Bridging Music via Sound Effects

Yin-Tzu Lin*, Chuan-Lung Lee†, Jyh-Shing Roger Jang*,†, Ja-Ling Wu*†
*Graduate Institute of Networking and Multimedia
†Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
{known,kane0986}@cmlab.csie.ntu.edu.tw, jang@mirlab.org, wjl@cmlab.csie.ntu.edu.tw

Abstract—The prevalence of digital technologies allows people to easily create and share their own media contents, but sometimes we do not have handy tools to manipulate the media we want to create. For example, while creating personal films, a user may separately find the music segment that matches each part of the video, and then concatenate the segments to create the soundtrack that matches the visual contents along the timeline. However, one may lose the temporal coherence between consecutive music segments if the user just choose music segments that best match parts of the video contents. In this study, we focus on how to smoothly connect not-so-coherent music clips and make the transition natural and pleasant to hear. In particular, we improve the temporal smoothness by bar alignment and dual tempo adjustment. To further fit in with the transition between clips, we incorporate “sound effect insertion” which is a commonly used technique in popular song composition/remixing. In order to provide pleasant listening experience and systematically analyze the effectiveness of the proposed music bridging method, we have conducted specifically designed experiments to collect subjective opinions and reduce the cognitive loads of the participants. The experimental results indicate that with proper arrangement for creating smooth transition via tempo adjustment and sound effect insertion, the listening experience can be largely enhanced.

Keywords—sound effect, concatenate music

I. INTRODUCTION

The prevalence of digital technologies allows people to easily create their own media contents and share them with friends through social websites like Facebook, Twitter, etc. These personally created contents relax our mind, vent our emotion, treasure our precious moments, and become more and more important in our daily life. Sometimes, however, we do not have handy tools to manipulate the media we want to create. For example, while making personal films or slideshows, suitable soundtracks usually add much interest to the visual only contents. In order to match the incidental music to the visual content along the timeline, one may, of course, hire a professional composer to make custom music, but the cost is high. One may exhaustively search and fortunately find a piece of music that exactly meet the order and the duration of every pieces of visual content. Or one may just pre-choose a piece of music as the background first and then put visual contents in; nevertheless, the order and the duration of visual contents will be limited for matching the chosen music [1]. Actually, a practical method is to separately find music segments from different music pieces that match to each part of the video, and then concatenate them one by one. Unfortunately, in this way, we may lose the temporal coherent between chosen music segments if we just select music segments that best match certain parts of video contents [2]. The aforementioned shortcoming motivates this study, we intend to remedy the feeling of abruptness between these kinds of chosen segments. The aforementioned problem is related to the automatic DJ system [3], the mashup creation [4], and the medley creation [5]. But in those studies, they often exclude the cases of not suitable-for-combination clips by choosing only timbre or chroma similar segments, while the challenge we are facing here is that we have to handle not only the suitable-for-connecting clips but also the ones that are not that suitable.

In this work, given any two music segments, bar alignment is first applied to generate a more smooth beat counting experience for users. Then, a fine-grained matching up of double/half or quadruple/quarter of the tempo is also considered to manage the situation about the occurrence of large tempo difference between clips. After temporal adjustment, some connected clips are already nice to hear, but some clips still not so smooth. We then incorporate “sound effect insertion” which is a commonly used technique in popular song composition/remixing for bridging the abrupt change between song sections [6]. We also found “sound effect” has been frequently used in producing movie trailers, e.g. in the trailer of the Disney movie: Frozen. The concept of “sound effect insertion” is similar to that of adding shot transition effects in videos, which can bridge the consecutive segments and eliminate the feeling of abruptness. From another point of view, even though we dramatically exaggerated the transition instead of making them smooth, but surprisingly, the listener will feel the music transition is very smooth and nature. To the best of our knowledge, there is no golden rule of how to choose suitable sound effects to bridge various segments from different music pieces. In this study, we try to figure out some general guidelines for choosing sound effects that could help the design of an advanced incidental music creation system.

1https://www.youtube.com/watch?v=FLzFtXQSPBo, 00’12, 00’32, 00’38, etc.
II. RELATED WORK

The studies related to mix different music pieces lie in the fields of automatic DJ system, mashup creation, and medley creation. In the field of automatic DJ systems, most studies dealt with concatenation of given songs by matching beats and rhythms since the resultant music creations are often used as dance music. Basu [7] aligned two clips through scaling and shifting so that the energy of the two songs becomes similar. Jehan [8] realized a DJ system by extracting auditory features, connecting the clips at rhythm-similar segments and aligning the beats of clips. Given a collection of tracks and a tempo trajectory of the tracks, Cliff [9] determined how the track sequence should be played in accordance with their tempi. In [3], audio clips with similar Mel-frequency cepstral coefficients (MFCCs) were first selected. Their tempi were then adjusted by computing their optimal tempo adjustment coefficients (OTACs), and the two clips were then aligned and concatenated by matching the strong beats of the clips.

In the studies of medley creation and mashup creation, music clips suitable for combination are first selected and then are concatenated or overlaid. Lin et al. [10] first filtered out clips with high dissimilarity of loudness, tempo, chroma histogram to the others, then connected the clips at chroma similar positions with tempi adjusted smoothly between consecutive clips. Liu et al. [5] analyzed the user provided collection and constructed a musical dice graph, where the system is able to flexibly generate various medley according user’s specification. Griffin et al. [11] used a phase vocoder to adjust the tempo of every user-specified clip, and overlaid them after synchronizing their beats. AutoMashupper was recently proposed to automatically create mashup music from multiple song tracks. In AutoMashupper [4], users first picked a song track as the basis song, then, the system segmented the picked track into short segments. For each segment, clips with the highest mashability—by chromagram similarity—will overlay with the corresponding phrases to create the final mashup.

To sum up, the studies in automatic DJ system only match or adjust the beat of the connecting songs, and the studies in medley and mashup creation only deal with clips that are similar enough to each other, i.e. those clips are suitable for being connected. In this work, we try to deal with the concatenation of the music segments that are not-so-coherent. In the following sections, two strategies for connecting music segments: temporal adjustment and sound effect insertion will be addressed in details, respectively.

III. TEMPORAL ADJUSTMENT

According to related studies, we knew that tempo smoothness of the connected music is important. All of the previous related studies have handled the tempo issue: beat matching. However, when concatenating music segments with low similarity, matching beats is not effective enough. In addition to beat matching, we further proceed two temporal adjustment steps: bar alignment and dual tempo adjustment.

A. Bar Alignment

As mentioned in [12], the transition between clips should occur at the end points of musical phrases. However, there are many songs contain pick up notes\(^2\). If we directly connect the songs at phrase boundaries, even with the beats matched [5], the connected clips will be still temporally unsmooth (sounds like losing tempo). So we suggest to cut the songs at phrase boundaries but connect the songs with bar alignment. We first use EchoNest API\(^3\) to extract bar and beat information. Then, we proceed the following two processes according to the conditions at the boundaries of phrases. For each pair of consecutive phrase clips \(a\) and \(b\), suppose that there are \(m\) remaining inter-beat intervals (IBIs) after the last bar of clip \(a\), and \(n\) IBIs before the first bar of clip \(b\). Let \(S\) be the average number of IBIs per bar of the former song, which can be treated as the time signature of clip \(a\). Let \(b_a\) and \(b_b\) respectively denote the last bar of the former clip and the first bar of the latter clip. If \(m+n < S\), we align \(b_a\) with \(b_b\), as shown in Figure 1(a), otherwise, if \(m+n \geq S\), we let the number of IBIs between \(b_a\) and \(b_b\) to be \(S\) (see Figure 1(b)).

B. Dual Tempo Adjustment

In order to well handle consecutive music segments that are with large tempo differences. We modified the method proposed in [10] to adjust the tempi for each pair of consecutive clips \(a\) and \(b\). In [10], the system gradually adjusted the tempi from the tempo of clip \(a\), \(T(a)\), to the tempo of clip \(b\), \(T(b)\) to provide temporal smoothness of the connected clips. To do so, a transition length of \(K\) IBIs is determined to ensure that the tempo change ratio at each beat, \(r\), is small enough so that the change in the speed of the songs would not sound abrupt to the listener. However, if the tempo difference of clips \(a\) and \(b\) is large, we may not find a long enough transition length to gradually change the tempi. Besides, the tempo at the overlap location will be far from the original tempi of the segments and lead to explicitly audible artifacts. Moreover, most beat detection algorithms have a common issue with double/half errors [3]. We incorporate a concept similar to what Ishizaki et al. mentioned in [3] to deal with it: take the dual tempo into account. That is, match the IBIs of clip \(a\) to its double/half or to its quadruple/quarter of that of clip \(b\). Accordingly, the bar alignment method presented in Section III-A should be modified as follows. As shown in Figure 2, \(S\), \(m\) and \(n\) are respectively weighted by factors of \(f_a\) and \(f_b\) for clips \(a\) and \(b\), where \(f_a = \arg \min_{i \in 1, 2, 4} |i \cdot T(a) - T(b)|, f_b = \)

\(^2\)One or more notes preceding the first metrically strong beat of a phrase. Also called anacrusis or upbeat [13].

\(^3\)http://echonest.github.io/remix/apidocs/echonest.remix.audio.AudioAnalysis-class.html
tempo values in the slower-tempo clip. So, we up-sample the tempo value sequence and weight the numbers of overlapped IBIs to be the same for both clips. to apply Lin et al. [10]'s algorithm, we should make the tempo value for each IBI instead of the whole clip. Then, grainedly adjust the speed of the clips, the authors calculated how to choose the most appropriate sound effect for any sound effect insertion feasible, we reduce the huge amount of possible sound effects to four recordings that represented major types of percussive instruments – cymbal, drum, wood, and metal. We did not choose pitched FXs because the factors affecting the suitability of inserting pitched sound could be very complicated, e.g. the sound FX needs to match the chord of the two to be connected clips. Our chosen short sound effect snippets are all conform to the characteristics of rising FX [6]. Figure 3 shows the waveforms and the spectrograms of the chosen sound effects. The snippets are all chosen from the sound recording library – EastWest/Quantum Leap Symphonic Orchestra. To reduce the influence of the durations of sound effects, we fixed the major audible parts of all the chosen snippets to about 3 seconds. The effect of the types of sound FX will be detailed in Section V.

### IV. Sound Effect Insertion

After temporal adjustment, some connected clip pairs are already nice to hear, but some clips still not smooth enough. So we incorporate sound effect insertion [6] – a frequently used technique in popular song composition/remixing for smoothing the abruptness between song sections.

#### A. The Rising FX

As mentioned by Langford [6], a rising FX is useful while dealing with abrupt section boundary (e.g. from a breakdown to a chorus) or at the ending. The rising FX is a kind of sound effect that creates a sense of "build up", usually with long duration, rising in volume and getting brighter, sometimes rising in pitch [6]. There are many sound effects have the characteristics of rising feeling, but not all of them can be used for connecting songs. Besides, how to choose the most appropriate sound effect for any two given clips is still an open problem. To make the user evaluation on sound effect insertion feasible, we reduce the huge amount of possible sound effects to four recordings that represented major types of percussive instruments – cymbal, drum, wood, and metal. We did not choose pitched FXs because the factors affecting the suitability of inserting pitched sound could be very complicated, e.g. the sound FX needs to match the chord of the two to be connected clips. Our chosen short sound effect snippets are all conform to the characteristics of rising FX [6]. Figure 3 shows the waveforms and the spectrograms of the chosen sound effects. The snippets are all chosen from the sound recording library – EastWest/Quantum Leap Symphonic Orchestra. To reduce the influence of the durations of sound effects, we fixed the major audible parts of all the chosen snippets to about 3 seconds. The effect of the types of sound FX will be detailed in Section V.

#### B. The Insertion

The sound effect we used are all percussive sounds, so the position where the volume level is maximum can be perceived as an onset of downbeat. As a result, given two

\[
\text{arg min}_{i \in [1, 2, 4]} (|i \cdot T(b) - T(a)|)
\]

(For example, in Figure 2(a), \(f_a = 2\) and \(f_b = 1\) while in Figure 2(b), \(f_a = 1\) and \(f_b = 2\).). Accordingly, \(b_a\) and \(b_b\) should be aligned if \(f_a m + f_b n < f_a f_b S\). On the other hand, if \(f_a m + f_b n \geq f_a f_b S\), we should align \(b_a\) and \(b_b\) in a way to make the number of IBIs (take IBIs of the faster clip as the base) between \(b_a\) and \(b_b\) to be \(f_a f_b S\). In Lin et al. [10]'s algorithm, to fine-grainedly adjust the speed of the clips, the authors calculated a tempo value for each IBI instead of the whole clip. Then, to apply Lin et al. [10]'s algorithm, we should make the numbers of overlapped IBIs to be the same for both clips. So, we up-sample the tempo value sequence and weight the tempo values in the slower-tempo clip \(s\) by its factor \(f_s\). After this pre-processing, we could apply Lin et al. [10]'s algorithm without any change while still taking the dual tempo into account.

![Figure 1. Bar alignment method.](image1.png)

![Figure 2. Bar alignment method.](image2.png)

4The used sound effect snippets can be found in [http://goo.gl/oBWfq8](http://goo.gl/oBWfq8)

5[http://www.soundsonline.com/Symphonic-Orchestra](http://www.soundsonline.com/Symphonic-Orchestra)
connected clips, the insertion location of sound effect is determined by aligning “the position with the maximum volume in the sound effect” with “the first beat onset of the second segment”, as shown in Figure 4.

V. Subjective Evaluation

In the subjective evaluation, we use the 100 best-selling English songs provided by Liu et al. [5]. In the dataset, every song is labeled with its musical phrases and the type of the phrases: vocal or instrumental. An instrumental phrase is a music phrase consisting of purely instrumental sounds, while a vocal phrase, is defined as a singing voice phrase with or without background music [5]. There are total 1409 musical phrases in the dataset. To systematically analyze the effectiveness of connecting methods, we proposed to use the similarity between “the latter clip” and “the phrase after the former clip in the original song” as the metric to measure the suitability of the consecutive clips for connection (c.f. Figure 5). A similar concept has been proposed in [5]. Three types of similarity are used, the chord sequence similarity [5], the timbre similarity (calculated by using the KL divergence of MFCC features), and tempo similarity (calculated based on tempo differences). Based on these metrics, we can measure the performance of different connecting methods on the clip pairs with various connecting-suitability. To choose the clip pairs that cover various suitability, we first compute the three types of similarity on each pair of phrases in the dataset. Then, we divide all the computed similarity values into 3 groups: the highest 30 % similar, the middle, and the lowest 30 % similar. After that, total 20 different types of music clip pairs are chosen, according to the relation between the two clips. The 20 types are composed of 5 similarity types multiplied by 4 different timbre types. The 5 similarity types consist of the clip pairs that their similarity characteristics between “the latter clip” and “the phrase after the former clip in the original song” belong to one of the following situations:

- **LLL**: low similarity\(^6\) in all dimensions (chord, timbre, tempo).
- **LLH**: low chord and timbre similarity, but high tempo similarity.
- **LHL**: low chord and tempo similarity, but high timbre similarity.

\(^6\)in the the lowest 30 % of all the similarity values
connecting clips will be compared, Liu et al. [5], Echonest7.

A. Temporal Adjustment

The 4 timbre types are:

- V-V: both of the clips are vocal.
- V-I: the former clip is vocal, but the latter clip is instrumental.
- I-V: the former clip is instrumental, but the latter clip is vocal.
- I-I: both of the clips are instrumental.

The former clips will be fixed for the five similarity types to reduce the influence of variation. In the following two subjective experiments, we randomly choose 3 clip pairs from each one of the clip pair types, that is, there are 60 clip pairs in total. And for each clip pair, we apply different connecting methods to it, as described in the following subsections. Some of the used test samples can be found in http://goo.gl/oBWfq8.

A. Temporal Adjustment

In this experiment, three temporal adjustment methods for connecting clips will be compared. Liu et al. [5], Echonest7, and the proposed method. In Liu et al. [5], selected clips are connected at phrase boundaries, beats are matched, and tempi are adjusted using the method of [10]. But bar information and dual tempo adjustment are not taken into account. Echonest connects clips at timbre similar positions, matches beats, and adjusts tempi linearly from clip to clip, and does take dual tempo adjustment into consideration. For each one of the 60 chosen clip pairs, we connect them based on the above-mentioned 3 different temporal adjustment methods, and there are 168 resulting connected clips8 in total. We divide the connected clips into 12 groups, each group contains 9 ∼ 15 connected clips of 3 ∼ 5 different similarity types and 1 timbre type. The former clips in the connected samples are the same for each group. Our user evaluations are performed through the aid of a web interface, and the tested clips are presented in random order. Users are invited to listen to one group of clips per time, taking about 10 minutes to finish each test. The questions are designed using a 7-point Likert scale [14]– users are asked to report their opinions of the connected clips from the following options: very pleasing, pleasing, somewhat pleasing, neutral, not so pleasing, not pleasing, and very unpleasing.

46 males and 11 females, aged around 20 to 40, participated in this experiment. Each user involved 1 to 2 groups, and each test sample was listened by 5 different people. Figure 6 shows the mean scores of each one of the connecting methods with all tested clips and with the test clips of each similarity type only. The paired Wilcoxon signed rank test is applied to analyze the results. The corresponding p-values are reported in Figure 6, each line reported the p-values of “Proposed vs. Liu“, “Proposed vs. Echonest“, and “Liu vs. Echonest“, respectively. The overall result shows that the proposed method did improve the smoothness of clip concatenation under a confidence level of 95%. For those low-tempo-similar (LLL, LHL, HLL) clips, the proposed method and Echonest’s approach out perform the method used in [5], which shows that dual tempo adjustment is relatively more important in tempo-dissimilar cases. In those high-tempo-similar (LLH, HHH) cases, the mean score of the proposed method is higher than those of both [5] and Echonest, which indicates that bar alignment did improve the temporal adjustment since, now, it is no need to apply dual tempo adjustment to high-tempo-similar clips.

B. Sound Effect Insertion

In this experiment, we examine the effects of sound FX insertion and the type of sound FX used. Five methods are tested: no sound effect, cymbal sound insertion, drum sound insertion, wood sound insertion and metal sound insertion. For each one of the 60 chosen clip pairs we mentioned in Section V, we connect them with the aforementioned 5 sound FX insertion methods. Total 295 connected clips are generated9. The number of test samples in this experiment. Each user involved 1 to 2 groups, and each test sample was listened by 5 different people. Figure 6 shows the mean scores of each one of the connecting methods with all tested clips and with the test clips of each similarity type only. The paired Wilcoxon signed rank test is applied to analyze the results. The corresponding p-values are reported in Figure 6, each line reported the p-values of “Proposed vs. Liu“, “Proposed vs. Echonest“, and “Liu vs. Echonest“, respectively. The overall result shows that the proposed method did improve the smoothness of clip concatenation under a confidence level of 95%. For those low-tempo-similar (LLL, LHL, HLL) clips, the proposed method and Echonest’s approach out perform the method used in [5], which shows that dual tempo adjustment is relatively more important in tempo-dissimilar cases. In those high-tempo-similar (LLH, HHH) cases, the mean score of the proposed method is higher than those of both [5] and Echonest, which indicates that bar alignment did improve the temporal adjustment since, now, it is no need to apply dual tempo adjustment to high-tempo-similar clips.

---

7https://github.com/echonest/remix/tree/master/examples/capsule

---

8Some clip pairs contain too short phrases so that the Echonest method cannot produce the result. So we remove them.

9Some clip pairs contain too short phrases, too add sound effect, so we remove them.

10We divide the 5 similarity types into 2 groups: “LLL, HHH” and “LLH, LHL, HLL”
The result indicate that, sound FX insertion did improve the smoothness of connected clips.

We then separately investigate every type of sound effects. Figure 8 shows the mean scores of each method with all data and with respect to each one of the 5 similarity types. We also applied paired Wilcoxon signed rank test to analyze the results. The corresponding p-values are reported in Figure 6, each line reported the p-values of “cymbal vs. no effect”, “drum vs. no effect”, “wood vs. no effect”, and “metal vs. no effect”, respectively. The results show that cymbal and metal sounds insertion outperform other methods in all conditions. This indicates that these two sounds are useful in smoothing the connecting clips for the used English song dataset, while drum sound is not so useful. For the lowest similarity case (LLL), the obtained user satisfaction of test samples with cymbal insertion is as high as that of the test samples without sound effects in higher similarity cases. By contract, in the highest similarity case (HHH), the sound effect insertion did not improve the listener’s satisfaction much. The user satisfaction even goes down if the inserted sound effect is not suitable (e.g. drum) for the connected clips.

C. Discussion

We further analyze the results by comparing connected clips across different timbre types, as shown in Figure 9. The case of I-I obtains the highest scores in no sound FX inserted test samples. When sound FXs are inserted, the cases of I-I still have high scores. In addition, we find another high scored case: V-V. Notice that the drum sound obtains higher scores only in V-V cases. As a result, sound effect maybe more suitable to be inserted for V-V cases.

To sum up, inserting sound effects is effective in making the connected clips heard smoother. However for high similarity cases, the improvement is limited. Cymbal and metal sounds are generally quite suitable for being inserted to the used English song dataset while drum sound is not. The probable reason is the used drum sound has prominent
onsets, which is sensitive to the appearance position of the sound effect and the correctness of beat detection. Besides, adding sound effect to V-V cases may be more suitable than other cases. The degree of suitability between sound effect and other characteristics of connected clips, e.g. styles, genres and emotions, needs to be further investigated in our future work.

VI. CONCLUSION

In this study, we focus on smoothly connect music clips which are not that suitable for concatenation, traditionally. We improve the temporal smoothness by using the techniques of bar alignment and dual tempo adjustment. To further fit in the transition between clips, we incorporate “sound effect insertion”, which is a commonly used technique in popular song composition/remixing. In order to systematically analyze the effectiveness of the proposed approach, we conducted experiments to collect subjective opinions and reduced the cognitive loads. The experiment result shows that the proposed method is effective in creating smooth transition via tempo adjustment and sound effect insertion. Many aspects can be further investigated. For example, we only discussed the effect of sound FX types and the relations between two connected clips, while the volume ratio of sound effect to music clips, the duration of sound effect, and the relation between the sound FX and the connected clips have not been covered. Besides, the influence of the style/genre/emotion of the music clips to the choice of sound effects are worthy of further investigations. Finally, more complex sound FXs can be included, such as pitched sound effects, or longer rhythmic sound FXs that played along both of the two connected clips, which will make the results sound more like human made.

REFERENCES