on-demand.

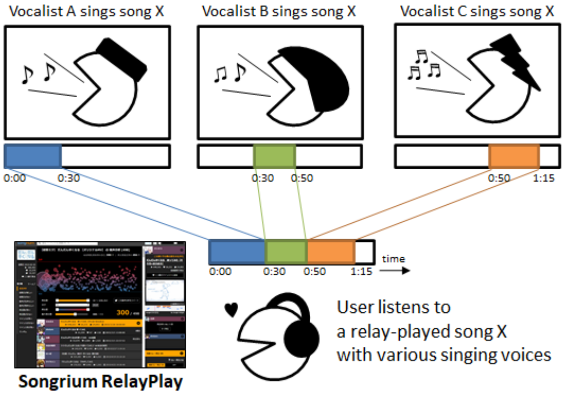

Figure 1. Concept of relay-playing. A song is played back by concatenating different parts of singing voices sung by various vocalists. Listeners can enjoy various singing voices by switching vocalists phrase by phrase.

Relay-playing has the potential to create rich experiences that encourage users not only to enjoy their favorite songs with fresh minds but also to discover new vocalists, as follows:

- People sometimes get bored of repeatedly listening to the same song even if they love it very much at the beginning. Our relay-playing interface could add a fresh taste to such a song and encourage users to listen to it with different vocalists again and again.

[1] https://www.google.com/search?q=%22let+it+go%22+cover+site%3Ayoutube.com

[2] https://youtu.be/Glny4jSciVI

- Since the relay-playing interface plays back singing voices of several vocalists within one song, it could increase opportunities to encounter unfamiliar but potentially interesting vocalists. Even if some singers might not like the basic concept of our interface, as it only partially (not fully) plays back their voices, they could warm to the idea of their increased exposure to potential listeners.

We developed new technologies to overcome difficulties in realizing the relay-playing interface with a focus on massive cover songs on the Web. First, since the interface needs to collect derivative singing videos of the same song on a video-sharing service, we developed a Web-mining technique to analyze unobvious derivative relationships. Second, since derivative singing videos often have different temporal offsets, we cannot simply concatenate their fragments to achieve the seamless playback switch. We therefore developed a new way of combining existing fingerprinting and beat-tracking technologies to align those singing videos with their original song. Third, since users are typically not familiar with a huge number of vocalist names of derivative singing videos, it is almost impossible to manually specify the order (i.e., create a playlist) of those videos one by one. Our interface provides interactive support for creating such a playlist of derivative singing videos of a same song without relying on vocalist names. For example, it uses a signal processing technology to analyze singing voice characteristics so that singing voices could gradually change from female-like voices to male-like voices after switches. It can also utilize the popularity (number of views) of online videos to create a playlist, and we have developed another technology to remove possible temporal biases of the popularity to make this function more effective.

We implemented "Songrium RelayPlay" where users can listen to derivative singing videos on *Niconico* [3], the most popular Japanese video-sharing service, and made it open to the public so that anyone can enjoy our interface [4].

2. RELATED WORK

Some researchers have investigated how to achieve smooth playback and song switching [3], while others have provided the experience of listening to a mash-up that features various overlapping songs [4, 5]. Music recommendation [6, 7] and playlist generation [8, 9, 10] are also popular research topics. None of these approaches, however, have focused on relay-playing to enjoy music while discovering new songs and vocalists.

Tsuzuki et al. [11] proposed Unisoner, an interface for assisting with the creation of derivative choruses in which different vocalists sing the same song. Although the target content is similar to ours, the goal and playing style are different: Unisoner simultaneously plays back several singing voices together to generate choruses, which is different from our relay-play concept. Since singing voices are overlapped, it is more difficult for users of Unisoner to notice singing characteristics of individual vocalists than it is for users of Songirum RelayPlay.

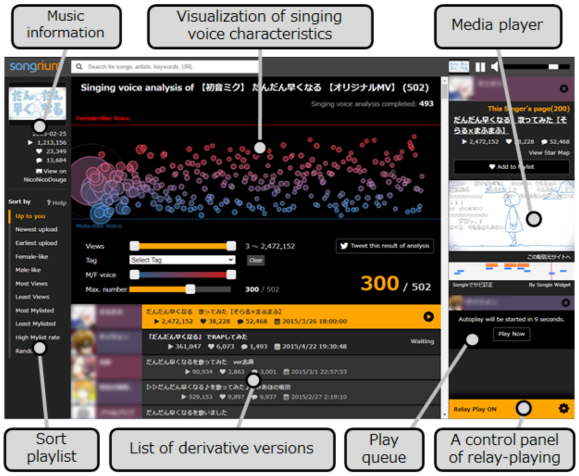

Figure 2. Screenshot of the "Songrium RelayPlay" interface. At the bottom-center, a list of derivative versions is shown. The relay-playing function uses this list as a playlist and plays back singing videos continuously and seamlessly. At the bottom-right, there is a control panel to change behaviors of the relay-playing function. The official embedded video player of the Niconico service, shown at the upper-right, can play back a video clip of the selected cover song. At the top-center, it visualizes cover song videos in the list of derivative versions. Each circle represents one cover song video, where the size of the circle indicates its number of page views and the color indicates the singing voice characteristics. Reddish circles indicate derivative versions with female-like singing voices and bluish circles those with male-like ones. The x-axis indicates a selected sort-criteria parameter which is chosen at the left-middle. Users can choose the number of page views, the number of favorites, publishing date, etc. as the order of a playlist.

3. SONGRIUM RELAYPLAY

We introduce user behaviors in the relay-play of derivative versions sung by various other vocalists. In the Songrium RelayPlay interface, each original song has its own "Song Page" (Figure 2), and a user first searches for and selects a favorite song to access its Song Page showing its derivative singing videos. On this Song Page, the user can start the automatic relay-play of cover songs. The user can additionally specify switching rules, such as "play at least T seconds" and "play until chorus part." At the bottom-center of the Song Page, a list of derivative versions is shown. The relay-playing function uses this list as a playlist and plays back songs continuously and seamlessly. Without specifying individual online singing videos, the playlist can be easily customized by changing various filter parameters, such as the number of page views (for playing back popular or less popular videos only), social tags (for playing back videos having the same tag), and male-like or female-like singing voice characteristics (for playing back videos with female-like voices only). The sorting order of the playback can also be customized by changing various sort-criteria parameters, such as the number of page views, the publishing date, and the number of page views with our original time-based correction (described later). In addition, the user is free to select the next played cover song directly from the list shown on the Song Page. The user can interactively take any of the above actions anytime while listening to relay-playing singing voices, thus easily enjoying the singing voices of a valiety of vocalists.

[3] https://nicovideo.jp
[4] https://songrium.jp/sings

By using this interface, for example, a user could first enjoy listening to a favorite song with various vocalists on the Song Page. If the user encounters an interesting singing voice, the user can also jump to its vocalist's page and listen to other songs sung by this vocalist. By listening to such songs sung by this vocalist, the user could further encounter another interesting song and then come back to its Song Page to relay-play various derivative singing videos of that found song. In this way, the user can repeatedly explore both vocalists and songs without navigating away from the proposed interface.

In our current implementation, users can listen to music video clips on Niconico. "Singing a VOCALOID song" is the target content category of our interface, as it is an extremely popular category of music-related user-generated content on Niconico. A VOCALOID song is an original song composed by using the singing synthesizer named VOCALOID [12]. Over 1,800 VOCALOID songs and 2,900 their cover song videos are uploaded to Niconico every month. Songrium RelayPlay has already collected and analyzed over 18,000 VOCALOID songs and their 0.4 million derivative singing videos. While Niconico is a rich environment in which to enjoy music, it is not always easy for users to encounter favorite songs and derivative singing videos given such a large number of videos. We therefore feel that users of Niconico would appreciate the idea of our interface.

4. IMPLEMENTATION

4.1 Collecting derivative versions

It is difficult to collect derivative versions of the same song because most derivative versions on video-sharing services are not clearly organized. It is a big burden on listeners to find and collect them.

In Songrium RelayPlay, derivative versions are automatically collected and classified so that they can be listed up as a playlist for relay-playing. The backend system classifies music videos into original songs and their derivative versions while analyzing their derivative relationships [1]. So far, the system has found 18,676 video clips as original songs and 465,911 video clips as their derivative versions.

We randomly selected 100 pairs of original songs and derivative versions and manually evaluated the accuracy of the automatic classification. We found that 99 pairs were correct and one was incorrect, but the wrong one was a medley video clip that included the original one. This result suggests that the accuracy is good enough for relay-playing, which is also in accordance with our experience using the proposed interface.

4.2 Estimating temporal offsets between singing videos

The derivative singing videos for a given song are usually based on the same melody and song structure. However, the lengths of the videos are often different, as a video clip may feature a title, a short chat at the beginning, and so on. To synchronously play back such singing videos, the system needs to estimate the temporal offset between the original song and each singing video. Songrium RelayPlay uses two technologies to estimate the offset.

The first one is a fingerprinting technology for audio signals. Audio signals of the derivative singing videos are not identical since different singing voices result in different audio signals. We therefore extend an existing fingerprinting technology to consider only audio signals without vocal sections (singing voices). Since those non-vocal sections (such as introductions and interludes) can be expected to have the same audio signals, the system can appropriately estimate the offset by using the fingerprints of such sections only. Since an arbitrary fingerprinting technology can be used, in our current implementation we use the open-source AUDFPRINT technology [5] .

The second one is a beat-tracking technology that can estimate beat positions from audio signals. If the system cannot estimate the offset by using the extended fingerprinting technology, it tries to estimate the offset by using the difference between a beat in the original song and its nearest beat in each derivative video clip. This method might not be able to estimate the correct offset if the difference is longer than the duration of one beat, but it is still useful for many cases where the timing of beats is almost identical. Since an arbitrary beat-tracking technology can be used, we use an open Web-based API for beat tracking [13].

By taking this two-pronged approach, the system can estimate the offset for each video clip having a different singing voice. We tested the relay-playing function with ten original songs and their ten derivative video clips (100 pairs) at random, and found that the offsets of 73 of the 100 pairs were successfully estimated by using the extended fingerprinting technology and that the median of offsets was 128 msec. The remaining 27 pairs could be dealt with by the second beat-tracking technology.

4.3 Singing voice characteristics

Since most derivative singing videos for a given song have the same accompaniments, excluding some special cases in which accompaniments are performed on different instruments, their differences lie solely in singing voices. Singing voice characteristics such as vocal timbre are therefore important to distinguish them. As shown in Figure 2, Songrium RelayPlay visualizes (changes the color of the circle-icon) derivative singing videos on the basis of their female-like or male-like characteristics.

To estimate these characteristics from video clips, Songrium RelayPlay analyzes audio signals to segregate singing voices and uses a probabilistic estimation model with extracted reliable frames including the vocal part [1].

4.4 Time-based correction for cumulative indices

Page views and the number of favorites are popular indices as a user rating on video-sharing services. Since these indices are cumulative in number, however, older videos tend to have more page views than recent videos even if their popularity might be comparable. We propose using a robust Z-score based on publishing date as a compensation for time-based cumulative indices. The robust Z-score can be obtained by $\frac{(r_x - med_x)}{niqr_x}$, where r_x is page views, med_x is the median, and $niqr_x$ is the normalized interquartile range (NIQR).

[5] https://github.com/dpwe/audfprint

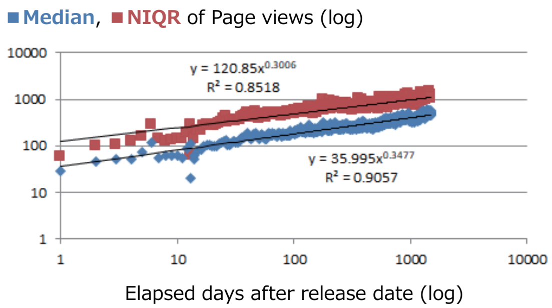

Figure 3. Log-log plot shows median (blue) and NIQR (red) of page views of video clips published since x days ago. Each formula is an approximation formula and R^2 is a coefficient of determination.

To achieve the time correction, we group singing videos by publishing date to obtain a robust Z-score. However, the number of published videos on the same day could vary widely, and very popular videos with extremely large numbers of page views could skew the data. We therefore use curve approximation to estimate the median and the NIQR from the number of days that have elapsed since the time of publishing. Figure 3 shows the results of power approximation of the median and the NIQR, both of which were quite high: $R^2 = 0.90$ and 0.85, respectively.

5. DISCUSSION

Accurate offset estimation is an essential function of our system. However, there are a few cases in which our approach does not work. For example, playing and remixing led to the misestimation due to the difference in BPM and acoustic features. To overcome its limitation, more advanced approaches are required, such as music structure analysis. This will be the focus of our future work.

When three participants tested the relay playback in a preliminary user study, they said that it was favorable as a new way to enjoy music. They also commented that their interest in known songs increased, and they wanted to listen to various new vocalists. On the other hand, there were requests for switching in terms of the song structure, lyrics, etc. In the future, we plan to provide more options for specifying the switching timing.

6. CONCLUSION

We have described the Songrium RelayPlay interface that enables users to continuously and seamlessly play back music videos of various vocalists singing the same song. In our current implementation, it visualizes VOCALOID music including original songs and covers on the Niconico video-sharing service and provides a listening interface for relay-playing and a search interface for exploring content.

For future work, we will continue to run the Songrium RelayPlay service and improve it on the basis of user feedback. In this paper, we focused on VOCALOID music, but derivative versions of the same song are available from many other sources, which we hope to utilize in the near future.

Acknowledgments

This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

7. REFERENCES

[1] M. Hamasaki, M. Goto, and T. Nakano, “Songrium: A music browsing assistance service with interactive visualization and exploration of a web of music,” in *Proc. WWW 2014*, 2014, pp. 523–528.

[2] J. B. L. Smith, M. Hamasaki, and M. Goto, “Classifying derivative works with search, text, audio and video features,” in *Proc. ICME 2017*, 2017, pp. 1422–1427.

[3] H. Ishizaki, K. Hoashi, and Y. Takishima, “Full-automatic DJ mixing with optimal tempo adjustment based on measurement function of user discomfort,” in *Proc. ISMIR 2009*, 2009, pp. 135–140.

[4] M. E. P. Davies, P. Hamel, K. Yoshii, and M. Goto, “AutoMashUpper: An automatic multi-song mashup system,” in *Proc. of ISMIR 2013*, 2013, pp. 575–580.

[5] J. Fan, W. Li, J. Bizzocchi, J. Bizzocchi, and P. Pasquier, “DJ-MVP: An automatic music video producer,” in *Proc. of ACE 2016*, 2016, pp. 1–8.

[6] B. Pardo, Ed., *Special issue: Music information retrieval*, ser. Communications of the ACM. ACM, 2006, vol. 49, no. 8, pp. 28–58.

[7] P. Knees and M. Schedl, “A survey of music similarity and recommendation from music context data,” *ACM TOMM*, vol. 10, no. 1, pp. 1–21, 2013.

[8] Y. Song, S. Dixon, and M. Pearce, “Survey of music recommendation systems and future perspectives,” in *Proc. CMMR 2012*, 2012, pp. 395–410.

[9] T. Nakano, J. Kato, M. Hamasaki, and M. Goto, “PlaylistPlayer: An interface using multiple criteria to change the playback order of a music playlist,” in *Proc. of IUI ’16*, 2016, pp. 186–190.

[10] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani, “Recsys challenge 2018: Automatic music playlist continuation,” in *Proc. of RecSys ’18*, 2018, pp. 527–528.

[11] 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,” in *Proc. of ICMC/SMC 2014*, 2014, pp. 790–797.

[12] 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 uxTV ’08*, 2008, pp. 165–168.

[13] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, “Songle: A web service for active music listening improved by user contributions,” in *Proc. ISMIR 2011*, 2011, pp. 311–316.