AUDIO SCENE ANALYSIS AND SCENE CHANGE DETECTION IN THE MPEG COMPRESSED DOMAIN

Sarat Venugopal	K.R.Ramakrishnan	S.H.Srinivas	N.Balakrishnan
Dept. E. E. , IISc*	Dept. E. E. , IISc	IRIS†	SERC, IISc
Bangalore	Bangalore	Bangalore	Bangalore
India	India	India	India

Abstract - The use of audio to retrieve and index the associated video is a relatively new approach. In this paper the focus is on MPEG video. For indexing and retrieval one needs to segment the audio stream associated with the video in terms of gender, speech, music and the speaker. This is called "Audio scene" analysis. The paper discusses techniques for such analysis in the MPEG audio compressed domain.

INTRODUCTION

The size and diversity of multimedia databases demand efficient methods of indexing and retrieving information of interest quickly. The earliest attempts segmented the data using the video, extracted key frames and used them as indices. This is normally known as the image-based approach. However certain events in a multiplexed stream are characterized by strong audio cues (e.g., The voice of the commentator or the audience reaching a crescendo, when a goal is scored in a game of football). The use of audio is relatively new for this kind of indexing, but is thought to be powerful. A combination of the imagebased approach and the audio-based approach will provide a more accurate and efficient tool for indexing multimedia documents. In this paper, we discuss techniques for Audio scene analysis in the **MPEG** streams in the compressed domain.

Audio Based Approach to Video Handling

Saraceno et al. [7]have attempted classification by segmenting audio into periods of speech, music, silence and noise. In [8]the authors try to detect not only music but types of music as well. Violence detection based on the audio present during such scenes is also attempted. More recently Minami et al. [10] had used speech-music discrimination for indexing video. There has been few reported efforts dealing with the **MPEG** multiplexed stream. We

*Indian Institute of Science

†Institute of Robotics and Intelligent Systems

0-7803-5610-1/99/$10.00 © 1999

also try to extract features from the segmented speech, to extract information relevant to content-based retrieval.

Audio Scene Analysis in the Compressed Domain

We define the "audio scene" as a sequence of shots, where the audio characteristic do not change. By change of characteristics, we mean, that the audio changes from speech to music, the speaker has changed or the gender of the speaker has changed. It is possible to include more than these transitions, but in this paper we deal with the three types of transitions in audio mentioned above. Audio scene analysis refers broadly to the class of operations which can be performed on the audio associated with the multiplexed stream, so that some semantic sense can be attributed to the segment(s) under consideration.

The audio data rates are much lower compared those of the video. The synthesis filter is the most time consuming module in the decoder. By doing away with this step, further reduction in the computations are achieved.

The MPEG compressed audio is nothing hut the output of a filter-bank. The samples at the output are allocated varying bits as determined by a psycho-acoustic model. A block of samples are quantized and the resulting scale factors and side information is encoded into the stream. At the decoder this information is used to reconstruct the quantized samples. Any of the established filter-bank analysis may be performed on these sub-band samples. Quantization adversely affects the performance, but the resulting error observed was negligible (as compared to the decoded audio), except at very low hit rates.

SEGMENTATION OF AUDIO

In this paper the audio is segmented into segments of speech, music and silence. The features used here were derived from the work reported in [2, 3, 4]. Those features which gave good discrimination and found suitable and efficient for compressed domain processing were used. The following properties were used to achieve a speech-music discrimination. Since the discrimination was intended **as** a pre-processing step, the complexity of computation was also a deciding factor in the selection of the features. The features themselves or their variances were used to effect segmentation.

Tonality. Music tends to be composed of a multiplicity of tones, each with a unique distribution of harmonics. Speech exhibits an alternating sequence of tonal and noise like segments.

Bandwidth. Speech is usually limited in frequency up to about $8KHz$, whereas music tend to occupy the full spectrum, up to $20KHz$.

Excitation **Patterns.** Pitch of speech spans only three octaves hut music usually spans about ***six*** octaves.

Tonal Duration. There is **a** syllabic rate in speech, owing to the fact that the vowels in speech are very regular.

Energy Sequences. That speech follows a pattern of high energy conditions of voicing followed by low energy conditions, is a reasonable generalization. Music is unlikely to exhibit these properties.

A measure of each of the properties listed above were calculated from the quantized samples. The filter-bank output enabled us to avoid extensive frequency analysis. The magnitude of these features determined probabilities of speech and music. These probabilities are weighted with respect to their robustness and summed to give the likelihood for speech and music data. Whichever has higher probability was retrieved.

SEGMENTATION BY SPEAKER IDENTIFICATION

Many a time, persons in a multiplexed stream are more easily distinguished by their voice. Many of the existing speaker identification techniques were considered for the purpose. We adapted the method suggested in [1]. Suitability to compressed domain processing, possibility of fast hardware implementation and length of data needed for training and identification (**as** short **as** possible) were some of the considerations in choosing the above algorithm. The method uses Gaussian mixture models (GMMs), which models the underlying sound classes in a person's speech, without imposing any Markovian constraints, unlike the HMM. Further, arbitrary densities can be modeled by a GMM, containing sufficient number of Gaussian s. The cepstrum was used **as** the feature.

Linear frequency cepstral coefficients were extracted directly from the quantized samples, normalized, and the GMM parameters were estimated using the Expectation-Maximization (EM) algorithm. **A** GMM model was created for each speaker of interest. The test data **was** applied to these models and a likelihood function was maximized to identify the speaker. The identification algorithm works well with $\mathbf{5} - 10sec$ of speech data. If the speaker matched the query, the corresponding segment was retrieved.

SEGMENTATION BY GENDER DISCRIMINATION

The idea is to declare whether a given segment of speech is spoken by a male or a female speaker. Parris and Carey [5] based their work on two types of processing, an HMM based model for the male and the female and an acoustic analysis, for telephone quality speech. Both knowledge sources were then combined and a decision **was** made based on the outputs of them. **A** similar approach was followed in this project, but used the pitch estimation algorithm used in the Multiband Excitation Vocoder (MBE) [6] by suitably modifying it for efficient implementation in the compressed domain. The

Name of the Speaker	No. of Segments Analyzed	No. of Segments Correctly Retrieved
Mohanty	15	9
Nisha	11	11
Lalitha	16	10
Prakash	9	7

Name of the Speaker	No. of Segments Analyzed	No of Segments Correctly Retrieved
Mohanty	**15**	**9**
Nisha	**11**	**11**
Lalitha	**16**	**10**
Prakash	**9**	**7**
Sarat	**13**	**13**

Retrieval of speech or music segments

In our method the stress **was** on giving a fast decision between speech and music **so** that either retrieval or further processing, **as** the case may be, can be performed. The speech segment retrieval **was** tried on pure speech data, and table **3** summarizes some of the results. The percentages in the third column indicate the portion of segments, which are correctly identified **as** speech.

Name of the Speaker	No. of Segments Analyzed	No. of Segments Correctly Retrieved
Mohanty	122	115
Nisha	130	109
Lalitha	**128**	92
Prakash	81	75
Sarat	**130**	**122**

Table **3:** Results of locating speech in the MPEG audio stream, with $1sec$ segments.

We made a system stream containing both speech and music segments, interleaved, and attempted to retrieve either music or speech at any given time. These were also correctly retrieved and played back.

Retrieval of a specific gender

The query placed is either Male ***or*** Female, and the system retrieves speakers of correct gender. Both $5sec$ and $2sec$ segments were used for analysis and the result for $5sec$ is summarized for some of the speakers in the database, in table **4.** $MorF$ against the speakers' **names** indicate whether male or female. Further the same video mentioned in the speaker location was used to retrieve a speaker of the desired gender.

Name of the Speaker	No. of Segments Analyzed	No of Segments Correctly Retrieved
Mohanty (M)	15	11
Madhavi (F)	17	14
Lalitha (F)	17	13
Prakash (M)	16	12
Sarat (M)	16	15

References

[1] Douglas A. Reynolds, Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models,"*IEEE Trans. Speech and Audio Processing,* vol. 3, no. 1, Jan. 1995.

[2] John D. Hoyt and Harry Wechsler,"Detection of Human Speech in Structured Noise", in *Proceedings of the ICASSP'94*, 1994.

[3] John Saunders, "Red-Time Discrimination of Broadcast Speech/Music", in *Proceedings* of *the ICASSP'96*, 1996.

[4] Eric Scheirer and Malcolm Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", in *Proceedings of the ICASSP'97*, 1997.

[5] Eluned Parris and Miachael J. Carrey, "Language Independent Gender Identification", in *Proceedings* of *the ICASSP'96*, 1996.

[6] Daniel W. Griffin and Jae S. Lim, "Multiband Excitation Vocoder", *IEEE Transactions on ASSP*, vol. 36, no. 8, Aug. 1988.

[7] Caterino Saraceno and Ricardo Leonardi, "Audio As a Support to Scene Change Detection and Characterization of Video Sequences", in *Proceedings of ICASSP'97*, 1997.

[8] Silvia Pfeiffer, Stephen Fischer and Wolfgang Effelsherg, "Automatic Audio Content Analysis", in *Proceedings* of *the ACM Conference on Multimedia,* 1996.

[9] Erling Wold, Thom Blum, Douglas Keisar and James Wheaton, "Content–Based Classification, Search and Retrieval of Audio", *IEEE Multimedia,* Fall 1996.

[10] Kenichi Minami, Akihito Akutsu, Hiroshi Hamada and Yoshinobu Tonomura: Video Handling with Music and Speech Detection, in *IEEE Signal Processing Magazine,* 1998.