A HISTOGRAM ALGORITHM FOR FAST AUDIO RETRIEVAL

Wei Liang
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China
wliang@hitic.ia.ac.cn

Shuwu Zhang
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China
swzhang@hitic.ia.ac.cn

Bo Xu
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, China
xubo@hitic.ia.ac.cn

ABSTRACT

This paper describes a fast audio detection method for specific audio retrieval in the AV stream. The method is a histogram matching algorithm based on structural and perceptual features. This algorithm extracts audio features based on human perception on the sound scene and locates the special audio clip by fast histogram matching. Experimental results based on the advertisement detection in TV program showed that the algorithm can achieve a very high overall precision and recall rate both about 97% with very fast search time about 1/40 on real time.

Keywords: audio Retrieval, histogram.

1 INTRODUCTION

TV broadcasting is a rich multimedia information source. There are large amounts of advertisements (Ad) in TV programs of different channels per day. Traditionally, the work of Ad monitoring is done mainly by human checking. However, it is really a stuffy work for human to listen and watch so many TV plays chronically. Thus, how to precisely detect and locate these Ad segments from long Audio and Video (AV) recordings becomes a real requirement for advertisers.

This paper address the problem of detecting and locating sound objects from a stream of TV audio data quickly, while using computationally inexpensive processing. This has potential applications for multimedia data management.

In recent years, there is an increasing interest on AV data retrieval. Relevant techniques on audio feature extraction, modeling and search algorithm have been widely studied. For instance, Wold et al. employed loudness, brightness, pitch, timbre as audio perceptual features [1]; Liu et al. used subband energy ratio feature [2]; Foote also proposed 12 dimensions MFCC and Energy features [4]; some audio modeling approaches are GMM, HMM, VSM, Histogram, and so on.

Aimed at TV Ad monitoring, this paper describes a fast histogram based audio retrieval algorithm. This algorithm extracts audio features based on human's sense of sound and locates the specific audio clip by fast histogram matching. Experimental results based on TV Ad detection showed that the algorithm can achieve a very high overall precision above 96% with very fast search time.

The paper is organized as follows: Section 2 explains the details of the algorithm including audio feature extraction and modeling. Section 3 introduces the experiment based on TV Ad monitoring. And section 4 gives the conclusion.

2 AUDIO RETRIEVAL ALGORITHM

2.1 Overview

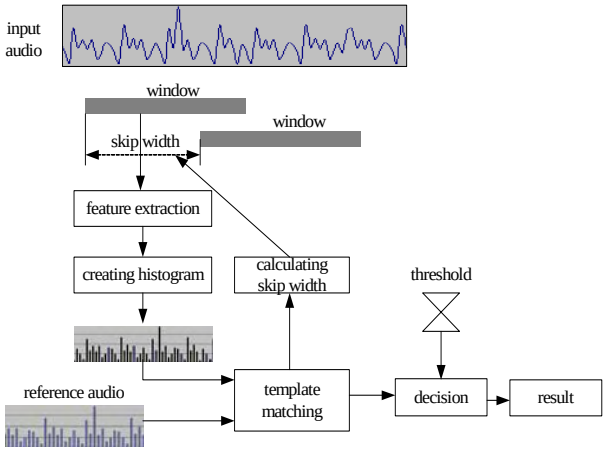

Figure1: Block diagram of the Audio Retrieval Algorithm

Figure 1 outlines the block diagram of proposed audio search algorithm. Firstly, the feature vectors are extracted from both reference audio and test audio. The window on the input signal is shifted with a range of overlapping. The Window length may be the same as the reference audio duration. Then, the histogram is created from the feature vectors. In the next step, the template matching is carried out to detect and locate the reference audio segment from input audio.

2.2 Feature Extraction

As mentioned in above section, there are some different types of audio features already used in audio retrieval. For example: Wold et al. used loudness, brightness, pitch, timbre [1]; Liu et al. used subband energy ratio [2]; Foote used 12 dimensions MFCC and Energy [4]. These features have reported a good precision in quiet envi-

ronment. However, it could be relatively distorted in strong noise environments. It, thus, is necessary to design a group of features, which could be beneficial for both computing cost and noise resistance, for robust modeling under noise environment.

In general, multimedia signal is a mixed signal of many different sources. Human ear has the nature to distinguish individual sound from mixed sources. This function of human ear is called stream segregation. Stream segregation involves two steps, which are decomposing the signal into its continent parts (partials) and grouping these parts into streams - one stream for each sound source. At a basic level, one can model audio representation in the human mind as a series of peak amplitude tracks in a time-frequency-intensity space [4].

Considering the unstable factor in TV signals, we adopt multiple bands energy relative ratio as basic audio feature. This type of feature may depict energy movement trend in a time-frequency-intensity space. Its mathematical description can be described as follows.

An audio feature vector $feature(n)$ is written as

$$feature(n) = (f(n), g(n)) \quad (1)$$

$$f(n) = (f_1(n), f_2(n), f_3(n), ..., f_M(n)) \quad (2)$$

$$g(n) = (g_1(n), g_2(n), g_3(n), ..., g_M(n)) \quad (3)$$

where n is the time frame. M denotes the number of frequency sub-bands. An element of $f(n)$ is the normalized short-time power spectrum, which is given as

$$f_i(n) = \alpha(n) \times E_i(n) \quad (4)$$

An element of $g(n)$ is the normalized short-time power spectrum change ratio by time, which is given as

$$g_i(n) = \beta(n) \times ECR_i(n) \quad (5)$$

$$ECR_i(n) = (E_i(n) - E_i(n-1)) / E_i(n-1) \quad (6)$$

where $E_i(n)$ denotes the energy of output of the I-th sub-band filter at n-th frame. Because short -time energy is sensitive to high level voltage, this algorithm uses short-time average amplitude to carve the change of signal amplitude, which is given as

$$E_i(n) = \sum_{t=n*L}^{(n+1)*L-1} |x_i(t)| \quad (7)$$

where, L is the length of a frame, and $x_i(t)$ denotes the amplitude of ith sub-band at sampling point t. And $\alpha(n)$ and $\beta(n)$ is a normalized constant defined as

$$\alpha(n) = \frac{1}{\max_i(E_i(n))} \quad (8)$$

$$\beta(n) = \frac{1}{\max_i(ECR_i(n))} \quad (9)$$

Figure 2 shows the short-time energy curve of an audio clip that is processed through multiple band pass filters.

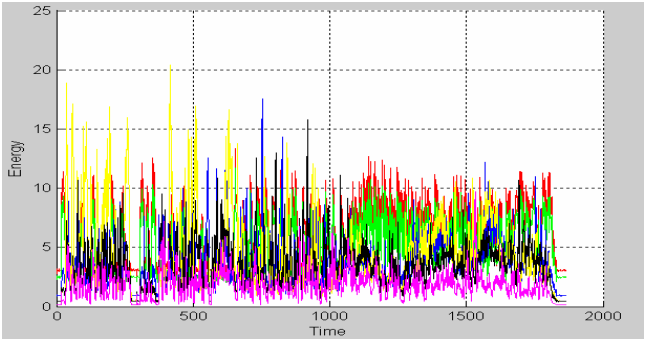

Figure 2: the short-time energy curve of an audio clip after passing multiple band pass filters

In order to reduce the bad influence of noise and volume and lift the perceptual features of the audio clip, the short-time energy curve need to be filtered by low pass filters again. Figure 3 shows the result filtered by low pass filters.

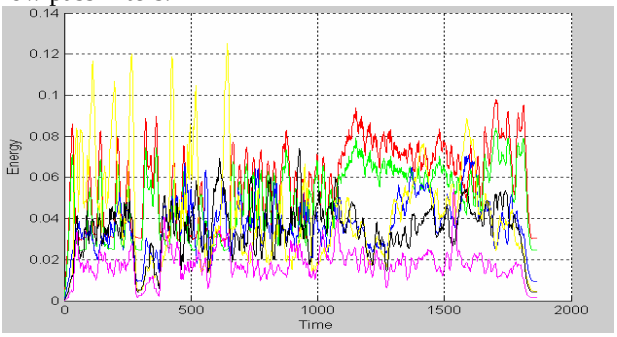

Figure 3: the curve of short-time energy curve through low pass filters.

Furthermore, normalization across frequency bands is taken to delete the influence of absolute energy. Figure 4 shows the result of normalization across frequency bands.

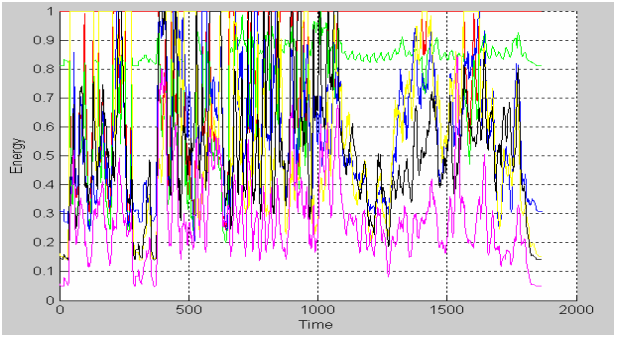

Figure 4: the result of normalization across frequency bands

2.3 Histogram Modeling

After feature extraction, we need to train model for each audio clip. Some modeling approaches, such as GMM, HMM, SVM, have already been employed for audio modeling. However, because of computationally expensive processing, it is hard to meet the speed demand of quick audio search and retrieval. Histograms can be used as a type of non-parametric signal model for both the reference and input signals over a shifted window. It

doesn't need computationally expensive processing while it is relatively stable under adverse environments [5]. We, thus, adopt histogram modeling for specific audio detection. .

For the sake of removing the influence of noises, the feature vector, firstly, need to be quantized (VQ) before modeling histogram. We use the codebook of VQ to build histogram. The similarity distance between the reference and input feature vector histogram can be measured by histogram intersection. The histogram intersection for a window is defined as

$$S_n(h^R, h^T(n)) = \frac{1}{L}\sum_{j=1}^{L} \min(h_j^R, h_j^T(n)) \qquad (10)$$

where h^R is the histogram for the reference; $h_j^T(n)$ is the histogram started from the i-th frame; and L denotes the number of bins.

2.4 The Prediction of Similarity Upper Bound and Skip Width

As the window for input signal shifts forward in time, the similarity based on reference and input feature vector histograms changes with regard to the correlated overlapping between reference and object segment in input stream. We, thus, may predict the next upper bound of the similarity in terms of current value. The upper bound on $S(h_i^R, h_i^T)$ can be defined as:

$$S_{ub}(h_i^R, h_i^T(n2)) = S(h_i^R, h_i^T(n1)) + \frac{n2 - n1}{p_i} \qquad (11)$$

Where $h_i^T(n1)$ and $h_i^T(n2)$ are the histograms for windows started from n1 and n2 frame respectively, $n1 < n2$; p_i denotes the total number of frames in each histogram. When the window is shifted the n2-th frame, the similarity should be no larger than $S_{ub}(h_i^R, h_i^T(n2))$. We, thus, may set the threshold to skip the durations where the similarity is lower than the threshold. Using Eq.(9), the skip width can be calculated as:

$$w = \begin{cases} floor(p_i(\theta - S)) + 1 & if\ (S < \theta) \\ 1 & otherwise \end{cases} \qquad (12)$$

Where $floor(x)$ means the greatest integral value less than x; θ is the threshold; and S denotes the value of current similarity.

3 EXPERIMENTS

The experiments on TV Ad retrieval have been conducted based on the recordings of real TV broadcast. We randomly picked out 200 different commercial Ad templates of different durations from real TV broadcasting and edited a 20 hours' test set of TV program from six channels. In the audio feature extraction, the audio of recording was first digitized at 8 kHz sampling frequency and 16 bit quantization accuracy.

The experimental platform we used was a workstation Pentium4 2.4G CPU, 256M memory. The performance was evaluated with regard to search speech and accuracy under the recording of real TV broadcast.

3.1 Search Speed

The time cost for the search consists of two parts: (1) the time cost during feature extraction, and (2) the search time based on the extracted feature vectors.

(1) The feature extraction in the experiment was performed on 20 hours' testing TV program and 200 commercial ads. It took about 240 seconds of CPU time totally;

(2) The search time depends on many factors, such as the length of reference clip, the length of shifted window of input signal, numbers of histogram bins, threshold, and so on. Averagely, it takes about 10-15 seconds for each reference Ad clip to search the number of occurrences through whole testing TV program.

3.2 Search Accuracy

The search accuracy was evaluated by the recall rate δ , precision rate ξ , and average accuracy η . These are defined as

$$\delta = \frac{the\ number\ of\ correctly\ retrieved\ objects}{the\ number\ of\ objects\ that\ should\ be\ retrieved} \qquad (13)$$

$$\xi = \frac{the\ number\ of\ correctly\ retrieved\ objects}{the\ number\ of\ all\ retrieved\ objects} \qquad (14)$$

$$\eta = \frac{\delta + \xi}{2} \qquad (15)$$

Table 1 lists the performance of histogram search in comparison with correlation coefficient matching. It has shown that the histogram search algorithm can achieve a very high precision about 97% on average with very fast search time about 1/40 times of real time within 200 Ad audio clips.

Table 1: The performance of histogram search compared to correlative coefficient matching.

Algorithm	Len. of Ads (sec)	Precision (%)	Recall (%)	Accuracy (%)	CPU time
Correl. Coeff.	<=5	64.8%	78.2%	71.5%	21hr56m
	6-10	91.1%	88.5%	89.8%	
	11-20	97.1%	85.8%	91.4%	
	>20	100%	89.7%	94.8%	
	On Ave.	88.3%	85.6%	86.89%	
Histogram	<=5	93.0%	94.2%	93.6%	30m14s.
	6-10	95.3%	96.0%	95.7%	
	11-20	99.2%	97.5%	98.4%	
	>20.	100%	98.2%	99.1%	
	On Ave.	96.9%	96.5%	96.7%	
* search time is the cost of detecting 200 Ads through 20 hrs TV program					

4 CONCLUSION

This paper has discussed a fast histogram search algorithm that can quickly detect and locate a reference audio segment in a long audio stream. The experiment based on TV Ad detection showed that the performance on both precision and speed are significantly acceptable for real applications. This algorithm has been truly applied in the Ad monitoring system by the Bureau of Beijing Business Administration.

A potential improvement point of the algorithm is that since the disturbance of AV signals in transmission line as well as the volume change of acceptor, the audio signal would be distorted in some sense. We are planning to study a kind of spectrum masking technique of robust feature selection for quick audio retrieval. We also will further revise the algorithm enable to search a segment in a reference clip by dynamically changing the length of search window.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (NSFC) under the grant No. 60475014 and National Hi-tech Research Plan under the grant No. 2003AA115520 & 2005AA114130.

REFERENCES

[1] Wold, E, Blum, T, Keislar, D, and Wheaton, J (1996), “Content-based classification, search and retrieval of audio”, IEEE Multimedia Mag.,Vol. 3, pp.27-36, July 1996.

[2] Liu. Z., Huang. J., Wang Y., and Chen T. (1997), “Audio feature extraction and analysis for scene classification,” in IEEE Signal Processing,1997.

[3] Foote J. et al. (1997), “Content-based retrieval of music and audio,”in Proc.SPIE Multimedia Storage Archiving Systems II, vol.3229,C.C.J.Kuo et al., Eds.,1997,pp.138-147.

[4] Ellis D. P. W. , and Vercoe B. L. (1992), “A Perceptual Representation of Audio for Auditory signal Separation”, presented at the 23th meeting of the Acoustical Society of America, Salt Lake City, May 1992.

[5] Smith G., Murase H., and Kashino K. (1998): “Quick Audio Retrieval Using Active Search”, Proc. of ICASSP-98, Vol.6(1998).

[6] Han K.-P., Park Y.-S., Jeon S.-G., Lee G.-C., and Ha Y.-H (1998). “Genre classification system of TV sound signals based on a spectrogram analysis”. IEEE Transactions on Consumer Electronics ,44(1):33-42, February 1998.

[7] Jiang D.-N., Lu L., Zhang H.-J., Tao J.-H., and Cai L.-H. (2002), “Music Type classification by spectral contrast feature”. Proceedings 2002 IEEE International Conference on Multimedia and Erpo (Cat.No.02TH8604), 1:113-116, August 2002. Conference date 26-29 Aug. 2002.

[8] Kimber D. and Wilcox L. (1997), “Acoustic segmentation for audio browsers”. Computing Science and Statistics. Vol.28. Graph-Image-Vision. Proceedings of the 28th Symposium on the Interface. Interface’96, (295-304),1997.

[9] Albiol A., Torres L., and Delp E. J. (2003), “The indexing of persons in news sequences using audio-visual data”. Proceedings of the 2003 IEEE International Conference on Acoustics,Speech, and Signal Processing, ICASSP’03, 3:137-140, April 2003.

[10] Abe M. and Nishiguchi N. (2002), "Self-optimized spectral correlation method for background music indentification". Proceedings 2002 IEEE International Conference on Multimedia and Expo(Cat. No02TH8604), 1(1):333-336, August 2002.

[11] Chien J.-T. (2003), “Linear regression based Bayesian predictive classification for speech recognition”. IEEE Transactions on Speech and Audio Processing, 11(1):70-79,January 2003.

[12] Gouyon F., Pachet F., and Delerue F. (2000), "On the use of zero-crossing rate for an application of classification of percussive sounds". Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), December 2000.