An Experiment in Audio Classification from Compressed Data

Roman Jarina [*], Noel O'Connor [#], Noel Murphy [#], Seán Marlow [#]

*University of Žilina, Faculty of Electrical Engineering, Veľký diel, 010 26 Žilina, Slovakia
email: jarina@fel.utc.sk

[#] Dublin City University, Centre for Digital Video Processing, Dublin 9, Ireland
email: oconnorn@eeng.dcu.ie, URL: www.cdvp.dcu.ie

Abstract – **In this paper we present an algorithm for automatic classification of sound into speech, instrumental sound/ music and silence. The method is based on thresholding of features derived from the modulation envelope of the frequency limited audio signal. Four characteristics are examined for discrimination: the occurrence and duration of energy peaks, rhythmic content and the level of harmonic content. The proposed algorithm allows classification directly on MPEG-1 audio bitstreams. The performance of the classifier was evaluated on TRECVID test data. The test results are above-average among all TREC participants. The approaches adopted by other research groups participating in TREC are also discussed.**

Keywords: speech, music, MPEG, TREC, audio features

1. INTRODUCTION

Recent advances in digital audio/video coding and in digital storage technologies have contributed to the emergence of a large number of digital multimedia databases. Content-based video segmentation, indexing and retrieval have recently become active research topics due to the enormous amount of unstructured video data available nowadays. Although there has been a lot of work on video analysis, the work on audio analysis has been more limited. The segmentation and classification of audio content is of interest for a wide range of applications. Speech/music/ silence discrimination is the most common task in audio classification. For example, it can be applied to a multi-mode audio coder (such as MPEG-4 audio) to select the most appropriate coding scheme according to the type of the signal. Speech/ non-speech segmentation may improve ASR systems. Audio analysis together with video analysis is crucial for semantic content-based navigation and retrieval of video. This applies not just to the speech information, which clearly provides semantic information, but also to generic audio. Many different approaches to audio classification have been reported recently. Some of them use only a few features calculated in the time and/or the frequency domain, followed by a thresholding procedure [1-3]. Other approaches use more complicated features, several of which are motivated by perceptual properties of audio, and they apply more sophisticated procedures for classification including Gaussian mixture model (GMM), k-nearest neighbour (kNN) and Support vector machine (SVM) [4-9].

Although the majority of audio content is available in compressed form (e.g. MPEG) very little work has concentrated on audio classification directly in the compressed domain. Recently several approaches have been reported [2,10,11]. If the analysis is performed directly on compressed data, significant computation time can be saved. Audio analysis and indexing can be done in parallel with decoding, which is also important for streaming applications.

Unlike an ASR system, it is very difficult to evaluate the performance of an audio classification system due to the lack of a widely available labelled audio database to act as ground truth. To address this, our work is carried out in the context of the U.S. NIST (National Institute for Standards and Technology) TREC (Text REtrieval Conference) initiative to benchmark information retrieval systems. We participated in the feature extraction task of the video track of TREC, known as TRECVID 2002 [12]. Ten different audio/video features were specified as test

[*] This work was completed while the first author was with Dublin City University

features in TRECVID. In this paper we describe our work on speech feature and instrumental sound/music feature detection. The results are compared with results among all TREC participants.

2. TASK DESCRIPTION

The task specified by TREC was as follows. Given a standard set of video shot boundaries for the feature extraction test collection (see section 3.A), participants were to return for each feature in the list, the top 1000 video shots (max) from the standard set, ranked according to the highest possibility of the presence of the feature. If the feature was detected for some sequence within the shot, then it was considered detected for the entire shot. This latter simplification was adopted in order to allow pooling of results and calculation of precision and recall. We examined detection of two audio features, namely speech and instrumental sound/music.

3. EVALUATION SETUP

A. Database

The video data (accompanied by audio) consisted of MPEG-1/VCD recordings from the Internet Archive and the Open Video Project. Within this corpus, different subsets were defined as the development sets and test sets. Ninety six videos (23.26 hours) randomly chosen from the total available data formed the feature development set (training database). The feature extraction test collection consisted of 5.02 hours of from twenty three different video. The video collections were accompanied by reference sets of video shot boundaries (7891 standard shots for the development set, and 1848 standard shots for the test set). Joint audio was stored in the MPEG-1 Layer II format (MP2). The audio component of the feature development test collections was not labelled. This means that ground truth information for the content (i.e. manually annotated speech, music, silence decisions for each shot) was not available. The feature extraction test collection was labelled at the video shot level. Shots containing given features were determined by NIST assessors (using pools of shots submitted by participants). A shot contains a feature if at least one frame within the shot matches the feature's description, and otherwise does not contain the feature. A more detailed description of the TRECVID database can be found in [12].

B. Evaluation metrics

A feature extraction run consisted of a ranked list of up to 1000 shots ordered by the likelihood that the shot contains the feature. Runs were evaluated using precision and recall, as well as uninterpolated average precision. Measures were computed for speech and music feature individually. Since the number of "true" shots (i.e. shots containing the feature) in the feature extraction test set exceeded 1000 (1382 true shots for speech and 1221 true shots for instrumental sound), an artificial upper bound on possible average precision was 0.724 for speech and 0.819 for music [12].

4. EXPERIMENTS

Our audio feature detector does not use an audio signal waveform as the input data, rather it utilises information taken directly from an MP2 audio encoded bitstream. The method is based on thresholding of low-level features derived from the modulation envelope of the frequency limited audio signal. Four characteristics are examined for discrimination between speech and instrumental sound/ music. They are occurrence and duration of energy peaks, rhythmic content and level of harmonic content. The volume contour of the signal in each of the Layer II subbands is estimated from the scalefactors [15]. By definition the scalefactors carry information about the maximum level of the signal in each subband.

The procedure for audio classification is depicted in Fig.1. Only the scalefactors from 7 of the low frequency subbands and coded samples from the 2nd subband were included in the processing. First, silence detection was carried out. An energy level of the signal was determined by the superposition of all relevant scalefactors. The frames in which the level was below the threshold, were assigned as silent frames. For further analysis a sliding window was used with a window length of 3.9 seconds (150 frames of the MP2 bitstream) and a 1.3 seconds (50 frames) shift.

A. Speech Detection

The envelope of the band-limited signal was estimated by summing relevant scalefactors from the 2nd to the 7th subbands only. This procedure was followed by the 5^{th} order median filtering to avoid rapid random changes in the amplitude. Due to the regular syllabic structure of speech., strong temporal variations in the amplitude of speech signals are observed.

Energy peaks were extracted by a simple thresholding procedure. The following two descriptors were chosen for speech detection: the duration of the widest peak L, and the number of peaks R in the analysis window. Each segment was assigned to speech or non-speech by using a simple rule-based decision procedure defined as follows [13].

If $L_m < t_L$ & $t_{R1} < R < t_{R2}$ then the frame is considered as speech where t_L ,t_{R1}, t_{R2} are empirically chosen thresholds, the relevance REL_S is set to one, otherwise REL_S=0.

We have derived values of the thresholds in [13] where an importance of various MPEG frequency subbands for speech/music discrimination is also discussed.

B. Music Detection

Unlike speech, musical sounds are very difficult defined due to their great variety and uncertain nature. But, musical signals have some unique characteristics, which may help to discriminate them from other sounds. Music

tends to be composed of a multiplicity of tones, each with an own distribution of higher harmonics. The energy contour has usually a much smaller number of "peaks" and "valleys" and it shows either very little change over a period of several seconds (e.g. classical music) or strong long term periodicity due to exact rhythm (e.g. dance music).

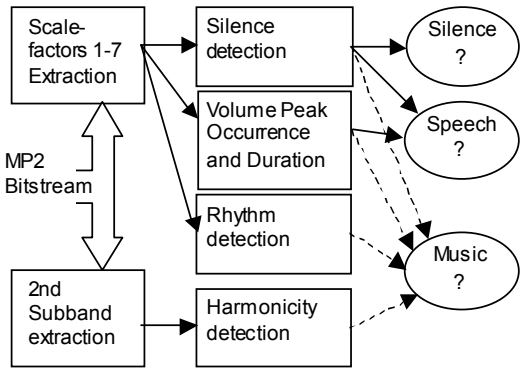

Fig. 1: Proposed procedure for audio classification

The descriptors *L* and *R* were also used for music detection. In addition, two other audio descriptors, corresponding to rhythm and harmonicity, were incorporated into the system to improve discrimination between musical sounds and other environmental sounds. We believe that the spectra of most sounds produced by instrumental music exhibit a harmonic structure, unlike noise-like environmental sounds. Rhythmic pulses are detected by applying a long-term autocorrelation on modulation envelope (subband energy contour) derived for each of the first 7 subbands. We searched the autocorrelation functions over the interval corresponding to time t = 0.2 - 1.75 seconds to find peaks. If $p(j)$ is the value of the highest peak in the j-th subband, the rhythm metric P is defined as follows

$$P = \max\{p(1), p(2), ..., p(7)\}, \quad (1)$$

The higher the value of P, the greater amount of rhythmicality in the signal. From the previous experiments [14] we hypothesize that if $P > 0.6$, the signal is not speech ($P < 0.6$ for 99.73% of the speech signal being examined in [14]).

Further we computed the *harmonicity ratio* similarly as it is defined in the MPEG-7 description schema [16]. At first, comb filtering was applied as follows

$$r(k) = \sum_j s(j)s(j-k) \Big/ \left(\sum_j s(j)^2 \cdot \sum_j s(j-k)^2 \right)^{0.5} \quad (2)$$

where s is sequence of coded samples of the band-limited signal. To speed-up the processing, we used only the 2nd MPEG frequency subband for this computation. Index k was changed up to the value corresponding to the maximum expected fundamental period (around 20 ms). The harmonicity ratio H was determined as the maximum value of $r(k)$ for each frame. H = 1 for a purely periodic signal, and it will be close to 0 for white noise. We used simple rule-based classification procedure as follows:

If $L_m < t_L$ & ($t_{R1} > R$ or $R > t_{R2}$) & $P > t_P$ & $H > t_H$ then the signal is considered as music and $REL_M = 1$, otherwise $REL_M = 0$. The thresholds were set at $t_L = 0.7$s, $t_{R1} = 2.5s^{-1}$, $t_{R2} = 5.5s^{-1}$, $t_P = 0.6$, and $t_H = 0.8$.

Final speech and music feature measures for the standard video shots were determined by averaging the relevance scores REL_S and REL_M over all the audio frames corresponding to a given video shot.

C. Comparison of results among TREC participants

The following is a list of some of the groups that took part in the audio feature detection tasks and a brief explanation of the approaches adopted.

CLIPS-IMAG Grenoble [5] computed 16 Mel-Frequency Cepstral Coefficients (MFCC) and log energy on 20 ms signal windows. Gaussian Mixture Models (GMM) were then applied to characterize speech and non-speech. The length of detected speech segments within a shot was used for ranking the results. A research group from *Fudan University* [6] applied audio analysis on a 1-second window. They used many features derived from zero-crossing rate (ZCR), short-time energy, LPC and MFCC coefficients. A Nearest Neighbour Model and GMM were trained on the TRECVID corpus.

IBM Research group [7] fused several methods for statistical modelling including SVMs and GMMs. For the models they used 24 MFCC together with video/image descriptors. The speech/music discriminator of the *MediaMill* group [3] is based on amplitude variation of frequency limited signal in two frequency bands. When the amplitude variation in either of the two spectral bands was above a threshold, the segment was identified as speech, otherwise as music.

Microsoft Research Asia [8] employed audio descriptors derived from ZCR, short-time energy together with linear spectral pairs distance and band periodicity. A SVM-based classifier was applied to classify the audio stream. The approach of the *MediaTeam Oulu* [9] is based on kNN classification. They used 3 energy-based features derived in the time domain for a 3-second analysis window.

Ten of thirteen groups submitted 13 runs for speech detection. Nine groups submitted 11 runs for instrumental sound detection [12]. The best five results for each task are summarized in Table 1 and Table 2 respectively. The average value is computed as the median of all the submitted results. Our results are referred as DCU. Of the participants listed in Table 1 and Table 2, we are the only one who performed audio classification entirely in compressed domain. For both features we obtained the highest precision among TREC participants for 100 results. According to average precision, our run is scored as the 3rd and the 4th for speech and instrumental sound detection tasks respectively. The results marked with * in Table 2 are unofficial (post-submission results). Since we identified many more relevant shots than we submitted for

judgment (originally only 300), we have re-calculated precision and average precision for our 1,000 top ranked shots.

From the results reported by the TREC participants, we can notice that the best speech detectors (CLIPS [5], IBM [7]) utilize only MFCCs as inputs, and the best musical sound detectors (MT Oulu [9], Fudan Univ.[6]) utilize either time domain energy-based descriptors or combination of many features.

Table 1: Top five speech feature detection results

Group	Average Precision	Precision at 100 results	Precision at 1000 results
CLIPS	0.721	1.000	0.997
IBM	0.713	0.990	0.990
DCU	**0.710**	**1.000**	**0.987**
MediaMill	0.681	0.960	0.970
Fudan Univ.	0.663	0.980	0.951
AVERAGE	0.656	0.980	0.944

Table 2. Top five instrumental sound feature detection results

Group	Average Precision	Precision at 100 results	Precision at 1000 results
MT Oulu	0.637	0.840	0.877
Fudan Univ.	0.564	0.850	0.799
MSRAsia	0.511	0.900	0.709
DCU	**0.494***	**0.970**	**0.650***
MediaMill	0.438	0.920	0.716
AVERAGE	0.347	0.845	0.667

5. CONCLUSION

We have proposed a very fast audio classification algorithm for MPEG bitstreams. For classification of speech, music and silent segments only a very small portion of the MPEG 1 layer II bitstream is required. The performance of the system was evaluated on the TREC 2002 Video track test collection. Results are compared with the results of other TREC Video track participants. The performance of the system for both speech feature detection and instrumental sound/music detection tasks is above average. Particularly the speech detector performs very well. We reached the precision 98.7 % for 1000 top ranked shots. Advantages of the proposed method are very fast processing and simple implementation. Thus the method is suitable for streaming and real-time applications.

ACKNOWLEDGMENTS

This research is supported by the Research Institute for Networks and Communications Engineering (RINCE) at Dublin City University, and EU Marie Curie Development Host funding.

REFERENCES

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. ICASSP'96*, vol. II, Atlanta, GA, May 1996, pp. 993–996.

[2] Y. Nakajima, et al., "A Fast Audio Classification from MPEG Coded Data", in *Proc. ICASSP'99*, Vol. 6, Phoenix, Arizona, May 1999.

[3] J. Vendrig, et al., "TREC Feature Extraction by Active Learning", in *Proc. Text Retrieval Conference TREC 2002,* Gaithersburg, Maryland, 19-22 Nov. 2002.

[4] L. Lu, H-J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation", *IEEE Trans. on Speech and Audio Processing*,, Vol.10, No. 7, Oct. 2002, pp.504-516.

[5] G.M. Quénot, et al., "CLIPS at TREC-11: Experiments in Video Retrieval", in *Proc. TREC 2002,* Gaithersburg, Maryland, 19-22 Nov. 2002.

[6] L. Wu, et al., "FDU at TREC2002: Filtering, Q&A, Web and Video Task", in *Proc. TREC 2002,* Gaithersburg, Maryland, 19-22 Nov. 2002.

[7] B. Adams, et al., "IBM Research TREC-2002 Video Retrieval System", in *Proc. TREC 2002,* Gaithersburg, Maryland, 19-22 Nov. 2002.

[8] X. -S. Hua, et al., "MSR-Asia at TREC-11 Video Track", in *Proc. TREC 2002,* Gaithersburg, Maryland, 19-22 Nov. 2002.

[9] M. Rautiainen, "TREC 2002 Video Track Experiments at Media Team Oulu and VTT", in *Proc. TREC 2002,* Gaithersburg, Maryland, 19-22 Nov. 2002.

[10] G. Tzanetakis, and P. Cook, "Sound Analysis Using MPEG Compressed Audio" , in *Proc. ICASSP'2000*, Istanbul, Turkey, June 2000.

[11] S. Kiranyaz, M. Aubazac, M. Gabbouj, "Unsupervised Segmentation and Classification over MP3 and AAC Audio Bitstreams", in *Proc. of Europ. Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 2003*, London, UK, April 2003.

[12] TREC2002 Video track. Available online at URL: http://www-nlpir.nist.gov/projects/t01v/t01v.html

[13] R. Jarina, N. Murphy, N. O'Connor, and S. Marlow, "Speech-Music Discrimination from MPEG-1 Bitstream", In *Advances in Signal Processing, Robotics and Communications*, V.V. Kluev, N.E. Mastorakis (editors), WSES Press, pp. 174-178, 2001.

[14] R. Jarina, R., N. O'Connor, S. Marlow, and N. Murphy, "Rhythm Detection for Speech-Music Discrimination in MPEG Compressed Domain", in *Proc. of IEEE Int. Conf. on Digital Signal Processing DSP'02*, Santorini, Greece, pp. 129-132, 1-3 July 2002.

[15] ISO/IEC 11172-3, "Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 Mbit/s, Part 3: Audio", 1992.

[16] ISO/IEC JTC 1/SC 29/WG 11, "Information Technology – Multimedia Content Description Interface – Part 4", Audio, March 2002.