BROADCAST NEWS SEGMENTATION BY AUDIO TYPE ANALYSIS

Tin Lay NWE and Haizhou LI

Institute for Infocomm Research, Republic of Singapore
tlnma@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

It is common that we define audio types according to human perception instead of audio spectral properties. In this paper, we analyse the spectral properties of audio types and propose the acoustic features based on spectral properties and harmonic enhancement to classify audio. By analyzing the spectral properties of sound types, a multi-model HMM is proposed to integrate the primitive spectral properties in statistical modeling. To validate the approach, we build a classifier to segment audio streams into speech, commercials, environmental sound, physical violence and silence in multiple steps. It is shown that proposed approach outperforms conventional methods. Experimental evaluations on 20 audio tracks of TRECVID broadcast news database have shown the effectiveness of the proposed approach.

1. INTRODUCTION

Rapid increase in the amount of audiovisual data demands an efficient method for automatic information retrieval. Segmentation of the audio streams into meaningful regions is an important step and can provide a clue towards detecting a particular event. For example, the presence of crowd cheering can be an important clue for detecting events such as soccer or basketball matches.

Most studies on audio segmentation focus on speech/music discrimination since these are the two most important types of audio. A large number of features have been suggested in the literature. These include 4 Hz modulation energy, spectral roll-off point, spectral flux, spectral centroid [1] and harmonic coefficients [2]. Features based on spectral analysis mainly aim to capture the spectral variations between speech and music signals [3]. Kim [4] states that the presence of harmonic contents in the spectrum is the indication of music.

Only a few research works have been done to detect the environmental sound and crowd cheering in comparison with speech/music discrimination task. Baillie [5] used MFCC features to detect crowd cheering events in the audio stream. Zhang [6] states that many environmental sounds are non-harmonic and these are detected using features such as energy function curve, average zero crossing rate [6], band periodicity, spectrum flux and noise frame ratio [7]. In general, features based on harmonic contents and spectral analysis seem to be important for audio segmentation.

Different strategies have been employed to segment audio into several categories. Most studies use pattern classifiers such as Gausian Mixture Model (GMM), Neural Network (NN) and Nearest Neighbor (NN) [1, 2, 3]. In those studies, a statistical model for each of the audio categories (for example, speech or music) is created using the available training samples. However, this approach is not effective for audio modeling of commercials, which typically include different audio types such as music, singing and speech with music background and have diverse spectral properties. In this case, spectral variations should be taken into consideration in statistical modeling.

In this research, we explore a new approach towards broadcast news audio segmentation. Broadcast news includes two main components: news story and commercial breaks. Several audio types can present within news story such as speech, environmental sound and physical violence. It is noted that the audio types are defined according to human perception instead of audio spectral properties. For example, environmental sounds include sounds of road traffic, flooding and storming. Sounds such as crowd cheering and shouting are categorized into the audio type named physical violence. We segment audio streams into five different types including speech, commercials, environmental sound, physical violence and silence. After analyzing the audio type, we propose a hierarchical method that exploits the merits of a particular feature in detecting a specific audio type and models primitive spectral features present in an audio type for audio segmentation.

The paper is organized as follows. In Section 2, characteristics of five audio types are analysed to derive suitable features for audio segmentation. Details of the multi-model HMM classifier formulation are discussed in Section 3. The experiment setup and results are discussed in Section 4. Section 5 concludes the paper.

2. ACOUSTIC FEATURES

Short-time energy of the audible sound is in general significantly higher than that of silence segments. Therefore, the short-time energy [8] is computed from the audio signal to detect silence.

Music as well as singing signals are rich in harmonic content in comparison with other audio signals such as speech [4]. Therefore, commercial audio is typically harmonic because of its musical background and non-commercial audio such as speech, physical violence and environmental sound is non-harmonic [4,6]. Examples of the frequency content of commercial and non-commercial (physical violence) audio are illustrated in Figures 1(a) and 2(a) respectively.

Furthermore, the spectral characteristics of commercial and non-commercial segments are different. For example, spectral energy of speech with music background is higher than that of pure speech.

0-7803-8874-7/05/$20.00 ©2005 IEEE

Therefore, harmonic content enhancement and analysis of energy distribution in different frequency bands will provide discriminative information between commercial and non-commercial audio. Next, we discuss how to derive the discriminative acoustic features from the audio segments.

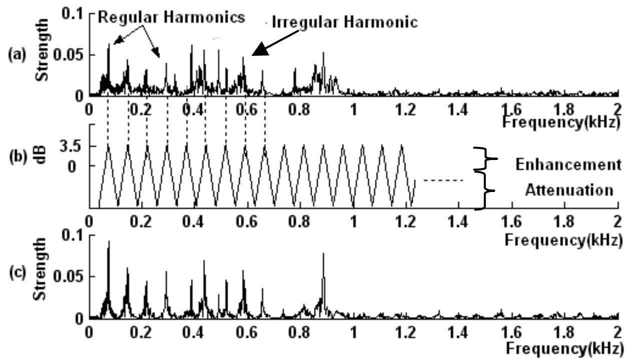

Figure 1: (a) Commercial audio in frequency domain (b) Frequency response of triangular bandpass filter (c) Commercial audio after harmonic enhancement

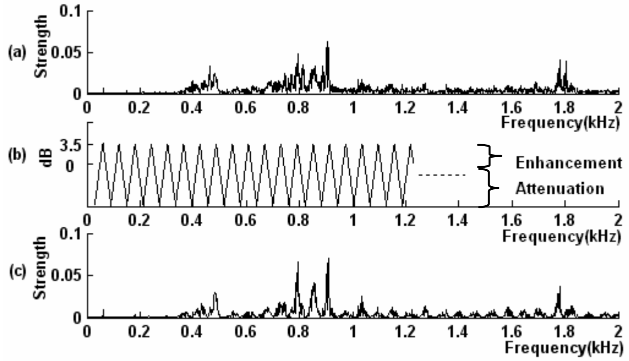

Figure 2: (a) Non- commercial audio (physical violence) in frequency domain (b)Frequency response of triangular bandpass filter (c) Non-commercial audio after harmonic enhancement

2.1. Harmonic enhancement

The audio signal is divided into frames of 20ms windows with 13ms overlaps. Each frame is humming-windowed to minimize signal discontinuities at the end of each frame. Harmonics are detected by using triangular bandpass filters shown in Figure 3, covering a frequency range from 0Hz to 16kHz. This frequency range is chosen because the audio stream, including signals such as music and physical violence, has high frequency energies. We locate the fundamental frequency with which the signal is most attenuated [4] by varying the bandwidth of the triangular filters.

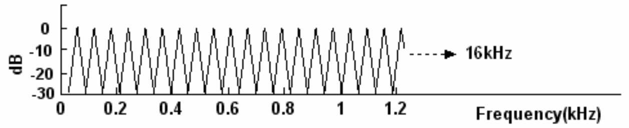

Figure 3. The frequency response of triangular bandpass filters

Using information of harmonic locations, the triangular filters shown in Figures 1(b) and 2(b) are implemented to enhance harmonic content and to attenuate other frequency content. According to [9], harmonic patterns of singing and music signals are not always regular. Therefore, the filter is implemented to have signal enhancement at regular harmonic as well as irregular harmonic frequencies as illustrated in Figures 1(a) and 1(b). This process enhances the energy content of commercial audio segment and suppresses that of non-commercial segment in general, as shown in Figure 1(c) and 2(c).

2.2. Energy distribution analysis

After harmonic enhancement, each audio frame passes through bandpass filters spaced logarithmically from 130 Hz to 16kHz. Subband based Harmonic Enhanced Log Frequency Power Coefficients (HE-LFPC) are then computed using Equations (1) and (2) which have been defined previously for LFPC calculation in [10].

$$S_t(m) = \sum_{k=f_m-\frac{b_m}{2}}^{f_m+\frac{b_m}{2}} X_t(k)^2, \quad m = 1.......12 \tag{1}$$

where, $X_t(k)$ is the k^{th} spectral components of the signal at frame index t , $S_t(m)$ is the output of the m^{th} subband, while f_m and b_m are the center frequency and bandwidth of the m^{th} subband respectively. The HE-LFPC parameters which provide an indication of energy distribution among subbands are calculated as follows:

$$HE-LPFPC_t(m) = 10\log_{10}\left[\frac{S_t(m)}{N_m}\right] \tag{2}$$

where N_m is the number of spectral components in the m^{th} subband. For each frame, 12 HE-LFPCs are obtained.

To further classify the non-commercial audio into speech, environmental sound and physical violence, their spectral properties are analyzed. Physical violence has higher tempo and volume in comparison with speech. Therefore, their spectral properties could be different. In addition, speech has pauses between words and/or phrases. However, environmental sound such as road traffic has no such pauses. To investigate their spectral properties, spectral energy distributions of these audio signals are obtained by computing LFPC features [10] and presented in Figure 4. The process of harmonic enhancement is not carried out since these signals are non-harmonic.

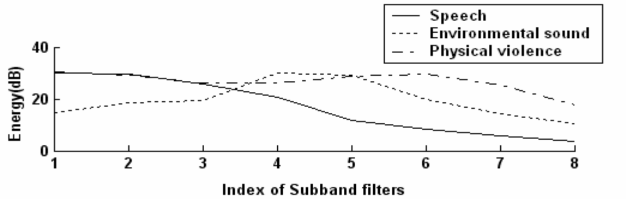

Figure 4. Energy distributions of speech, environmental sound and physical violence

Figure 4 shows that physical violence has the highest high frequency energy followed by environmental sound while speech signal has the lowest high frequency energy. Therefore, LFPC feature is an effective feature for the discrimination among three non-commercial audio types.

3. CLASSIFIER FORMULATION

Most studies on speech/music discrimination use statistical pattern classifiers [1, 2, 3]. However, to our knowledge, none of the studies takes into account the variations in signal characteristics within each audio type. An important observation is that music as well as singing signals have different tempos and intensity levels. In general, soft music signals have lower tempo and intensity than loud music. We assume the tempo of the music to be constrained between 40~185 beats per minutes (BPM). We divide music into high and low tempo classes according to a fixed threshold, which is 70BPM in our current implementation. Similarly, we divide music into loud and soft classes according to a threshold, which is determined by music and singing segments in the training data set. In addition, an audio type of speech includes male and female speech, with or without noise. Speech of an anchorperson is usually free from noise and speech of a reporter is often coupled with noise such as environmental or channel noise. Instead of creating only one model for speech using all noise clean and noisy samples of male and female speech, it could be more effective to model the primitive spectral sub-classes explicitly in a multi-model HMM. In Figure 5, manual labeling of commercial and non-commercial audio training data are illustrated.

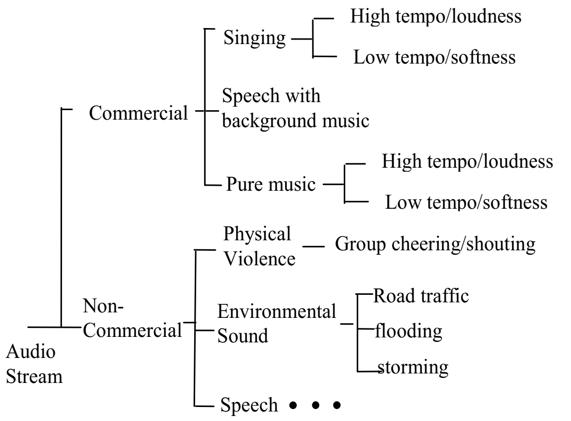

Figure 5. Manual audio classification tree based on similarity of audio properties for creation of multi-model HMM

A model is created using the signals of similar properties. For example, the commercial audio class has five models including one speech with background music model, two singing models (high and low tempos) and two music models (high and low tempos). In our current implementation, we use 13 models in total. This process results in multiple HMM models for each audio type. Several models for each class form an HMM model space to allow more accurate modeling as compared to the single model baseline.

We construct a hierarchical classifier by having a tree of nodes, including a root node, branching nodes and leaves. At root node, silence vs. non-silence is decided using short-time energy threshold. Each of the branching nodes and leaves is associated with an MM-HMM model. The classifier makes a decision by comparing the scores from all possible nodes at the same level to select the best matched node. The decision tree classifier follows the similar structure to Figure 5 for audible sounds.

4. EXPERIMENTS

4.1. Experimental set up

Our experimental database includes 44 audio tracks of ABC World News Tonight and CNN Headline News extracted from TREC Video Retrieval Evaluation 2003 (TRECVID 2003) development data. Each audio track is 30 minutes long and the database comprises 22 hours of audio. Each audio track is annotated manually to obtain commercial, speech, physical violence, environmental sound and silence segments to provide ground truth data. This ground truth data is used to evaluate the system performance. We use the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models in the experiments. Using the training data set, the MM-HMM classifier is trained to obtain several variants of HMM models for each audio type as shown in Figure 5. The frame log-likelihoods from the HMM over a test unit are accumulated for decision making. The unit length is set to one second to provide significant statistics [11].

First, the test audio track is blocked into one second test frames, and then, short-time energy is computed from 20ms with 13ms overlapping subframes. Then, silence detection is performed by an energy threshold. If the test frame is audible sound, the HE-LFPC feature is obtained for commercial/non-commercial classification. If the signal is non-commercial, it is further classified into speech, environmental sound and physical violence by using the LFPC feature.

Several experiments are conducted. First, five audio types are detected on a database including 10 hours of audio using proposed hierarchical method. The training and test dataset include 6 hours and 4 hours of audio respectively. Then, MM-HMM training method is compared with baseline HMM training approach using different sizes of training data.

4.2. Rule based post processing

Speech without music background sometimes presents within the commercial audio. These segments are often misclassified as pure speech. To correct this, we apply rule based error correction methods. We use a statistical method to build a statistical duration model based on the manual annotation of 5 audio tracks from our training data. The duration models (the normalized histogram) of all 5 audio categories are depicted in Figure 6.

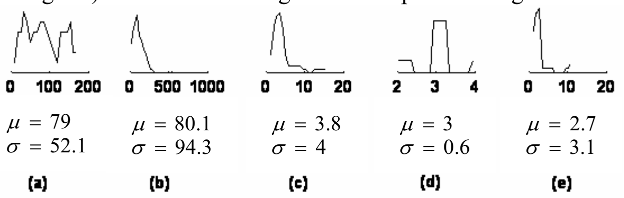

Figure 6: Duration distributions of (a) commercials, (b) speech, (c) environmental sound, (d) physical violence, (e) silence segments. X-axis represents duration in seconds.

The statistics in Figure 6 are translated into rules in the post-processing. For example, it is observed that commercials and speech have significantly longer durations than others. Therefore, short commercials and speech segments are merged into their neighbors and the segments is re-aligned.

4.3. Results and discussion

The first column of Table 1 shows the average detection accuracy of proposed approach. To compare performance of our proposed hierarchical method with other features, experiments are conducted using only LFPC (without harmonic enhancement) [10], MFCC [12] and LPCC [13]. It can be seen that the proposed hierarchical method outperforms other features. It is observed that performing the harmonic enhancement and hierarchical process gives (2.8%) improvement in performance over simple LFPC (85.8% to 83.0%).

Table 1: Performance comparison between proposed feature set and traditional features (%)

Audio Type	FEA1	FEA2	FEA2	FEA4
Commercials	78.9	73.8	51.7	85.7
Speech	90.3	87.5	91.5	70.9
Environmental Sound	81.5	88.9	88.9	59.3
Physical Violence	78.9	73.7	84.2	68.4
Silence	99.3	91.3	92.4	81.2
Average	**85.8**	**83.0**	**81.8**	**73.1**

FEA1=Energy/ HE-LFPC/ LFPC, FEA2=LFPC,
FEA3=MFCC, FEA4=LPCC

Next, we investigate the effectiveness of using MM-HMM. We compare the performance of the MM-HMM against the baseline HMM training method in which only one model is created for each audio class. First, the experiments are conducted using the same number of mixtures per state (2mixtures/state) for both MM-HMM and baseline HMM on 6 hours of training data. The results in Table 2 show that the MM-HMM training method outperforms the baseline HMM training approach.

Table 2: Comparison between MM-HMM and baseline HMM using training data of 6 hours, 12 hours and 18 hours

HMM settings	No. of mixtures	Average Accuracy (%)		
		6 hrs	12 hrs	18 hrs
MM-HMM	2	85.8	85.6	86.2
Base-line HMM	2	81.7	82.2	80.5
Base-line HMM	10	73.9	75.7	78.1

The baseline HMM has about 7 times less free parameters than the MM-HMM as MM-HMM has 13 models in total. To ensure fair benchmarking between the baseline HMM and the MM-HMM, we perform further experiments using a baseline HMM with 10 mixtures per state. Since HMM with 10 mixtures per states has much more free parameters, a large amount of training data are needed for the model to perform well. Hence, we conduct experiments using 12 hours and 18 hours of training data. The results presented in Table 2 show that HMM with 10 mixtures per state perform better when given more training data. However, MM-HMM training method performs the best in all experiments. In the baseline HMM, we expect the HMM training to self-organize the audio class in an unsupervised way. The modeling turns out to be less effective than supervised MM-HMM training using manual labeled data with less mixture per state.

5. CONCLUSION

Audio classes can be defined at different levels of granularity in different ways. It is common that we define audio type according to human perception instead of acoustic spectral properties. As a result, an audio type could include diversified spectral properties. Many conventional methods in audio segmentation try to model the ill-defined audio type using statistical modeling with spectral features. In this paper, we propose MM-HMM training methods by using several variants of HMM models for each audio type to capture spectral variations. In addition, we propose a hierarchical segmentation method using energy, HE-LFPC and LFPC features. The proposed approach outperforms the traditional methods and achieves a segmentation accuracy of 85.8%.

6. REFERENCES

[1] E. Scheirer and M. Slaney, **"**Construction And Evaluation of A Robust Multifeature Speech/Music Discriminator," *Proc. ICASSP*, pp. 1331-1334, 1997.

[2] W. Chou, and L. Gu, "Robust Singing Detection in Speech/Music Discriminator Design," *Proc. ICASSP*, pp. 865-868, 2001.

[3] K. El-Maleh, M. Klein, G, Petrucci, and P. Kabal, "Speech/Music Discrimination for Multimedia Applications," *Proc. ICASSP*, pp. 2445-2448, 2000.

[4] Y. Kim, and B. Whitman, "Singer Identification in Popular Music Recordings Using Voice Coding Features," *Proceedings of Int. Symposium on Music Information Retrieval*, 2002.

[5] M. Baillie and J.M. Jose, "An Audio-based Sports Video Segmentation and Event Detection Algorithm," *IEEE Workshop on Event Mining,* Washington DC, USA, July 2, 2004

[6] T. Zhang, C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 441 – 457, 2001.

[7] L. Lu, H. Jiang and H.J. Zhang, "A Robust Audio Classification and Segmentation Method". *ACM Multimedia*, pp. 203-211, 2001.

[8] J.R. Deller, J.G. Proakis and J.H.L. Hansen. Discrete-Time Processing of Speech Signals, Macmillan Pub. Co, Toronto, 1993.

[9] F. Alton Everest, The Master Handbook of Acoustics, McGraw-Hill, New York, 2001

[10] T.L. Nwe, F.S. Wei, and L.C. De-Silva. "Stress Classification Using Subband Based Features," *IEICE Trans.on Info.and Systems*, pp. 565-573, 2003.

[11] W.H. Tsai, H.M. Wang, D. Rodgers, S.S. Cheng, and H.M. Yu, "Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics," *4th International Conference on Music Information Retrieval*, Maryland, USA, 2003.

[12] C. Becchetti and L. Ricotti. Speech Recognition Theory and C++ Implementation. John Wiley &Sons, New York, 1998.

[13] L. Rabiner and B. Juang. Fundamentals of Speech Recognition. Prentice Hall, Englewood, N.J.,1993.