AUTOMATIC SINGING VOICE TO MUSIC VIDEO GENERATION VIA MASHUP OF SINGING VIDEO CLIPS

Tatsunori Hirai
Waseda University
tatsunori_hirai@asagi.waseda.jp

Yukara Ikemiya
Kyoto University

Kazuyoshi Yoshii
Kyoto University

Tomoyasu Nakano
National Institute
of Advanced Industrial
Science and Technology
(AIST)

Masataka Goto
National Institute
of Advanced Industrial
Science and Technology
(AIST)

Shigeo Morishima
Waseda Research Institute
for Science and Engineering
/ CREST, JST
shigeo@waseda.jp

ABSTRACT

This paper presents a system that takes audio signals of any song sung by a singer as the input and automatically generates a music video clip in which the singer appears to be actually singing the song. Although music video clips have gained the popularity in video streaming services, not all existing songs have corresponding video clips. Given a song sung by a singer, our system generates a singing video clip by reusing existing singing video clips featuring the singer. More specifically, the system retrieves short fragments of singing video clips that include singing voices similar to that in target song, and then concatenates these fragments using a technique of dynamic programming (DP). To achieve this, we propose a method to extract singing scenes from music video clips by combining vocal activity detection (VAD) with mouth aperture detection (MAD). The subjective experimental results demonstrate the effectiveness of our system.

1. INTRODUCTION

Many people consume music by not only listening to audio recordings, but also watching video clips via video streaming services (*e.g.*, YouTube[1]). Thus, the importance of music video clips has been increasing. Although a lot of music video clips have been created for promotional purposes, not all existing songs have their own video clips. If a video clip could be added to an arbitrary song, people could enjoy their favorite songs much more. Note that one of the most important parts of popular music is the vocal part. Thus, to enrich music listening experience, the automatic generation of "singing" video clips for arbitrary songs is a big challenge worth tackling.

Since there are a large number of music video clips available on the Web, these clips can be considered as an audio-visual dictionary covering almost all sound events. Given an audio clip, we could figure out what happens in a visual manner by searching for a video clip including similar sounds. The key idea in this paper is that we could make a music video clip for an arbitrary song by searching for video clips including singing voices that are acoustically similar to that in the target song.

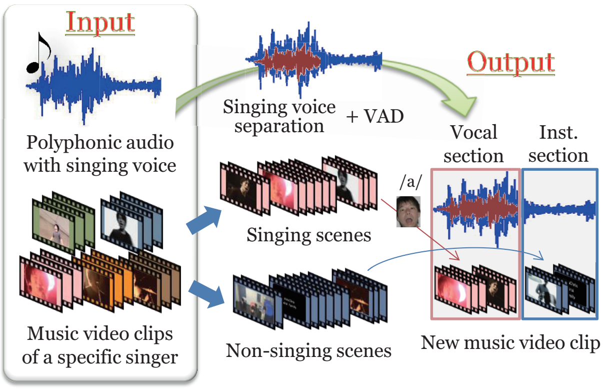

Figure 1. Conceptual image of our system.

To achieve automatic video generation, it is important to search for similar singing voices in a database of existing music video clips. If a similar singing voice can be found in an existing clip within the database, the singing actions of the singer in the clip can be expected to match the input singing voice. In this paper, we try to find multiple short video fragments that match the input singing voice and concatenate these fragments together to make a new singing video clip. As typical music video clips include a number of scene changes, the system output is allowed to contain frequent scene changes in output clips, as long as the singer remains unchanged.

Another solution to automatic singing video generation is to construct an audio-visual association model for lip sync. This approach requires clean video clips (e.g., simple background, stabled camera and target) recorded in an ideal environment for the precise analysis of audio-visual association. However, real video clips are noisy and are difficult to construct a reliable model. We aim to deal

[1] http://www.youtube.com/

with real video clips rather than video clips recorded in an ideal environment. It is extremely difficult to precisely detect objects in real video clips. Therefore, constructing a model from such clips is an unreasonable approach. We achieve automatic "singing voice to singing video" generation by focusing on singing voices and acoustic similarity between these voices. Fig. 1 shows a conceptual image of our system. Given an input song sung by an arbitrary singer and existing music video clips in which the singer appears, our system automatically generates both singing and non-singing scenes by mashing up[2] existing singing and non-singing scenes. This paper has two main contributions: audio-visual singing scene detection for music video clips, and singing video generation based on singing voice similarity and dynamic programming (DP).

2. RELATED WORK

The research topic of automatic music video generation has recently become popular, in fact a competition was held at the ACM International Conference on Multimedia 2012. Several methods have been proposed for the automatic generation of music video clips by focusing on *shallow* audio-visual association [1–4]. Foote *et al.* [1] proposed an audio-visual synchronization method based on audio novelty and video unsuitability obtained from camera motion and exposure. Hua *et al.* [2] proposed a system of automatic music video generation based on audio-visual structures obtained by temporal pattern analysis. Liao *et al.* [4] proposed a method to generate music video by extracting temporal features from the input clips and a piece of music, and casts the synthesis problem as an optimization. Although these methods consider audio-visual suitability [1], temporal patterns [2], or synchronization [4], higher-level information (*e.g.*, the semantics of video clips) was not taken into account. Nakano *et al.* [3] proposed a system called *DanceReProducer* which automatically generates dance video clips using existing dance video clips. For audio-visual association, the system uses an audio-to-visual regression method trained using a database of music video clips. Since they use low-level audio and visual features for regression, higher-level information (*e.g.*, dance choreography) was not taken into account.

In this paper, we tackle the problem of audio-visual synchronization between a singing voice and a singer's singing action (*e.g.*, lip motion). This enables more *semantic* synchronization than conventional methods. Yamamoto *et al.* [5] proposed a method that automatically synchronizes band sounds with video clips in which musical instruments are being played. Although this method can be considered as semantic synchronization, it does not mention synchronization of the vocal part, and requires manual input for sound source separation. In contrast, we focus on the vocal parts of a song and automate all processes.

In terms of synchronizing voice and lip motion, lip sync animation has been intensively studied in the field of computer graphics (CG). Various lip sync methods have been applied to 3DCG characters, including image-based photo-realistic human talking heads [6–9]. Basically, these talking heads are obtained by 3D face reconstruction or 3D face capture, methods that are not easily applicable to general video clips. Therefore, to make a video clip of a specific singer, users must prepare an ideal frontal face image or a sufficient amount of ideal video sequences of the singer. Our goal is to use the abundance of existing video clips so that users do not need to record new data to generate a video clip.

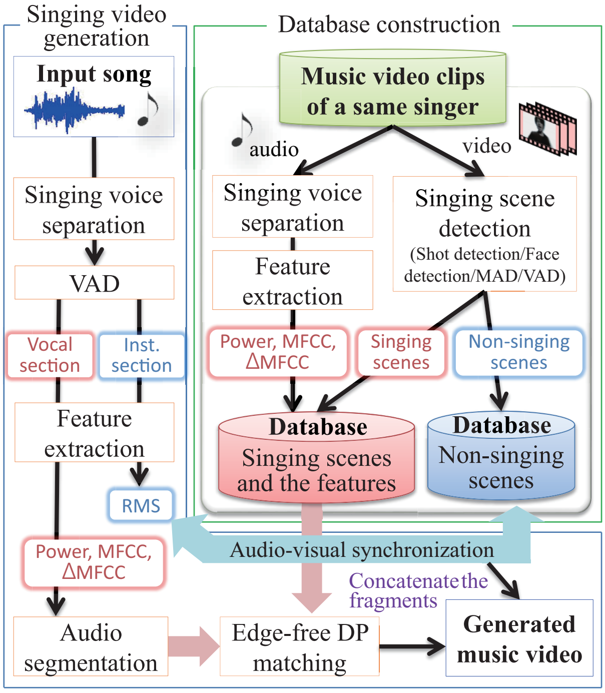

Figure 2. Overview of our system.

To mash up existing music video clips for synthesizing new singing video clips, we require a method to automatically detect singing scenes. Video event detection is a popular research topic in both the multimedia and pattern recognition communities. Many promising video analysis methods have been presented at the International Workshop on Video Retrieval called TRECVid [10]. The test data for the semantic indexing task in TRECVid include video clips with the label "Singing." To distinguish singing events from other events, most teams used acoustic features such as the mel frequency cepstral coefficient (MFCC) in addition to video features. These methods were designed for general-purpose event detection, not for the specific detection such as singing scene detection. To extract a target activity (*e.g.*, singing) from music video clips, we propose an automatic singing scene detection method that constructs a database of such scenes from existing music video clips.

3. SYSTEM IMPLEMENTATION

Our system consists of two processes: database construction and singing video generation. The system flow is shown in Fig. 2. The only data required by the system are the input song and music video clips for a database.

[2] The term "mashup" refers to the mixture of multiple existing video clips.

The database video clips must include singing scenes of the singer of the input song. A larger number of database clips will result in better output video quality.

To construct the database, singing scene detection will be applied to music video clips. Specifically, we employ an algorithm that combines vocal activity detection (VAD) from polyphonic musical signal and mouth aperture detection (MAD) based on facial recognition in a video clip. At the same time, singing voice separation is applied to the audio part of the database music video clips, and the singing voice feature is extracted. As a result, the singing scenes in the database clips and the singing voice and the features will be stored as a database for the system.

The singing video generation starts with singing voice separation and VAD. At this point, an input singing voice and the singer's singing scenes with the singing voices are available. For the vocal section of an input song, the system searches for acoustically similar singing voices from a database of singing scenes. For example, if part of the input singing phrase (query) is "oh," the system searches for a singing voice with a similar sound, such as the "o" from the word "over" or "old," on the basis of the similarity of singing voice features. Note that the system does not consider lyrics. It is difficult to find good matches between longer queries and the database. Therefore, the output singing video will be a mashup of small fragments of singing scenes. The length of each fragment will be automatically determined on the basis of the automatic singing voice segmentation. For the instrumental section of an input song, the system automatically adds the best synchronized video fragments. These are calculated on the basis of the matching between the accents of both the input song and the database clips. In this case, reference video scenes will be narrowed down to the non-singing scenes in the database clips. By mixing separately generated singing scenes and non-singing scenes, system generates a new music video clip. Further details of each process will be described in a later section.

4. SINGING SCENE DETECTION

We apply singing scene detection to music video clips for database construction. The singing scenes in a music video clip are one of the highlights of the clip. We define a singing scene as one in which a singer's mouth is moving, and the corresponding singing voice is audible. Therefore, not only audio analysis but also video analysis is necessary to detect such scenes. Our approach is to combine the existing VAD method [11] with a new MAD method that is customized for handling faces in a video clip by using continuity of video frames.

4.1 Vocal activity detection (VAD)

VAD is applied to the polyphonic audio signal of a music video clip. We apply the HMM-based VAD method proposed by Fujihara *et al.* [11]. This method trains models for both vocal and non-vocal states by GMM. Using HMM to express the audio signal as a transition of both states, the method detects vocal sections on the basis of probability.

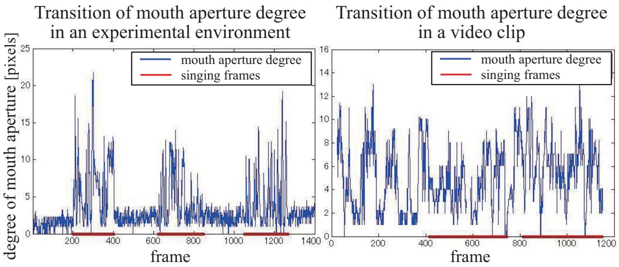

Figure 3. Transition of the mouth aperture degree in an experimental environment and a real video clip. Manually labeled singing scenes are shown in red.

4.2 Mouth aperture detection (MAD)

To detect mouth activity, we require a face recognition method for video clips. For facial detection, we apply the method proposed by Irie *et al.* [12]. This method uses a global fitting of the active structure appearance model (ASAM) to find face areas and a local fitting model to detect each facial part and facial feature points. Therefore, the feature points of the mouth can be detected.

This face detection method is comparatively robust to facial expressions and transitions in the direction of the face. However, the detection of facial feature points in a music video clip is difficult, as there is considerable noise in the detected results. Fig. 3 illustrates the appearance of noise in a real music video clip by comparing with the mouth aperture degree in a video recorded in an experimental environment. Hence, we cannot directly use the mouth aperture degree based on the mouth feature points to detect singing scenes in a real video clip. Instead of directly using the mouth aperture degree, we use the standard deviation of the distance between the upper and lower lip in each consecutive sequence of the same person's face as a mouth feature.

To acquire suitably consecutive face sequences, the system detects shot boundaries of a video clip. A shot is a consecutive video sequence that has no scene changes or camera switches. According to the continuity of video frames, a person in one shot is always the same when there is no movement of the person or camera. However, even if there is a movement of the person or camera, the same person can be captured by tracking their movement. To detect the boundary of a shot, the system subtracts consecutive luminance histograms, and uses their summation as a shot detection feature. When the value of the feature is higher than other values, the frame is considered to be a shot boundary. It is possible to detect the same person's face by tracking the spatial trajectory of the detected face across one shot. If the mouth feature is greater than a threshold[3] , the system classifies the shot with the entire consecutive face sequence as a singing scene.

[3] The value is experimentally fixed to 10 when the face size is normalized to 512×512 pixels.

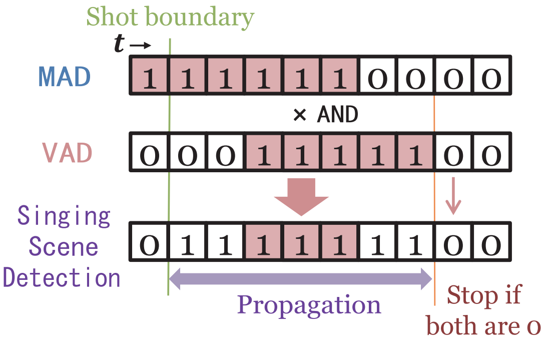

Figure 4. Combination of detected results.

	Precision	Recall	F score
VAD	0.632	0.732	0.672
MAD	0.609	**0.823**	0.677
VAD+MAD	**0.662**	0.759	**0.690**

Table 1. The accuracy of singing scene detection.

4.3 Combination of VAD and MAD

Our singing scene detection method combines the results of both VAD and MAD. Fig. 4 illustrates our method of combining the results. Both VAD and MAD results can be expressed in binary (1 denotes singing, and 0 denotes not singing). By taking the logical AND of both results, we can classify singing scenes with high reliability, however, only part of them can be detected. To detect more singing scenes, we use the continuity of video frames. The method propagates the results to consecutive frames until both the VAD and MAD results of the frame are 0.

To evaluate the accuracy of our singing scene detection method, we performed an experiment to detect singing scenes in real music video clips. We manually add annotations to 10 music video clips of professional musicians, and compared the detection accuracy with VAD and MAD. Table 1 lists the average accuracy of singing scene detection with each method. These results show that our combined algorithm gave the best performance in terms of accuracy (F score).

5. SINGING VIDEO GENERATION

To generate singing video, users prepare an arbitrary input song for singing video generation and the same singer's existing music video clips for database construction. Here, singing scenes should be included in database clips. The system calculates the similarities between the input singing voice and the database singing voices to search for well-synchronized singing sequences. Previous studies on talking heads ([13], [14]) have used both audio and visual features to construct an audio-visual association model from which a talking head is generated. In contrast, we use audio information alone to retrieve video fragments. This is because the extraction accuracy of visual features, especially the mouth feature, is not high enough to construct an audio-visual model from real music video clips. Most talking head research uses video captured in an experimental environment which is different from real music video clips. Therefore, we do not use visual features in generation part but focus on an acoustic similarity of singing voices.

5.1 Database construction

Our system constructs a database from the user-prepared music video clips. The database includes the singing scenes from the clips and the singing voice features extracted from the audio. The singing scenes are extracted with our singing scene detection method, and singing voice separation is performed on the audio part of the clips.

5.1.1 *Singing voice separation*

To extract singing voice features, a singing voice separation method is required. We apply the singing voice separation method proposed by Ikemiya *et al.* [15], which achieved the best separation performance in the Music Information Retrieval Evaluation eXchange (MIREX2014), a singing voice separation task.

This method uses a robust principal component analysis (RPCA) to separate non-repeating components, such as a singing voice, from a polyphonic spectrogram. By estimating the F0 contour from the separated components including the singing voice, we can obtain a binary mask that passes only the harmonic partials of the F0 contours. This method further improves the singing voice separation accuracy by combining the binary masks obtained using RPCA and F0 harmonics.

5.1.2 *Singing voice feature extraction*

From a separated singing voice, the system extracts the singing voice features that represent the characteristics of a singer's voice and prosody. Our goal is to generate a singing video that is well-synchronized to an arbitrary input song. The lyrics are an important factor in synchronizing a singing voice and a singing video, especially with regard to the motion of the mouth. However, it is difficult to obtain lyrics that are aligned with the audio signal for all existing songs. Therefore, we employ the MFCC which is related to prosody, and the power which is related to the dynamics of singing voice, as the singing voice feature.

The system extracts the 12 dimensional MFCC (excluding zeroth order which corresponds to the power), ΔMFCC, and the one dimensional power of the audio signal from the singing voice (25 dimensions in total).To realize better lip-sync, we handle power feature separated from MFCC features. The length of audio analysis frame and the analysis time step is 1/29.97 seconds in order to synchronize audio and video analysis time step. At this point, the singing voice feature values are normalized to have a mean value of 0 and a variance of 1.

Thus, singing scenes from the user-prepared music video clips and the singing voice features can be stored in the database.

5.2 Video fragment retrieval

From the input song sung by the same singer as database clips, the system retrieves singing video fragments that synchronize well with the input song. Because the input and database clip are of the same singer, we expect the mouth shape to be alike for similar singing voice features. To search for singing video fragments with similar singing voices, we extract the same 25 dimensional singing voice features from the singing voice of the input song as in the database. After the feature extraction, VAD is carried out to determine the section to which the singing video should be added.

Because our system will be used to generate new music video clips that do not exist in the database, there is no chance of finding a video clip with perfect synchronization. Therefore, we search for small fragments of singing video clips, and concatenate these fragments to achieve good synchronization. The intensity of synchronization is a trade-off between the temporal consistencies in the output clip. As we imagine, the output clip will be visually inconsistent, because it is a mixture of multiple video clips. However, many music video clips consist of more than one scene and the occurrence of frequent scene changes is not unusual in case of music video clips.

The length of each fragment is automatically determined by the singing voice segmentation. By manually specifying the minimum and maximum fragment lengths, the system automatically finds segmentation boundaries of an input singing voice on the basis of the power of the singing voice. This segmentation is performed by searching for the minimum power point within a user-specified range. Thus, singing voice can be segmented based on the phrases.

To retrieve the best-synchronized singing scene fragments, we employ edge-free DP. Though the normal DP matching require two sequences to be the same length, edge-free DP searches for the shortest path with arbitrary length. We fixed the length of the input singing voice feature and made the length of database features variable with edge-free DP. The Euclidean distance between the input and database singing voice features is used as the cost of DP. A weight is added to the cost in order to change the priority of the MFCC and the power. Adding larger weight to the power realizes better synchronization in terms of timing. Whereas the MFCC features affect prosody (shape of mouth), the power feature affects the onset and offset (timing) of a voice which is important in lip syncing. We assign a weight of 0.2-0.5 (20-50%) to the power feature, and spread the rest across the MFCC features (sum of the weight will be 1.0 in total). The system searches for a database singing scene with the minimum cost for each input singing voice fragment by shifting the DP start point in the database clips frame by frame. Edge-free DP makes it possible to adopt the end point of DP on the basis of the cost, so that the start point is fixed but the end point of DP depends on the cost. By concatenating all the retrieved fragments, the system generates the singing video output. Although we do not consider the mouth shapes of a singer, the mouth shape in a retrieved fragment tends to correspond to the phoneme of the input singing voice.

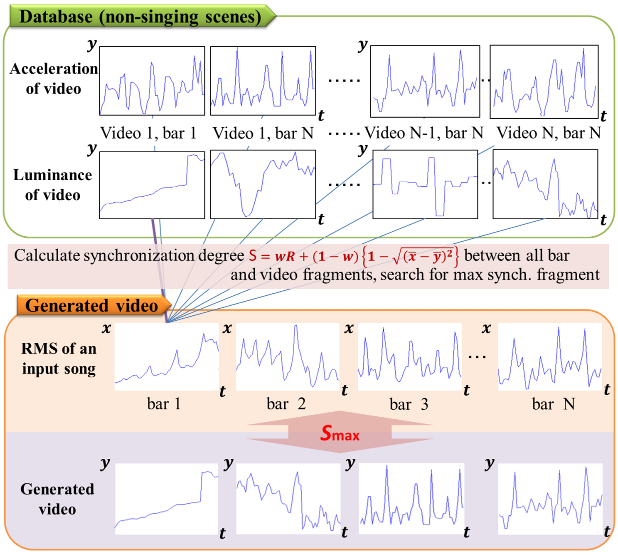
Database (non-singing scenes)
Acceleration of video
Luminance of video
y
t
Video 1, bar 1
Video 1, bar N
Video N-1, bar N
Video N, bar N
Calculate synchronization degree $S = wR + (1-w)\left\{1 - \sqrt{(\bar{x}-\bar{y})^2}\right\}$ between all bar and video fragments, search for max synch. fragment
Generated video
RMS of an input song
x
bar 1
bar 2
bar 3
bar N
S_{max}
Generated video

Figure 5. Generation of non-singing scenes.

5.3 Non-singing scene generation

The generation of singing scenes alone is not sufficient for music video generation. We therefore implement an automatic music video generation method for non-singing scenes to complement that for singing scenes. We employ the metric proposed by Hirai *et al.* [16], which considers the synchronization between the music and the video based on a subjective evaluation. This synchronization method considers accents in the video, such as the acceleration or transition in luminance, and the audio accent of the input music, which is calculated by the root mean square (RMS) of the audio signal. By applying this method to instrumental sections detected by VAD, non-singing scenes can be generated. Here, the database for this part only contains non-singing scenes from the database video clips, and the length of each fragment is fixed to one musical bar of the input song. Here, the length of one musical bar is extracted based on the same method as Hirai *et al.* [16].

Fig. 5 shows how these non-singing scenes are added to the instrumental sections of the input song. The synchronization degree S represents a measure for calculating the synchronization, as proposed in [16].

5.4 Mixing singing and non-singing scenes

To generate the final output music video clip, we mix the singing and non-singing scenes. However, the generation of singing scenes relies on the detection of singing scenes in database clips and the separation of the singing voice, which are not perfect. Thus, some mistakes may occur in these processes. For example, the system may detect an instrumental section as a vocal section in the VAD process, leading to the generation of singing scenes for instrumental sections. Therefore, we only use reliable singing scenes.

To improve the output clip, the system replaces unreliable singing scenes with non-singing scenes. This reliability may be either the DP cost, which represents the synchronization degree, or the VAD likelihood, which represents the reliability of VAD. Because there is no rule that singing scenes must be accompanied by a vocal section in a music video clip, we can instead use better synchronized fragments. This process makes it possible to generate only well-synchronized singing scenes, and replace other scenes with acceptable (non-singing) scenes.

6. EVALUATION

We performed a subjective evaluation experiment to compare the generated results with music video produced by another method. Twenty subjects were asked to watch the music video clips automatically generated by our proposed method and the comparison method by Hirai *et al.*([16]), and to determine which was better in terms of audio-visual synchronization. All subjects are not the professional of music video editing. Since we used the comparison method to generate non-singing scenes, the generated video clips using the same song are the exactly same in non-singing scenes. The differences between the clips generated by each method are therefore in the singing scenes. As input, we used five songs by one singer, and constructed the database from clips of this singer's music video clips. The subjects watched a total of ten 30-second video clips, and scored them from 1 (Hirai *et al.* 2012 is better) to 5 (proposed method is better) by comparing the clips generated with each method in an aspect of audio-visual synchronization. The clips show the beginning of the songs, and all include an instrumental section.

Fig. 6 shows the evaluation score for each song. The baseline is 3.00, and higher scores indicate that proposed method is better than the comparison method. The evaluation score for each song is the average of all subjects' evaluation scores, and the average of the five songs is 3.66. Seventeen out of twenty subjects pointed out that the synchronization between singing voice and the singer's mouth is one of the factors for the audio-visual synchronization. From this result, we can say that our method is better with regard to singing scene generation. However, some subjects mentioned that the frequent scene changes and unnatural scene transitions were distracting. It is difficult to evaluate music video clips because people focus on many factors within a single clip. Taking this into account, the evaluation results suggest that our method exhibits respectable audio-visual synchronization.

7. CONCLUSION

This paper presented a system that can automatically generate a "singing" video clip for an arbitrary song. The main contribution of this work is our proposal for an automatic generation method that employs singing voice similarity and edge free DP to semantically synchronize singing scenes from existing video clips with a singing voice. Our automatic detection technique for singing scenes in music video clips, especially the MAD technique based on the standard deviation of the consecutive face sequences, represents another contribution.

Our future work is to handle inter-singer generation. The current singing voice features limit our method, as the singer's voice of the database video clips must sound similar to the voice of the input song, because the feature values tend to change with each individual. Initial tests suggest that inter-singer generation is feasible between two singers with a similar voice, but this is not the case with voices that are not similar. These problems may be solved by applying a voice conversion technique. In the future, we aim to generate a music video clip in which an arbitrary singer sings an arbitrary song. It might be possible to create a video clip in which a deceased singer sings a new song, which we can sometimes see in a film concert.

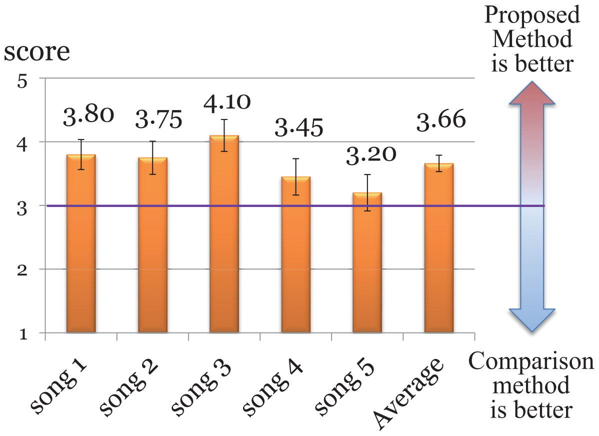

Figure 6. Subjective evaluation results comparing proposed method and comparison method by Hirai *et al.* 2012 [16].

Our future work will also include semantic audio-visual synchronization that is not limited to a singing voice, but instead considers other audio-visual objects (events), using existing video clips. Because we have access to large numbers of video clips, the style of watching might change if we can use them to create new experiences. Extending our framework, it might be possible to add video for unknown sounds to help us understand their sound source. We are investigating how music video clips can be reconstructed using existing clips, and are on the way to achieving this. The time may come when people will enjoy automatically generated digital content as well as human created content.

Acknowledgments

This work was supported by OngaCREST, CREST, JST and partially supported by JSPS Grant-in-Aid for JSPS Fellows.

8. REFERENCES

[1] J. Foote, M. Cooper, and A. Girgensohn, "Creating music videos using automatic media analysis," in *Proc. ACMMM*, 2002, pp. 553–560.

[2] X.-S. Hua, L. Lu, and H.-J. Zhang, "Automatic music video generation based on temporal pattern analysis," in *Proc. ACMMM*, 2004, pp. 472–475.

[3] T. Nakano, S. Murofushi, M. Goto, and S. Mor-

ishima, “DanceReProducer: An automatic mashup music video generation system by reusing video clips on the web,” in *Proc. SMC*, 2011, pp. 183–189.

[4] Z. Liao, Y. Yu, B. Gong, and L. Cheng, “audeosynth: Music-driven video montage,” *ACM Transactions on Graphics (SIGGRAPH)*, vol. 34, no. 4, 2015.

[5] T. Yamamoto, M. Okabe, Y. Hijikata, and R. Onai, “Semi-automatic synthesis of videos of performers appearing to play user-specified music,” in *Proc. WSCG*, 2013, pp. 179–186.

[6] C. Bregler, M. Covell, and M. Slaney, “Video rewrite: Driving visual speech with audio,” in *Proc. SIGGRAPH*, 1997, pp. 353–360.

[7] S. Kawamoto *et al.*, “Galatea: Open-source software for developing anthropomorphic spoken dialog agents,” in *Life-Like Characters*, ser. Cognitive Technologies. Springer Berlin Heidelberg, 2004, pp. 187–211.

[8] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, “Expressive visual text-to-speech using active appearance models,” in *Proc. CVPR*, 2013, pp. 3382–3389.

[9] L. Wang and F. Soong, “Hmm trajectory-guided sample selection for photo-realistic talking head,” *Multimedia Tools and Applications*, pp. 1–21, 2014.

[10] A. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in *Proc. MIR*, 2006, pp. 321–330.

[11] H. Fujihara, M. Goto, J. Ogata, and H. Okuno, “Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, pp. 1252–1261, 2011.

[12] A. Irie, M. Takagiwa, K. Moriyama, and T. Yamashita, “Improvements to facial contour detection by hierarchical fitting and regression,” in *Proc. ACPR*, 2011, pp. 273–277.

[13] B.Theobald and N. Wilkinson, “On evaluating synthesised visual speech,” in *Proc. Interspeech*, 2008, pp. 2310–2313.

[14] S. Deena, S. Hou, and A. Galata, “Visual speech synthesis using a variable-order switching shared gaussian process dynamical model,” *Multimedia, IEEE Transactions on*, vol. 15, no. 8, pp. 1755–1768, 2013.

[15] Y. Ikemiya, K. Yoshii, and K. Itoyama, “Singing voice analysis and editing based on mutually dependent f0 estimation and source separation,” in *Proc. ICASSP*, 2015.

[16] T. Hirai, H. Ohya, and S. Morishima, “Automatic mash up music video generation system by perceptual synchronization of music and video features,” in *Proc. SIGGRAPH (poster)*, 2012.