Semantic Segmentation of Documentary Video Using Music Breaks

Aijuan Dong and Honglin Li
Department of Computer Science, North Dakota State University Fargo, ND 58105
{aijuan.dong, honglin.li}@ndsu.edu

ABSTRACT

Many documentary videos use background music to help structure the content and communicate the semantic. In this paper, we investigate semantic segmentation of documentary video using music breaks. We first define video semantic units based on the speech text that a video/audio contains, and then propose a three-step procedure for semantic video segmentation using music breaks. Since the music breaks of a documentary video are of different semantic levels, we also study how different speech/music segment lengths correlate with the semantic level of a music break. Our experimental results show that music breaks can effectively segment a continuous documentary video stream into semantic units with an average F-score of 0.91 and the lengths of combined segments (speech segment plus the music segment that follows) strongly correlate with the semantic levels of music breaks.

1. INTRODUCTION

Recent advances in computer power, network bandwidth, information storage, and signal processing techniques have led to a proliferation of video data. To support flexible video indexing and ensure effective exploitation of these video assets, the first and critical step is to segment videos into semantically tractable units.

Video segmentation has been researched for many years. Earlier research in this field has focused on video visual features. Recently, researchers have begun to realize that audio information, including speech and acoustic properties, is as important as visual information in video content understanding. Audio-based processing requires much less complex processing and audio analysis results can always aid visual processing in video segmentation. In this paper, we are particularly interested in investigating music breaks in semantic segmentation of documentary video.

Two levels of audio features have been studied in audio content analysis: short-term frame level and long-term clip level [1]. An audio frame is defined as a group of neighboring samples that last about 10 to 40 ms, within which we can assume that the audio signal is stationary. Frequently cited frame-level audio features include volume, Zero Crossing Rate (ZCR), pitch, frequency centroid [2], Bandwidth [2], spectral rolloff point [3], and MFCC [4]. To study the semantic content of a video, we need to observe clip-level audio features. An audio clip is defined as a group of neighboring samples that last about one second to several tens seconds. Four types of clip-level audio features have been studied in the literature[1]: volume based, ZCR based, pitch based and frequency-based. Music break, the audio feature introduced in this paper, is built on top of clip-level audio features, but is a higher level audio feature than both frame-level and clip-level audio features.

To use music breaks in semantic video segmentation, we need to differentiate background music from speech. A wide variety of audio features for speech/music discrimination have been studied in the literature and considerable success has been reported [3, 5, 6]. However, most of these studies either focus on segmenting audio into different classes such as speech, music, environment sound and silence or concern with Automatic Speech Recognition (ASR). The effect of music breaks in semantic video segmentation has not been studied in the literature.

The concept of music break comes from the study of documentary video structure. In our study, we found that most documentary videos have background music. Filmmakers intentionally use music to structure the content and communicate the semantic. There is a temporally repetitive pattern of interleaving speech and music in videos. In general, videos start with a music segment that introduces the context and end with a music segment that summarizes the video content, or predicts the future or possibly presents copyright information. In the middle, speech and music segments of variable length alternate. In short,

$$\textit{Documentary} :== \textit{Intro-music}\{\{\textit{speech}\}\{\textit{music}\}\}^{n}\textit{End-music}$$

Based on this observation, we propose to employ music breaks in semantic video segmentation.

The major aim of the work is to investigate how effective music breaks can semantically segment documentary videos. The rest of the paper is organized as follows. In Section 2, we first define video semantic units based on the speech text that a video/audio contains, then describe in detail a three-step procedure for video segmentation using music breaks, i.e., feature calculation, speech/music classification, and video segmentation. Since the music breaks of a documentary video are of different semantic levels, we investigate how different speech/music segment lengths correlate with the semantic level of a music break using Association Rule Mining (ASR) in Section 3. Experiment results are presented in Section 4 and the paper concludes in Section 5.

2. VIDEO SEGMENTATION WITH MUSIC BREAKS

2.1 Definition of semantic units

We view a continuous documentary video stream consisting of a sequence of tractable semantic units. We define semantic units from the standpoint of the speech text that a video/audio contains. A semantic unit is a video/audio segment that matches a topically coherent speech text block. A text block can be either a phrase, or a sentence or a group of sentences. If a semantic unit is parallel in content to its previous semantic unit, then we consider it as a low-level semantic unit, otherwise, a high-level semantic unit.

1-4244-0367-7/06/$20.00 ©2006 IEEE

Formally, let V be a continuous documentary video stream, W be the speech text it contains, R be the set of text blocks on W, T be the video length in time, $v_{(i)}$ be i^{th} semantic unit, $w_{(\mathrm{i})}$ be the text block for $v_{(\mathrm{i}),}$ $st_{v(i)}$ and $et_{v(i)}$ be the start time and end time of i^{th} semantic unit respectively, and $st_{w(i)}$ and $et_{w(i)}$ be the start time and end time of i^{th} text block respectively, then the following defines semantic units of V :

$$V = (v_n) = (v_{(1)}, v_{(2)}, \ldots, v_{(n)}), \text{ (n = \# of semantic units)} \qquad (1)$$

$$w_{(\mathrm{i}),} \in R, \;\; 1 \leq i \leq \mathrm{n} \qquad (2)$$

$$st_{v(i)} \leq st_{w(i)} < et_{w(i)} \leq et_{v(i)}, \;\; 1 \leq i \leq \mathrm{n} \qquad (3)$$

$$st_{v(i)} < et_{\,v(i)} \leq st_{\,v(i+1)} < et_{\,v(i+1)}, \;\; 1 \leq i < i + 1 \leq \mathrm{n} \qquad (4)$$

$$\Sigma\, w_{(\mathrm{i})} = W \text{ and } \Sigma\, [st_{\,v(i)}, et_{\,v(i)}] = T, \;\; 1 \leq i \leq \mathrm{n} \qquad (5)$$

In the definition above, formula (1) indicates that V consists of a sequence of semantic units; formula (2) specifies that every semantic unit matches one topically coherent text block, which is an element of set R; formula (3) stipulates that the time interval of i^{th} text block is equal to or within that of i^{th} semantic unit; formula (4) indicates that semantic units do not overlap, and formula (5) means that the concatenated text blocks is the complete speech text and the sum of time intervals is the complete video length in time.

This definition of semantic units attempts a linear video segmentation. It is generic in that most documentary videos have speeches. In addition, this topic-based segmentation does not have common problems stemming from the lack of context as in other non-topic based segmentation methods.

2.2 Feature calculation

Fixed-length audio clips are used for feature calculation. Grounded on the work [3,5], we select three groups of clip-level audio features: ZCR based, volume based and spectral flux. These features describe the variations of ZCR, short time energy and spectrum of an audio clip.

ZCR based: Zero Crossing Rate is the number of time-domain zero crossings within an audio frame. Three statistical measures of ZCR are used. They are i) standard deviation of the first order difference, ii) the third central moment about the mean, and iii) the difference between the number of audio frames whose ZCRs are above and below the mean value of an audio clip.

Volume based: volume is approximated by the Root Mean Square (RMS) of the signal magnitude within an audio frame. Four statistical measures of RMS are employed. They are i) standard deviation of the first order difference, ii) the third central moment about the mean, iii) the difference between the number of audio frames whose RMSs are above and below the mean value of an audio clip, and iv) low short-time energy ratio, which is the percentage of frames with RMS less than 50% of the mean value.

Spectral Flux: spectral flux indicates frame-to-frame spectral amplitude difference and is represented using $\| \, |X_i| - |X_{i+1}| \, \|$. The sum of the differences of one audio clip is used in this paper.

After feature extraction, the last step in feature calculation is to normalize feature values using Gaussian Normalization [7] across both classes, i.e., speech and music, so that equal emphasis is put on every feature.

2.3 Speech/music classification

Due to its simplicity and comparable accuracy [3], we use KNN to classify an audio stream into speech and music segments.

In our study, we found the probability to observe a single speech segment (≤ 1s) surrounded by music segments is very low, and vice versa. Based on this empirical observation, we perform a simple fine-tuning on classification results. For each spurious segment, i.e. a single speech segment surrounded by music segments or a single music segment surrounded by speech segments, we adjust and predict its class label to be the same as its surrounding segments. After this adjustment, all speech/music segments are at least 2 seconds long.

The last step is to cluster neighboring segments of the same class and form continuous, long speech/music segments. The result of this step is a collection of temporal intervals of interleaving speech and music segments, i.e. ,

$$\{[st_{m(1)}, \; et_{m(1)}], \; [st_{s(1)}, \; et_{s(1)}], \; \ldots, \; [st_{m(n)}, \; et_{m(n)}], \; [st_{m(n+1)}, \; et_{m(n+1)}]\}$$

with *st* represents start time and *et* represents end time, subscript $m(i)$, $s(j)$ ($1 \leq i \leq \mathrm{n} + 1$, $1 \leq j \leq \mathrm{n}$) indicate i^{th} music segment and j^{th} speech segment respectively.

2.4. Video segmentation

In contrast to segmentation for speech recognition, the accuracy of semantic segmentation required here is not very high because this segmentation targets at human information consumption and humans may easily tolerate minor temporal variation in segmentation accuracy. Therefore, for each music segment $[st_{m(i)}, et_{m(i)}]$, we define its corresponding break point $t_{m(i)}$ as the midpoint of that temporal interval (Formula (6)). These music break points then segment continuous video stream into semantic units.

$$t_{m(i)} = st_{m(i)} + (et_{m(i)} - st_{m(i)})/2 \text{ where } 1 \leq i \leq \mathrm{n} \quad (6)$$

3. CORRELATION ANALYSIS

In our study, music breaks segment a video stream into semantic units. We observe that these music breaks are of different semantic levels. To predict the semantic levels of these breaks, we investigate how different segment lengths, represented by their time span, correlate with the semantic levels. Three types of correlations are studied using ARM algorithm, i.e. the correlation between music segment lengths and semantic levels of music breaks, the correlation between speech segment lengths and semantic levels of the music breaks that follows, and the correlation between combined segments lengths (speech segment plus the music segment that follows) and semantic levels of the music breaks.

Two videos, i.e., “How Water Won the West” and “Take Pride in America” from Open-video project (http://www.open-video.org/), are employed for this analysis. Preprocessing involves several steps. We first manually segment each video stream into speech and music segments that last more than one second, and determine the semantic level of each music break. Specifically, if a semantic unit is a high-level semantic unit, we mark the corresponding music break as a higher level semantic break (*h*), otherwise, a lower level semantic break (*l*). Then, we calculate music segment lengths, speech segment lengths, and combined segment lengths. After that, we categorize segments. Specifically, we calculate the average segment length of each type for each video. If a segment length is greater than its average, we mark it as a bigger segment (*bs*), otherwise, mark it as a smaller segment (*ss*). After preprocessing, we perform correlation analysis using Apriori algorithm [8]. Table 1 lists common best rules found in every experiment. From Table 1, we can see that bigger segments strongly correlate with higher level semantic breaks (rule 1), smaller segments correlates well with lower level semantic breaks (rule 2). Among three correlation types investigated, the correlation between the lengths of combined segments and the semantic levels of music breaks is the strongest as indicated with shaded confidence and support values. Rule 3, 4 and 5 justify these statements.

Table 1. Correlation Analysis

CONFIDENCE & SUPPORT	CORRELATION TYPES					
	Music segment lengths vs. semantic levels		Speech segment lengths vs. semantic levels		Combined segment lengths vs. semantic levels	
Best Rules	Conf.	Supp.	Conf.	Supp.	Conf.	Supp.
1. *bs* == > *h*	75%	25%	97%	45%	100%	54%
2. *ss* == > *l*	60%	40%	74%	51%	75%	53%
3. *ss* == > *h*	61%	40%	43%	21%	46%	25%
4. *l* == > *ss*	77%	26%	98%	33%	100%	34%
5. *h* == > *bs*	38%	24%	68%	45%	63%	46%
6. *h* == > *ss*	62%	40%	32%	21%	37%	25%

4. EXPERIMENTS

4.1 Experiments set up

The experiments use eight videos from Open-video Project, which are “Exotic Terrain”, “NASA 25th Anniversary Show”, “Airline Safety and Economy, Report #265”, “Lake Powell”, “Energy Gas”, “How Water Won the West”, “Take Pride in America” and “The Colorado”. Out of the eight videos, the first five are used for training while the remaining three are used for testing. 250 training samples, each of one second long, are prepared for each class, i.e., speech and music. For feature calculation, a sampling rate of 44.1 KHz is used. Fixed-length one-second audio clip is taken as basic unit for feature calculation and speech/music classification, which is further divided into 25ms non-overlapping audio frames.

F-score is adopted for performance evaluation. It is defined as $F = \dfrac{2 \cdot P \cdot R}{P + R}$, where P is precision, defined as the ratio of the number of hits to the total number of detected breaks, and R is recall, defined as the ratio of the number of hits to the number of actual breaks. The higher the F-score is, the better the segmentation accuracy is. To claim a hit, the corresponding music break point as defined in Formula (6) has to be within the temporal interval of the nearest music segment that is determined manually. Formally, let $t_{m(i)}$ be any music break point, $[st_{m(i)}, et_{m(i)}]$ be the nearest music segment determined manually, then,

$$\text{Hits} = \{ t_{m(i)} \mid t_{m(i)} \in [st_{m(i)}, et_{m(i)}], 1 \le i \le \text{n} + 1 \} \quad (7)$$

4.2 Results and Discussions

The aims of these experiments are to evaluate how effective music breaks can segment a continuous video stream into semantic units and how well combined segment lengths can predict different semantic levels of the breaks. Testing data consists of nineteen video samples from three testing videos, most of which are between 1 minute and 3 minutes long. Three steps as described in Section 2 are performed. The effectiveness of the approach is shown in Table 2 and Table 3. Figure 1 illustrates one segmentation example.

In Table 2 and 3, “MBs” is a shorthand for music breaks and “SUs” is for semantic units. Table 2 tells how effective the approach can pick out the music breaks while Table 3 indicates how well these music breaks can segment video streams into semantic units. The average F- score for music breaks detection is about 0.94 while that for semantic unit predication is 0.91. While these two scores are very close, the former is slightly higher than the later. These experimental results indicate music breaks are effective in semantic video segmentation, but there are a small portion of semantic units that can not be detected by music breaks. That is because filmmakers do use other approaches for topic changes as well. In addition, we note that the F-scores of the video titled “The Colorado” (No. 7 – 13) are lower than those of the other two (No. 1- 6 and No. 14 – 19). By comparing the background music, we found that the background music of “The Colorado” has lots of abrupt and striking changes, which causes the algorithm to mistaken music for speech. This reasoning is justified by its low precision and high recall.

To test how well combined segment lengths can predict semantic levels of music breaks, we pick four audio samples from the testing set. For each sample, we calculate combined segment lengths and categorize these segments as *bs* or *ss* based on the average segment length of each audio sample. We then predict higher semantic breaks *h* using *bs* and lower semantic breaks *l* using *ss*. The results (Table 4) show that combined segment lengths are reliable in predicting semantic levels of music breaks with an average F-score of 0.89 in both cases.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated semantic segmentation of documentary video using music breaks. We defined video semantic units from the standpoint of the speech text that a video contains, proposed three-step procedure for video segmentation using music breaks, and studied correlations between different segment lengths and semantic levels of music breaks. We found that music breaks can effectively segment continuous video streams into semantic units with an average F-score of 0.91 and combined segment lengths, i.e., speech segment plus music segment that follows, strongly correlates with semantic levels of music breaks. Bigger segments correlate well with higher level semantic breaks, smaller segments correlates with lower level semantic breaks.

Using music breaks for documentary video segmentation is a generic approach in that most documentaries use background music to help communicate. In the future, we will investigate effective mechanisms to integrate this approach into multi-mode video segmentation. In addition, more salient features need to be studied for accurate speech/music classification with short time interval (≤ 1s).

6. RERERENCES

[1].Y. Wang, Z. Liu and J. Huang, "Multimedia Content Analysis," *IEEE Signal Processing Magazine,* 17(6), pp. 12-36, Nov. 2000.

[2].Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *J. VLSI Signal Processing Syst. Signal, Image, Video Technol.*, vol. 20, pp. 61-79, Oct. 1998.

[3].E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeatures speech/music discriminator," *Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-97)*, vol. 2, pp. 1331-1334, Munich, Germany, Apr. 1997.

[4].L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, *Englewood Cliffs*, NJ: Prentice Hall, 1993.

[5]. J. Saunders, "Real-time discrimination of broadcast speech/music," *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-96)*, vol. 2, Atlanta, GA, May, 1996, pp. 993-996.

[6].Lie Lu, Hao Jiang, Hong-Jiang Zhang, "A Robust Audio Classification and Segmentation Method. *Proc. ACM Multimedia 01*, Ottawa, Canada, pp203-211, 2001.

[7].Q. Iqbal and J. Aggarwal, "Combining Structure, Color and Texture for Image Retrieval: A Performance Evaluation", *Proc.of the 16th Int. Conf. on Pattern Recognition*, Quebec City, Canada, vol. 2, 2002, pp. 438-443.

[8]. R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules," *Proc. of 20th International Conference of Very Large Data Bases (VLDB)*, pp. 487—499.

Table 2. Music breaks detection

No.	# of MBs (≥2s)	# of Detected MBs(≥2s)	# of Hits	P	R	F
1	7	7	7	1.00	1.00	1.00
2	4	4	4	1.00	1.00	1.00
3	4	3	3	1.00	0.75	0.86
4	4	4	4	1.00	1.00	1.00
5	2	2	2	1.00	1.00	1.00
6	3	3	3	1.00	1.00	1.00
7	3	3	3	1.00	1.00	1.00
8	6	7	6	0.86	1.00	0.92
9	3	3	3	1.00	1.00	1.00
10	3	4	3	0.75	1.00	0.86
11	4	9	4	0.44	1.00	0.62
12	4	5	4	0.80	1.00	0.89
13	3	5	3	0.60	1.00	0.75
14	2	2	2	1.00	1.00	1.00
15	7	7	7	1.00	1.00	1.00
16	8	6	6	1.00	0.75	0.86
17	7	6	6	1.00	0.86	0.92
18	14	13	13	1.00	0.93	0.96
19	7	7	7	1.00	1.00	1.00
Ave.						0.93

Table 3. Semantic breaks detection

No.	# 0f SUs	# of Detected SUs	# of Hits	P	R	F
1	6	6	6	1.00	1.00	1.00
2	4	3	3	1.00	0.75	0.86
3	5	4	4	1.00	0.80	0.89
4	3	3	3	1.00	1.00	1.00
5	2	2	2	1.00	1.00	1.00
5	2	2	2	1.00	1.00	1.00
7	3	3	3	1.00	1.00	1.00
8	5	6	5	0.83	1.00	0.91
9	2	2	2	1.00	1.00	1.00
10	2	3	2	0.67	1.00	0.80
11	3	7	3	0.43	1.00	0.60
12	3	4	3	0.75	1.00	0.86
13	3	5	3	0.60	1.00	0.75
14	1	1	1	1.00	1.00	1.00
15	8	7	7	1.00	0.88	0.94
16	9	7	7	1.00	0.77	0.87
17	5	5	5	1.00	1.00	1.00
18	14	13	13	1.00	0.93	0.96
19	6	6	5	0.83	0.83	0.83
Ave.						0.91

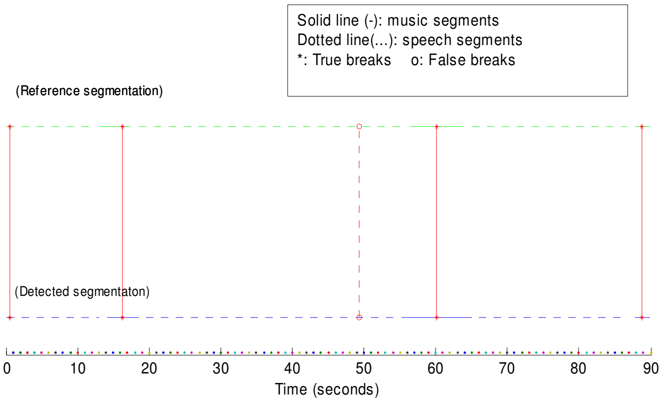

Figure 1. Segmentation illustration

Table 4. Semantic levels predication

No.	Predicting *h* with *bs*			Predicting *l* with *ss*		
	P	R	F	P	R	F
1	1.00	0.75	0.86	0.67	1.00	0.80
15	1.00	1.00	1.00	1.00	1.00	1.00
16	0.67	0.67	0.67	0.75	0.75	0.75
18	1.00	1.00	1.00	1.00	1.00	1.00
Ave.			0.89	Ave.		0.89