HMM-Based Audio Keyword Generation*

Min Xu[1], Ling-Yu Duan[2], Jianfei Cai[1], Liang-Tien Chia[1],
Changsheng Xu[2], and Qi Tian[2]

[1] School of Computer Engineering,
Nanyang Technological University, Singapore, 639798
{mxu,asjfcai,asltchia}@ntu.edu.sg
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613
{lingyu,xucs,tian}@i2r.a-star.edu.sg

Abstract. With the exponential growth in the production creation of multimedia data, there is an increasing need for video semantic analysis. Audio, as a significant part of video, provides important cues to human perception when humans are browsing and understanding video contents. To detect semantic content by useful audio information, we introduce audio keywords which are sets of specific audio sounds related to semantic events. In our previous work, we designed a hierarchical Support Vector Machine (SVM) classifier for audio keyword identification. However, a weakness of our previous work is that audio signals are artificially segmented into 20 ms frames for frame-based SVM identification without any contextual information. In this paper, we propose a classification method based on Hidden Markov Modal (HMM) for audio keyword identification as an improved work instead of using hierarchical SVM classifier. Choosing HMM is motivated by the successful story of HMM in speech recognition. Unlike the frame-based SVM classification followed by major voting, our proposed HMM-based classifiers treat specific sound as a continuous time series data and employ hidden states transition to capture context information. In particular, we study how to find an effective HMM, i.e., determining topology, observation vectors and statistical parameters of HMM. We also compare different HMM structures with different hidden states, and adjust time series data with variable length. Experimental data includes 40 minutes basketball audio which comes from real-time sports games. Experimental results show that, for audio keyword generation, the proposed HMM-based method outperforms the previous hierarchical SVM.

1 Introduction

With the increasing multimedia data available from Internet, there is a need to develop intelligent multimedia indexing and browsing systems. To facilitate high-level abstraction and efficient content-based access, semantics extraction is becoming an important aspect of multimedia-understanding. Recently, video

* This research is partially supported by Singapore A*STAR SERC Grant (032 101 0006).

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3333, pp. 566–574, 2004.

semantic analysis attracts more and more research efforts [1,2,3]. Their works attempt to extract semantic meaning from visual information but little work has been done on the audio parts of multimedia streams.

Audio, which includes voice, music, and various kinds of environmental sounds, is an important type of media, and also a significant part of video. Recently people have begun to realize the importance of effective audio content analysis which provides important cues for semantics. Most of the existing works try to employ audio-visual compensation to solve some problems which can not be successfully solved only by visual analysis [4,5,6,7]. Nepal et al. [5] employed heuristic rules to combine crowd cheer, score display, and change in motion direction for detecting "Goal" segments in basketball videos. Han et al. [6]] used a maximum entropy method to integrate image, audio, and speech cues to detect and classify highlights from baseball video. An event detection scheme based on the integration of visual and auditory modalities was proposed in [4, 7]. To improve the reliability and efficiency in video content analysis, visual and auditory integration methods have been widely researched.

Audio content analysis is the necessary step for visual and auditory integration. Effective audio analysis techniques can provide convincing results. In consideration of computational efficiency, some research efforts have been done for pure audio content analysis [8,9]. Rui et al. [8] presented baseball highlight extraction methods based on excited audio segments detection. Game-specific audio sounds, such as whistling, excited audience sounds and commentator speech, were used to detect soccer events in [9].

In [4,7,9], we used hierarchical Support Vector Machine (SVM) to identify audio keywords. The audio signals were segmented into 20 ms frames for frame-based identification while the audio signals are time continuous series signals rich in context information. By using SVM, we did not take into account the contextual information which is significant for time series classification. HMM is a statistical model of sequential data that has been successfully used in many applications including artificial intelligence, pattern recognition, speech recognition, and modeling of biological sequences [10]. Recently, HMM were introduced to sports video analysis domain [11,12,13,14]. Assfalg et al. [12] used HMM to model different events, where states were used to represent different camera motion patterns. In [14], Xie et al. tried to model the stochastic structures of play and break in soccer game with a set of HMMs in a hierarchical way. Dynamic programming techniques were used to obtain the maximum likelihood play/break segmentation of the soccer video sequence at the symbol-level. These works demonstrated that HMM is an effective and efficient tool to represent time continuous signals and discover structures in video content.

In this paper, we present our recent research work of audio keywords detection by using Hidden Markov Models (HMM) as an improved work for [4,7,9]. In Section 2, we briefly introduce audio keywords and HMM-based generation scheme. Section 3 discusses the audio feature extraction work. Our proposed HMM structure is presented in Section 4. Some comparison experiments and results are listed in Section 5. In Section 6, we draw conclusions and discuss some future work.

Table 1. Audio keywords' relationship to potential events.

Sports	Audio Keywords	Potential Events
Tennis	Applause	Score
	Commentator Speech	At the end (or the beginning) of a point
	Silence	Within a point
	Hitting Ball	Serve, Ace or Return
Soccer	Long-whistling	Start of free kick, penalty kick, or corner kick, Game start or end, offside
	Double-whistling	Foul
	Multi-whistling	Referee reminding
	Excited commentator speech or excited audience sound	Goal or Shot
	Plain commentator speech or plain audience sound	Normal
Basketball	Whistling	Fault
	Ball hitting backboard or basket	Shot
	Excited commentator speech or excited audience sounds	Fast break, Drive or Score
	Plain commentator speech or plain audience sound	Normal

2 Brief Introduction of Audio Keyword Generation System

Audio keywords are defined as some specific audio sounds which have strong hints to interesting events. Especially in sports video, some game-specific audio sounds (e.g. whistling, excited commentator speech, etc.) have strong relationships to the actions of players, referees, commentators and audience. These audio sounds may take place in the presence of interesting events as listed in Table 1. Generally, excited commentator speech and excited audience sounds play important roles in highlight detection of sports video. Other keywords may be specific to a kind of sports game.

Audio signal exhibits the consecutive changes in values over a period of time, where variables may be predicted from earlier values. That is, strong context exists. In consideration of the success of HMM in speech recognition, we propose our HMM based audio keywords generation system. The proposed system includes three stages, which are feature extraction, data preparation and HMM learning, as shown in Fig. 1.

As illustrated in Fig. 1, selected low-level features are firstly extracted from audio streams and tokens are added to create observation vectors. These data are then separated into two sets for training and testing. After that, HMM is trained then reestimated by using dynamic programming. Finally, according to maximum posterior probability, the audio keyword with the largest probability is selected to label the corresponding testing data. We next introduce the proposed system in detail.

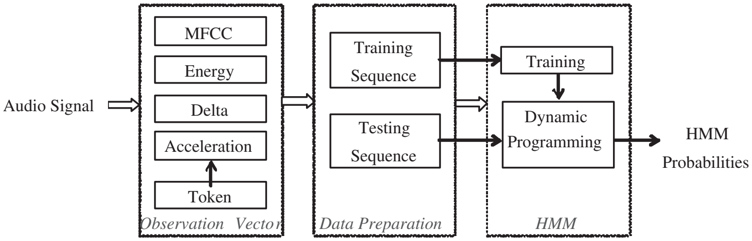

Fig. 1. Proposed audio keywords generation system.

3 Feature Extraction

We segment audio signal at 20 ms per frame which is the basic unit for feature extraction. Mel-Frequency Cepstral Coefficient (MFCC) and Energy are selected as the low-level audio features as they are successfully used in speech recognition and further proved to be efficient for audio keyword generation in [9]. Delta and Acceleration are further used to accentuate signal temporal characters for HMM [15].

3.1 Mel-Frequency Cepstral Coefficient

The mel-frequency cepstrum is highly effective in audio recognition and in modeling the subjective pitch and frequency content of audio signals. Mel scale is calculated as

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}), \tag{1}$$

where $Mel(f)$ is the logarithmic scale of the normal frequency scale f. Mel scale has a constant mel-frequency interval, and covers the frequency range of 0 Hz - 20050 Hz. The Mel-Frequency Cepstral Coefficients (MFCCs) are computed from the FFT power coefficients which are filtered by a triangular band pass filter bank. The filter bank consists of 12 triangular filters. The MFCCs are calculated as

$$C_n = \sqrt{\frac{2}{k}} \sum_{k=1}^{K} (\log S_k) \cos[n(k - 0.5)\pi/k], \quad n = 1, 2, \cdots, N \tag{2}$$

where $S_k (k = 1, 2, \cdots K)$ is the output of the filter banks and N is total number of samples in a 20 ms audio unit.

3.2 Energy

The energy measures amplitude variations of the speech signal. The energy is computed as the log of the signal energy, that is, for audio samples $\{s_n, n = 1, \cdots, N\}$

$$E = \log \sum_{n=1}^{N} s_n^2 \tag{3}$$

3.3 Delta and Acceleration

Delta and acceleration effectively increase the state definition by including first and second order memory of past states. The delta and acceleration coefficients are computed using the following simple formula (C_t is the coefficients from feature vector at time t).

$$\Delta(C_t) = C_t - C_{t-1}; Acc(C_t) = \Delta(C_t) - \Delta(C_{t-1}) \tag{4}$$

4 Our Proposed Hidden Markov Model

As for the HMM generation, we need to determine the HMM topology and statistical parameters. In this research, we choose the typical left-right HMM structure, as shown in Figure 2, where $S = \{s_1, \cdots, s_5\}$ are five states; $A = \{a_{ij}\}$ are the state transition probabilities and $B = \{b_i(v_k)\}$ are the observation probability density functions which is represented by a mixture Gaussian density. In our case, each audio frame f_i is regards as one observation o_i. One HMM sample $A = \{f_1, f_2, \cdots, f_n\}$, including n frames, is regards as an observed sequence, $O = \{o_1, o_2, \cdots, o_n\}$. The resulting audio features from each frame form the observation vectors. We use $\lambda = (\Pi, A, B)$ to denote all the parameters, where $\Pi = \{\pi_i\}$ are the initial state probabilities. In training stage, observation vectors are separated into classes to estimate initial B firstly. Then, to maximize the probability of generating an observed sequence, i.e. to find $\lambda^* = arg\ max_\lambda\ p(O|\lambda)$, we use Baum-Welch algorithm to adjust the parameters of model λ.

The recognition stage is shown in Figure 3, where l audio keywords are associated with pre-trained HMMs. For each coming audio sample sequence, the likelihood of every HMM is computed. The audio sequence A is recognized as keyword k, if $P(O|\lambda_k) = max_l\ P(O|\lambda_l)$ [15].

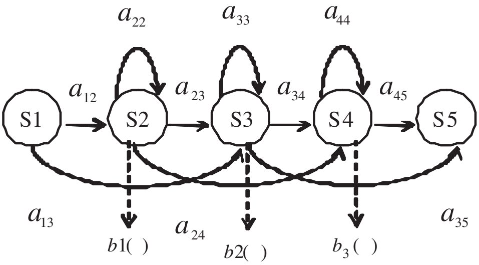

Fig. 2. The Left-right HMM with 5 States.

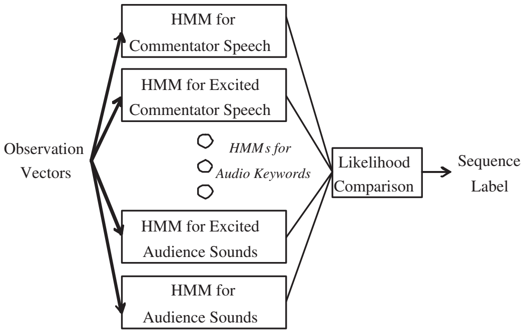

Fig. 3. The HMM overview structure.

In the following experiment, we are concerned about two issues. One is how many states are suitable for a HMM. The other one is the HMM sample length selection.

5 Experiments and Results

Excited commentator speech and excited audience sounds directly correspond to sports highlight which attracts audience's interests mostly. Compared with whistling and hitting ball, the recognition of these two keywords is quite challenging as excited parts always interlace with plain parts. Therefore, in our experiments, we concentrate on excited commentator speech and excited audience sounds.

The audio samples come from a 40 minutes real-time basketball game. They are collected with 44.1 kHz sample rate, stereo channels and 16 bits per sample. We used two third for training and one third for testing.

For the HMM learning, different number of states may model different states transition process, which could influence results. Moreover, as each kind of audio keywords have their own durations, we need to choose appropriate sample length for different keyword training. Therefore, we conduct some experiments to compare HMM structures with various states and change HMM sample length to achieve the best performance of our proposed audio keyword generation system.

5.1 HMM with Different Hidden States

At present, we change the states from 3 to 5. The experimental results are listed in Table 2. We find that 3-state HMM is good while 4-state HMM provides better performance for excited commentator. In some sports games, when the environment is very noisy, we can not detect sports highlights only by excited audience sounds while excited commentator speech is able to provide the most important cues. Therefore, higher performance of excited commentator speech identification is necessary. Based on the above criteria and performance results, we thus use 4-states HMM to generate audio keywords.

Table 2. Performance of various HMMs with different states for audio keyword generation.

Audio Keywords	States Number	Recall(%)	Precision (%)
Audience	5 States	95.74	95.74
	4 States	95.74	95.74
	3 States	100	100
Commentator	5 States	100	91.07
	4 States	98.04	94.34
	3 States	100	92.73
Excited Audience	5 States	85.71	85.71
	4 States	85.71	85.71
	3 States	100	100
Excited Commentator	5 States	66.67	100
	4 States	86.67	100
	3 States	73.33	100

Table 3. Performance of different sample length for audio keyword generation (5 states HMM).

Audio Keywords	Sample Length	Recall(%)	Precision (%)
Audience	0.2 Sec	95.39	96.61
	1 Sec	95.74	95.74
Commentator	0.2 Sec	96.52	83.33
	1 Sec	100	91.07
Excited Audience	0.2 Sec	83.33	75.95
	1 Sec	85.71	85.71
Excited Commentator	0.2 Sec	31.65	73.53
	1 Sec	66.67	100

Table 4. Audio keyword generation results (HMM vs. SVM).

Audio Keywords	Methods	Recall(%)	Precision (%)
Whistling	SVM	99.45	99.45
	HMM	100	100
Audience	SVM	83.71	79.52
	HMM	95.74	95.74
Commentator	SVM	79.09	78.27
	HMM	98.04	94.34
Excited Audience	SVM	80.14	81.17
	HMM	85.71	85.71
Excited Commentator	SVM	78.44	82.57
	HMM	86.67	100

5.2 HMM with Different Sample Length

Observation of real sports games reveals that the shortest keyword whistling lasts slightly longer than 0.2 second. Therefore, we segment audio signals into 0.2 second as samples for whistling detection. However, other audio keywords, such as commentator speech, excited audience sounds and etc., last much longer than 0.2 second. Table 3 lists the results of different sample length for several types of audio keywords. The Experimental results show that 1 second sample length is much better than 0.2 second for audience sounds and commentator speech related audio keyword generation. The main reason is that longer sample length provides much more contextual information for HMM to learn in order to differentiate among different audio keywords.

5.3 Comparison Between HMM and SVM

We further do a comparison between the HMM-based method and the SVM-based method [7]. According to the previous experimental results, 4-state left-right structure is selected to build HMM. We choose 0.2 second as sample length for whistling generation and 1 second for other audio keywords (i.e., commentator speech, audience sounds etc.). Compared with SVM-based audio keyword generation, the proposed HMM-based method achieves better performance as listed in Table 4. For the excited keywords generation, which are more significant for highlight detection, the recalls and precisions are improved at least 5%.

6 Conclusion Remarks

Our proposed HMM-based method for audio keyword generation outperforms the previous SVM based method, especially for the excited commentator speech and excited audience sounds. This conforms to the fact that the HMM-based method effectively captures rich contextual information so as to improve different keywords' separability.

As plain/excited commentator speech and plain/excited audience sound are quite general for extensive sports games, we are trying to design adaptive HMMs and combine visual features to boost performance among different kinds of sports games.

References

1. Gong, Y.H., Sin, L.T., Chuan, C.H., Zhang, H.J., Sakauchi, M.: Automatic parsing of TV soccer programs. In: International Conference on Multimedia Computing and System. (1995) 167–174
2. Tan, Y.P., Saur, D.D., Kulkarni, S.R., Ramadge, P.J.: Rapid estimation of camera motion from compressed video with application to video annotation. IEEE Trans. on Circuits and Systems for Video Technology **10** (2000) 133–146

3. Xu, P., Xie, L., Chang, S.F., Divakaran, A., Vetro, A., Sun, H.: Algorithms and systems for segmentation and structure analysis in soccer video. In: IEEE International Conference on Multimedia and Expo. (2001) 22–25
4. Duan, L.Y., Xu, M., Chua, T.S., Tian, Q., Xu, C.S.: A mid-level representation framework for semantic sports video analysis. In: ACM Multimedia. (2003)
5. Nepal, S., Srinivasan, U., Reynolds, G.: Automatic detection of goal segments in basketball videos. In: ACM Multimedia. (2001)
6. Han, M., Hua, W., Xu, W., Gong, Y.H.: An integrated baseball digest system using maximum entropy method. In: ACM Multimedia. (2002) 347–350
7. Xu, M., Duan, L.Y., Xu, C.S., Kankanhalli, M., Tian, Q.: Event detection in basketball video using multiple modalities. In: IEEE Pacific Rim Conference on Multimedia 2003. (2003)
8. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV baseball programs. In: ACM Multimedia. (2000) 105–115
9. Xu, M., Maddage, N.C., Xu, C., Kankanhalli, M., Tian, Q.: Creating audio keywords for event detection in soccer video. In: IEEE International Conference on Multimedia and Expo. (2003) 6–9
10. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
11. Pan, H., Beek, P., Sezan, M.I.: Detection of slow-motion replay segments in sports video for highlights generation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. (2001) 1649–1652
12. Assfalg, J., Bertini, M., Bimbo, A.D., Nunziati, W., Pala, P.: Soccer highlights detection and recognition using HMMs. In: IEEE International Conference on Multi-media and Expo. (2002) 825 – 828
13. Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.S.: Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. (2003) V–632 – V–635
14. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with domain knowledge and hidden markov models. Pattern Recognition Letters **25** (2004) 767–775
15. Young, S., et al: The HTK Book (for HTK Version 3.1) http://htk.eng.cam.edu/. Cambridge University Engineering Department (2002)