A FAST AUDIO CLASSIFICATION FROM MPEG CODED DATA

Yasuyuki Nakajima, Yang Lu, Masaru Sugano, Akio Yoneyama, Hiromasa Yanagihara, and Akira Kurematsu**

KDD R&D Labs.
2-1-15, Ohara, Kamifukuoka
Saitama, 356-8502 JAPAN

*University of Electro-Communications
1-5-1, Chofugaoka, Chofu
Tokyo, 182-8585 JAPAN

ABSTRACT

Audio information classification becomes a very important task for such purposes as automatic keyword spotting and other content-based audio-visual query system. In this paper, we describe a fast and accurate audio data classification method on MPEG coded data domain. Firstly silent segments are detected using a robust approach for different recording conditions. Then the non-silent segments are classified into three types, music, speech, and applause using temporal density, bandwidth and center frequency of subband energy. In order to be robust for a variety of audio sources as much as possible, we use Bayes discriminant function for multivariate Gaussian distribution instead of manually adjusting a threshold for each discriminator. In the experiment, every one-second MPEG audio data is classified and about 90% of audio and speech segments have been successfully detected. As for the detection speed, less than 20% of MPEG audio decoding processing power is required.

1. INTRODUCTION

Construction of efficient audio-visual data analysis is very important for content-based indexing, browsing, and retrievals from multimedia databases. For example, as a fundamental analysis for video signal, shot segmentation algorithms from video signal such as MPEG video have been proposed[1,2]. In these algorithms, shot boundaries including video effects such as dissolve and wipe transitions are detected on coded data domain.

Recently research on video segmentation combined with audio signal analysis[3-7] has also been reported. For example, in [3] video is segmented into shot level and audio information in each shot is classified into dialog, non-dialog, and silent. Then in [4], the above method is incorporated with speaker identification where only dialog shot is investigated. The other proposals in [5,6] directly use audio information to enhance video shot detection accuracy. For example, it was reported in [6] that since most of shot boundaries in TV news correspond to silent segments, silent detection may improve scene change detection accuracy in this kind of video source.

As for audio indexing, several methods have been reported[3,6,8-10]. For example, in [9], audio source is classified by measuring a similarity between input sound and predefined variety of sounds. In [8], speech/music classification is performed by exploiting lopisidedness of ZCR distribution where speech signals show a marked rise that is not common for music signals. In [10], 5 to 13 feature vectors including such features as cepstral coefficients are used to classify audio source in order to enhance the classification accuracy. Although all the above analyses have been conducted for PCM data so that decoding process is required for coded data before the analysis, reference sited in [3] proposed audio classification on MPEG subband domain. In this algorithm, classification is performed thresholding such discriminants as pitch and band energy ratio using subband data which can be extracted from coded bitstream without any high processing power[11]. Although thresholding technique is relatively easy to handle when the number of thresholds is small, it becomes very difficult to handle when the number of thresholds is increased in order to enhance the classification accuracy or increase the number of classes.

In this paper, we propose a fast and accurate audio classification algorithm from MPEG coded data based on statistical discriminant functions. MPEG audio data is classified into 4 classes, silent, speech, music, and applause segment. Although noise or other sound effects like rain and car engine sounds may be also important features for audio indexing, we include applause sound for the classification on the reason that applause segments can be used for such semantic segmentation as boundaries of music pieces[9] and talk shows. Temporal resolution of audio classification was empirically chosen to be one second because the length of speech and music is much longer than that and subjectively manual classification can be carried out without any difficulty for more than one second.

In the following sections, classification algorithm of silent, music, speech, and applause sounds from MPEG coded audio data are discussed. Then the classification experiments and their results are shown using several TV programs.

2. CLASSIFICATION ALGORITHM

2.1 Silent Segment Detection

When compared with silent and non-silent segments, silent segments usually have much smaller energy distributions than that of non-silent segments. Therefore, most of silent segments can be distinguished from that of non-silent by simple thresholding of total energy in a segment. However, this method may fail to detect non-silent segments with low loudness by fade effect or different recording conditions.

The other differences of silent and non-silent segments are that a non-silent segment normally has a certain level of variation in its loudness in the lower frequency, whereas a silent segment has a much smaller variation. Figure 1 (a) and (b) show examples of silent and non-silent MPEG audio subband energy in time and frequency domain. As can be seen in the figures, most of

significant subband energy is confined to lower subbands in both cases and the variations of subband energy can be compared easily. We use the variance of subband 0 energy for silent segment detection. In addition, since instantaneous audio spectrum can be regarded as stationary, we extracted one of subband 0 data for each MPEG audio frame. This subsampling of subband data in time–frequency domain can enhance detection speed (38 data for 1-second MPEG layer II data at 44.1kHz) while it can maintain detection accuracy.

The variance of subband energy, σ_e^2 can be obtained as

$$\sigma_e^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left(sb_{o,j}^2(n) - \left\langle sb_{o,j}^2(n)\right\rangle\right)^2 \quad (1)$$

where, N is a total number of frames in one second and <> is averaging operation. $sb_{i,j}(n)$ is a subband sample at band i, group j in a frame n (we chose j=0 in the experiment). For example, in MPEG1 layer II, there are 1152 samples in a frame, therefore, j=0,..35.

The silent segment is declared if σ_e^2 is smaller than the predetermined threshold. The further processing in the following sections is applied only for the non-silent segments.

2.2 Music / Speech Characteristics

Figures 2(a) and (b) show examples of temporal energy distributions of speech and music, respectively. Here, subband energy is accumulated for all 32 bands and defined as *r-sample*. Therefore the *r-sample* in the horizontal axis represents every other 32-subband sample data. As can be seen in the figures, when speech is compared with music, speech has intermittent energy distribution whereas music has continuous distribution except for such cases as dram solo play. In addition, the number of silent *r-sample* varies widely in speech, whereas in music it is small and its variation is also small.

In order to measure the temporal energy density, we firstly normalized subband energy data. Each subband energy value of *r-sample* is compared with a predetermined threshold and normalized to "1" if it has a higher value, otherwise set to "0". Normalized *r-sample* data for Figure2 (a) and (b) are shown in Figure2 (c) and (d), respectively. Then each silent *r-sample* ("0" sample) section is defined as s_i (i is section number) and the number of *r-sample* within the silent section is counted. The energy density D_e is measured as log value of the variance of series s_i.

$$D_e = \log_{10}\left[\frac{1}{M}\sum_{i=1}^{M}\left(s_i - \langle s_i\rangle\right)^2\right] \quad (2)$$

where M is number of silent sections in one second. Logarithm form is used in order to accommodate to Gaussian distribution described in later section.

Other characteristics of speech and music are that the bandwidth of speech is usually narrower than that of music. This property can be easily understood in subband domain. For example, in our preliminary experiment, music signal usually has wide range distribution with up to subband number 20 (which corresponds to about 14kHz when 44.1kHz sampling) or more, when coded at around 100kbit/s per channel. On the other hand, the subband range in speech rarely goes beyond subband number 10. In order to investigate the bandwidth of subband quantitatively, we calculate an average number of subbands (AN_{sb}) with significant level. If a one-second segment has broad bandwidth like in music, then AN_{sb} becomes large. This value can be obtained as follows. Firstly, a group of whole subbands ($sb_{0,j}(n)$ - $sb_{31,j}(n)$, j is a group number as stated in section 2.1) is sampled from MPEG frame data (j=0 is used in the experiment). Then the following normalization is employed for subband energy in order to absorb sound level dependency on audio source. Here, normalized subband energy, $EN_{i,j}(n)$, is defined as

$$EN_{i,j}(n) = 10\log_{10}(sb_{i,j}^2(n)/\max(sb_{i,j}^2(n))) \quad (3)$$

where max() operation is performed for subband i=0,1..31 at group j, frame n. Then significant subband $ssb_{i,j}(n)$ is determined as "1" when a normalized subband energy has higher value than the predetermined threshold, otherwise set to "0".

Then, AN_{sb} is obtained as

$$AN_{sb} = \frac{1}{nsf}\sum_{n\in non-silent}\sum_{i=0}^{31} ssb_{i,j}(n) \quad (4)$$

Here, *nsf* is number of frames with at least one significant subband.

2.3 Applause Characteristics

As described earlier, applause sound can be regarded as start or end point of concert, talk show, or sit-com. Therefore, the detection of applause sound may be useful for detection of "audio scene change" as well as for the further content-based analysis. When compared with speech and music, applause sound has continuous similarity and center frequency is relatively stable. Therefore we estimate the center frequency of subband by calculating the subband centroid for each MPEG audio frame and obtain its average and variance for one second.

The subband centroid $c_f(n)$ of MPEG audio frame n can be calculated as

$$c_f(n) = \sum_{i=0}^{31} isb_{i,j}^2(n) / \sum_{i=0}^{31} sb_{i,j}^2(n) \quad (5)$$

Then its average and variance, $<c_f>$ and σ^2_{cf} can be obtained as

$$< c_f >= \frac{1}{N}\sum_{n=0}^{N-1} c_f(n) \quad (6)$$

$$\sigma_{cf}^2 = \frac{1}{N}\sum_{n=1}^{N-1}\left(c_f(n) - < c_f >\right)^2 \quad (7)$$

Figure 3 shows the distribution of applause sound in average subband centroid and its variance domain. As can be seen from the figure, both the ranges of average frequency and its variance are smaller than the other sounds.

2.4 Discriminant Function

In order to discriminate music, speech, and applause, we applied Bayes discriminant function for multivariate Gaussian distribution[12].

$$f_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k) + [\log\ p(\omega_k) - \frac{1}{2}\log|\mathbf{C}_k|] \quad (8)$$

Here, $\mathbf{x}$ is input data vector, $\mathbf{m}_k$ is mean vector of class k, $\mathbf{C}_k$ is covariance matrix of class k, and $p(\omega_k)$ is *a priori* probability of class ω_k. We use four dimensional vectors described in the above sections 2.3 – 2.4. These are subband energy density D_e, average number of subbands AN_{sb}, and average and variance of subband centroid, $<c_f>$ and σ^2_{cf}, respectively.

3. EXPERIMENT

Firstly, two TV programs for 1000sec in total coded by MPEG1 layer II at 112kbit/s per channel with 44.1kHz sampling were used for supervised training. Each one-second audio was labeled manually by listening, and data were collected for each class. Although there were mixed sound segments in time and frequency domain, subjectively dominant sound was chosen in the labeling. Then the parameters in discriminant function for each class in Equation (8) was determined. The threshold values for normalization are also determined in the above sequences.

Figure 4 shows a block diagram of detection flow. MPEG coded data is decomposed into subband domain and each one-second data is investigated. Only non-silent segments are forwarded further, and then speech, music and applause discrimination is performed. Experiment has been conducted using several MPEG coded TV programs different from the training sequences. They include TV news program, talk show, and music program with audience. Table 1 shows the detection results. Here, correct detection ratio is defined as (number of correctly detected) / (number of correct segments) x 100. Similarly, false detection ratio is defined as (number of false detection) / (number of detected segments) x 100. Although silent and speech segments are successfully detected, detection accuracy of music and applause is not so high as that of speech. In addition, false detection ratio of applause is large. After analyzing these results, we found that miss classification was mainly occurred in between music and applause. Many music segments were classified as applause. Since applause sound may also have high subband energy density like music, it resulted in low detection accuracy of music and high false detection ratio of applause.

Improved performance can be obtained by separating applause detection from speech/music detection. Figure 5 shows the block diagram of improved detection flow. After the silent segment detection, applause sound detection is performed using Equation (8). In this case, applause and non-applause classification is performed. Then only non-applause segments are further classified into speech and music. The results are shown in parentheses in Table 1. When compared with the previous method, detection accuracy of music and false detection of applause are greatly improved although detection accuracy of speech and applause has been slightly decreased.

As for the errors in silent segment detection, they were mainly occurred in mixed sound segments. For example, when one-second sounds are composed of end of speech and silent segment, some of the results are opposite in manual and proposed method. For these kinds of sounds it may be appropriate to classify them as transition sound by introducing discriminant like similarity to neighboring segments. This also applies to miss classification of applause. In music classification, segments with intermittent sound like drum solo are often classified as speech. In this case, improvement of classification accuracy can be expected by incorporating such feature vector as rhythm detection with a certain length temporal windows.

As for the detection speed from MPEG bitstream, 16.1% of CPU time is used when compared with full decoding of MPEG audio data using about 160MIPS HP9000 workstation.

4. SUMMARY

We have proposed a very fast and accurate MPEG audio classification algorithm on subband data domain. Classification was performed for silent, speech, music and applause segments at one-second unit. After discriminating non-silent segments, MPEG audio stream was classified using Bayes discriminant function for multivariate Gaussian distribution. In the experiment, although 3% to 16% false detection ratios are found in each classification, music and speech have been successfully detected at around 90% accuracy. Since less than 20% of MPEG audio decoding CPU time is used in the classification, it is expected to be used for preprocessing of automatic transcription in digital audio archive, or it can be realized as fast audio-visual indexing tool of MPEG data by combining such analysis as video scene change detection on MPEG data domain.

5. REFERENCES

[1] B.L.Yeo and B.Liu, "Rapid scene analysis on compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol.5, pp.533-544, Dec. 1995.

[2] M.Sugano, Y.Nakajima, H.Yanagihara, and A.Yoneyama, "A fast scene change detection on MPEG coding parameter domain," presented at *IEEE ICIP*, MP.10.09, Oct. 1998.

[3] N.V.Patel and I.K.Sethi, "Audio characterization for video indexing," *SPIE*, vol.2670, pp.373-384, 1996.

[4] N.V.Patel and I.K.Sethi, "Video classification using speaker identification," *SPIE*, vol.3022, pp.218-225, 1997.

[5] J.Nam and A.H.Tewfik, "Combined audio and visual streams analysis for video sequence segmentation," *IEEE Proc.ICASSP*, pp.2665-2668, 1997.

[6] C.Saraceno and R.Leonardi, "Audio as a support to scene change detection and characterization of video sequences," *IEEE Proc.ICASSP*, pp.2597-2600, 1997.

[7] J.S.Boreczky and L.D.Wilcox, "A Hidden Markov Model framework for video segmentation using audio and image features," *IEEE Proc.ICASSP*, pp.3741-3744, 1998.

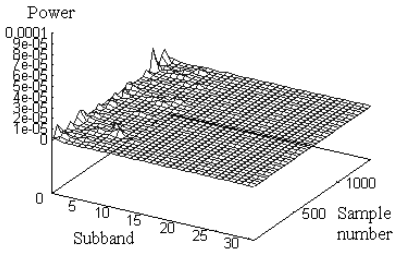
[8] J.Saunders, "Real-time discrimination of broadcast speech/music," *IEEE Proc.ICASSP*, pp.993-996, 1996.

[9] E.Wold, T.Blum, D.Keislar, and J.Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, pp.27-36, Fall, 1996.

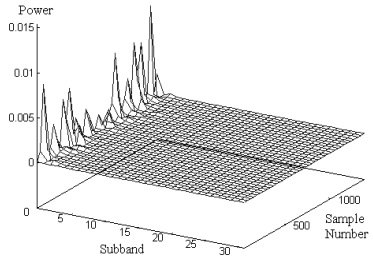
[10] E.Scheirer and M.Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *IEEE Proc.ICASSP*, pp.1331-1334, 1997.

[11] Y.Nakajima, H.Yanagihara, A.Yoneyama, and M.Sugano, "MPEG audio bit rate scaling on coded data domain," *IEEE Proc.ICASSP*, vol.6, pp.3669-3672, May 1998.

[12] S-T. Bow, "Pattern Recognition and Image Preprocessing," Marcel Dekker, 1992.

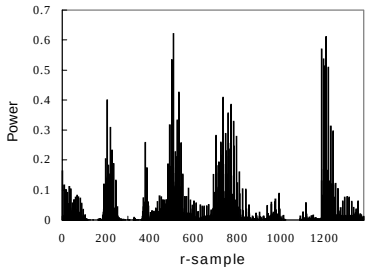


(a) Silent source

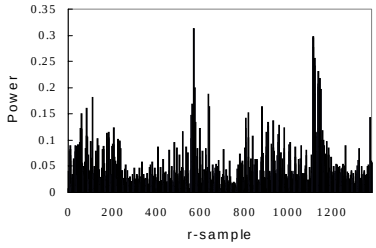


(b) Non-silent source

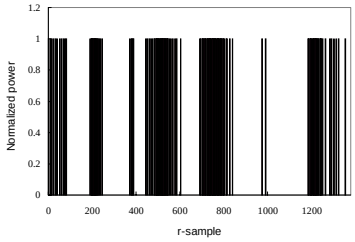
Figure 1. Examples of **(a)** silent, and **(b)** non-silent MPEG audio subband energy.



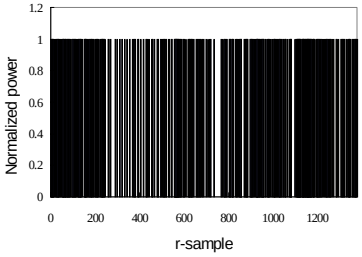
(a) Energy distribution of speech



(b) Energy distribution of music



(c) Normalized energy distribution of speech



(d) Normalized energy distribution of music

Figure 2. Energy distribution and normalized energy distribution of *r-sample* for speech and music sources

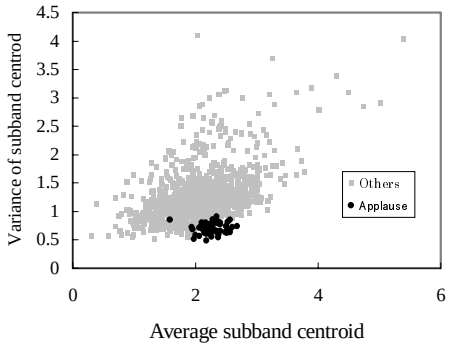


Figure 3. Applause and other sounds in average subband centroid and its variance domain

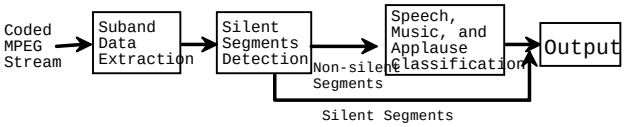


Figure 4. Block diagram of MPEG audio classification flow

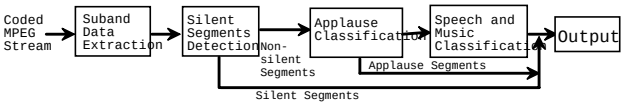


Figure 5. Block diagram of improved MPEG audio classification flow

Table 1. Classification results

Class	Silent	Speech	Music	Applause
Correct(s)	43	1053	177	27
Detect(%)	90.7	97.1(93.3)	76.3(88.1)	81.5(74.1)
False(%)	13.0	3.4(3.5)	9.8(16.4)	41.0(14.8)