AUDIO FEATURE EXTRACTION & ANALYSIS FOR SCENE CLASSIFICATION

Zhu Liu, Jincheng Huang, Yao Wang
Polytechnic University
Brooklyn, NY 11201
{zhul,jhuang,yao}@vision.poly.edu

Tsuhan Chen
AT&T Lab - Research
Holmdel, NJ 07733
tsuhan@research.att.com

Abstract - Analysis and classification of the scene content of a video sequence are very important for content-based indexing and retrieval of multimedia databases. In this paper, we report our research on using the associated audio information for video scene classification. We describe several audio features that have been found effective in distinguishing audio characteristics of different scene classes. Based on these features, a neural net classifier was quite successful in separating audio clips from different TV programs.

I. INTRODUCTION

Video scene analysis and classification are required in many applications, such as information indexing and retrieval in multimedia databases, video editing, etc. Research in this area in the past several years has focused on the use of speech and image information. These include the use of speech recognition and language understanding technology to produce keywords for each video frame or a group of frames, the use of image statistics (color histogram, texture descriptor, shape descriptors) for characterizing contents of individual frames, etc. [1]. We believe that analysis of the accompanying audio signal can also be beneficial to scene classification. For example, the audio in a sports video is very different from that in a news report. Various sports programs also have very different background sounds. To accomplish scene classification using audio information, the very first and crucial step is to determine appropriate features that can differentiate audio clips associated with various scene classes. In this paper, we describe several audio features that we have explored. We also evaluate the scene discrimination capability of these features using the intra- and inter-class scattering matrices. Finally, we report the classification results based on these features using a neural network.

II. AUDIO FEATURE ANALYSIS

There are many features that can be used to characterize audio signals. Generally they can be separated into two categories: time-domain and frequency domain. In this section, we describe several audio features that we have explored.

Volume Distribution The volume distribution of an audio clip reveals the temporal variation of the signal's magnitude, which is important for scene

classification. To compute volume, we divide an audio clip into many overlapping frames and use the root mean square (RMS) of the signal magnitude within each frame to approximate the volume of that frame. The mean and standard deviation of the volume within a clip are used as descriptors of the volume distribution. We also determine whether a frame is silent or not, by comparing its volume to a threshold determined based on the volume distribution of the entire clip. From the result of silence detection, we calculate the *silence ratio*, which is the ratio of the silence interval to the entire period. We found that this ratio varies quite significantly in different video sequences. In news reports there are regular pauses in the reporter's speech; on the other hand in advertisement programs there are always some background music which results in a low silence ratio.

To measure the variation of an audio clip's volume, we define the *volume dynamic range* (VDR) as $VDR = (\max(v) - \min(v)) / \max(v)$, where $\min(v)$ and $\max(v)$ represent the minimum and maximum volume within an audio clip. We have found that in sports programs, since there is a nearly constant level of the background sound, the volume does not change a lot. On the other hand, in news and weather reports, there are silent periods between speech, so the VDR is much higher.

Pitch Contour Pitch is the fundamental period of a human speech waveform, and is an important parameter in the analysis and synthesis of speech signals. In an audio signal, which generally consists of pure speech as well as many other sounds, the physical meaning of pitch is lost. But we can still use pitch as a low-level feature to characterize changes in the periodicity of waveforms in different audio signals. Among the many available pitch determination algorithms, we choose one that uses the short time Average Magnitude Difference Function (AMDF)[2] to determine the pitch of each frame. This method is simple and yet can give quite accurate results. Sometimes we cannot find a pitch in the search range chosen based on the speech signal. Such segments are marked as non-speech. After computing the pitch of each frame, we obtain a pitch contour for the entire clip. A median filter is applied to this contour to eliminate falsely detected pitches which often appear as spikes in the contour. We have found that the pitch level itself is primarily influenced by the speaker (male or female) rather than the scene content. On the other hand, the pitch difference between adjacent frames appears to reveal scene content more. We therefore used the mean and standard deviation of the pitch difference as two additional audio features.

Based on the pitch estimation results, we also detect which frames correspond to speech. Because a speech segment usually has a relatively constant pitch, only those frames which have smooth (compared to the previous frame) pitch periods are considered as speech frames. The *speech ratio*, which is defined as the ratio of the length of the speech frames to the entire audio clip, is used as another audio feature. Notice that some of the detected speech frames may actually correspond to music signals that have similar pitch to the human speech.

Frequency Features To obtain frequency features, we first calculate the spectrogram of an audio clip, which is a 2D plot of the short-time Fourier transform

(over each audio frames) along the time axis. Let $S_i(\omega)$ represents the short-time Fourier transform of the *ith* frame. We define the *frequency centroid*, $C(i)$, and *bandwidth*, *B(i)* of this frame as:

$$C(i) = \frac{\int_0^{\pi} \omega |S_i(\omega)|^2 d\omega}{\int_0^{\pi} |S_i(\omega)|^2 d\omega} \qquad B^2(i) = \frac{\int_0^{\pi} (\omega - C(i))^2 |S_i(\omega)|^2 d\omega}{\int_0^{\pi} |S_i(\omega)|^2 d\omega}$$

The mean centroid and bandwidth of an audio clip are used as two frequency domain features. Similar features have been proposed for audio classification in [3].

Since the energy distribution in different frequency bands varies quite significantly among different audio signals, we also use ratios of the energies in different subbands to the total energy as frequency domain features, which are referred to as *subband energy ratios*. Considering that the lower frequency band posses more energy in most audio signals, we divide the entire frequency band [0, *F*] into four subbands: [0, *F*/8], [*F*/8, *F*/4], [*F*/4, *F*/2] and [*F*/2, *F*], where *F* is half of the sampling rate.

III. FEATURE SPACE EVALUATION AND CLASSIFICATION USING NEURAL NETWORKS

Intra- and Inter-Class Separability of the Feature Space To evaluate the scene discrimination capability of the selected features, we calculated the intra-class and inter-class scattering matrices [4]. The intra-class scattering matrix reveals the scattering of samples around their respective class centroids, and is defined by

$$S_{intra} = \sum_{i=1}^{N} P(\omega_i) E\left\{ (X - M_i)(X - M_i)^T | \omega_i \right\},$$

where, $P(\omega_i)$ is the *a priori* probability of class ω_i, X is the sample feature vector, M_i is the mean feature vector (centroid) of class ω_i, *N* is the number of classes. On the other hand, the inter-class scattering matrix is defined as:

$$S_{inter} = \sum_{i=1}^{N} P(\omega_i)(M_i - M_0)(M_i - M_0)^T \text{, where } M_0 = \sum_{i=1}^{N} P(\omega_i) M_i .$$

The diagonal items in these two matrices characterize the intra- and inter-class separability of individual features. If the diagonal item in the intra-class scattering matrix is small while that in the inter-class matrix is large, then the corresponding feature has good class separability. The off-diagonal items in these two matrices reveal the correlation between different features. We can use these measures to eliminate highly correlated features and reduce the dimensionality of the feature space.

A Neural Network Classifier Feedforward neural networks have been used successfully as pattern classifiers in many applications. Conventional multilayer perceptron (MLP) use the all-class-in-one-network (ACON) structure. But such a

network structure has the burden of having to simultaneously satisfy all the desired outputs for all classes, so the number of hidden units tends to be large. Here, we use the one-class-in-one-network (OCON) structure, where one subnet is designated for recognizing one class only [5]. This structure is illustrated in Fig. 1. Each subnet is trained individually using the back-propagation algorithm so that its output is close to 1 if the input pattern belongs to this class, otherwise the output is close to 0. Given an input audio clip, it is classified to the class whose subnet gives the highest score. An advantage of the OCON structure is that one can accommodate a new class easily by adding a subnet trained for this class.

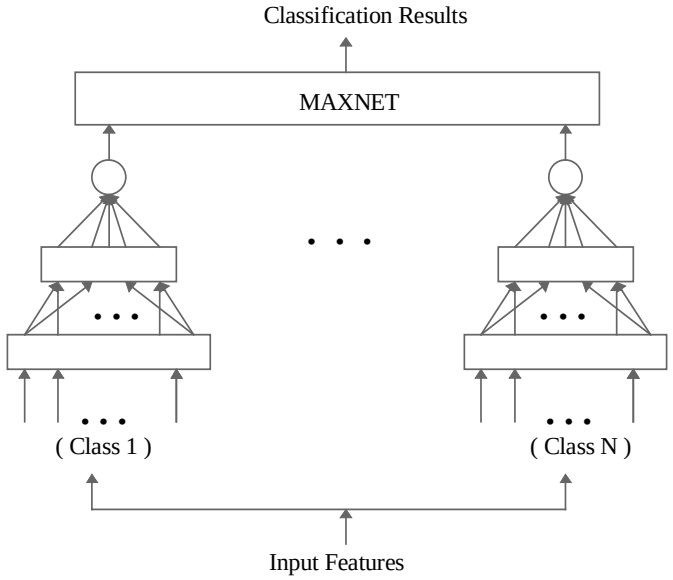

Figure 1 Structure of the Neural Network Used for Audio Classification

IV. EXPERIMENTAL RESULTS

We have collected audio clips from TV programs containing the following five scene classes: news reports by anchor persons, weather reports, advertisement, basketball games and football games. For each scene class, we have 70 audio clips (each one second long and sampled at 22 KHz), 50 of which are used in training the classifier and the others are used in testing. The frame size is 512 samples, each shifted by 128 samples from the previous one. A total of 13 features are extracted for each clip: 1) silence ratio, 2) volume mean, 3) volume standard deviation, 4) VDR, 5) speech ratio, 6) pitch difference mean, 7) pitch difference standard deviation, 8) frequency centroid, 9) frequency bandwidth, and 10)—13) energy ratios of subbands 1— 4.

To see how the above features differ among separate scene classes, Table 1 lists the mean feature vectors for the five different scene classes, obtained by averaging the features extracted from all the clips in the same class. We can see that for most features, football and basketball games have similar values, so do the news and weather reports. Features 3, 4, 9 and 10 differ significantly among advertisement, basketball/football, and news/weather. We note that the clips for advertisement are extracted from 3 different TV commercials and the mean feature vectors extracted here may not be representative of other commercials. Also, the sports clips contain the voices of the commentators. Table 2 lists the diagonal items of the intra-class and inter-class scatter matrices. When calculating these matrices, and when performing classification, all features are normalized by the maximum values of respective features in the training set. We can see that features 3, 4, 9, 10 and 11 have better class separability, while feature 5, 6 and 13 have worse class separability. However, in the classification results reported below, all the features are used.

Tables 3 shows the performance of the neural net classifier on testing data. In this experiment, one hidden layer is used in each subnet and the neuron number of the hidden layer is 7. The classifier can recognize the advertisement, basketball and football successfully, but it cannot separate the news and weather reports very well. This is not surprising because these two types of audio clips contain primarily speech and are hard to differentiate. To distinguish between these two classes, some higher level correlation between successive clips that reflect the flow of the conversation may be needed. We have also used a ACON neural net (1 hidden layer with 14 nodes) to perform the classification. The results are not as good as those obtained with the OCON structure. The ACON classifier cannot separate news and weather report at all for the testing data.

V. CONCLUDING REMARKS

In this paper, we have described several audio features for characterizing the scene content. Among these features, volume standard deviation, VDR, frequency bandwidth, energy ratios in subbands 1 and 2 appear to have a good scene discrimination capability. The effectiveness of the proposed audio features have been confirmed by the classification results using a OCON neural network classifier.

The classification results reported here are meant to show the promise of using audio features for scene classification. Better results should be obtainable with more optimized classification/training algorithms. More rigorous feature space analysis is still required to yield a most efficient set of features.

ACKNOWLEDGMENT This work is supported in part by the National Science Foundation and by the New York State Center for Advanced Technology in Telecommunications at Polytechnic University, Brooklyn, New York

Table 1 Mean Features in Different Classes

Feature	1	2	3	4	5	6 (ms)	7 (ms)
Advertisement	0.139	0.495	0.208	0.845	0.485	0.407	1.18
Basketball	0.000	0.618	0.127	0.615	0.576	0.574	1.19
Football	0.006	0.589	0.156	0.686	0.317	0.629	1.34
News	0.252	0.418	0.268	0.962	0.621	0.167	0.512
Weather	0.315	0.421	0.280	0.970	0.573	0.211	0.634
Feature	8 (Hz)	9 (Hz)	10	11	12	13	
Advertisement	761.9	1016.1	0.885	0.069	0.025	0.021	
Basketball	1575.7	1176.9	0.566	0.271	0.138	0.025	
Football	1275.4	1185.3	0.658	0.161	0.166	0.015	
News	838.6	536.6	0.910	0.055	0.011	0.024	
Weather	611.6	404.4	0.960	0.025	0.007	0.008	

Table 2 Diagonal Entries of Intra/Inter-Class Scattering Matrices

Feature	1	2	3	4	5	6	7
Intra-Class ($x10^{-3}$)	13.36	14.28	9.95	8.37	23.22	12.10	16.72
Inter-Class ($x10^{-3}$)	16.16	10.43	33.68	21.17	14.11	7.83	9.44
Feature	8	9	10	11	12	13	
Intra-Class ($x10^{-3}$)	7.79	10.03	9.01	17.15	5.86	21.21	
Inter-Class ($x10^{-3}$)	11.52	39.54	24.37	38.77	7.6	3.197	

Table 3 Classification Results for Testing Data (unit: 100%)

Data / **Result**	Advertisement	Basketball	Football	News	Weather
Advertisement	100	0	0	0	0
Basketball	0	95	10	0	0
Football	0	5	90	0	0
News	0	0	0	65	10
Weather	0	0	0	35	90

REFERENCES

[1] S. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval," *IEEE Multimedia Magazine* , Vol. 1, No. 2, pp. 62 - 72, Summer, 1994

[2] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.

[3] E. Wold, et al. "Content-based Classification, Search, and Retrieval of Audio," *IEEE Multimedia Magazine*, Vol. 3, No. 3. pp. 27-36, 1996.

[4] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, 1972.

[5] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face Recognition/ Detection by Probabilistic Decision-Based Neural Network," *IEEE Trans. Neural Networks*, Vol. 8, Jan. 1997.