Speech-Music Discrimination from MPEG-1 Bitstream

ROMAN JARINA, NOEL MURPHY, NOEL O’CONNOR, SEÁN MARLOW
Centre for Digital Video Processing / RINCE
Dublin City University, Dublin 9
IRELAND
jarinar@eeng.dcu.ie http://www.eeng.dcu.ie

Abstract: This paper describes a proposed algorithm for speech/music discrimination, which works on data directly taken from MPEG encoded bitstream thus avoiding the computationally difficult decoding-encoding process. The method is based on thresholding of features derived from the modulation envelope of the frequency-limited audio signal. The discriminator is tested on more than 2 hours of audio data, which contain clean and noisy speech from several speakers and a variety of music content. The discriminator is able to work in real time and despite its simplicity, results are very promising.

Key-Words: audio, video, classification, speech, music, signal processing, MPEG

1 Introduction

The work described in this paper under the auspices of the Centre for Digital Video Processing (CDVP) in cooperation with the Research Institute for Networks and Communications Engineering (RINCE) at Dublin City University. The CDVP research work concentrates on digital video analysis and multimedia content management. The current results are demonstrated on our Web-based digital video system called *Físchlár* [1],[2]. *Físchlár* is able to record selected broadcast television programmes into MPEG-1 format and process recorded data to enable convenient browsing and playback of all of the video content.

Until recently, research in the centre has been oriented primarily to the visual aspect of video multimedia content. We decided to improve the efficiency and utility of the demonstration system by also incorporating audio analysis, which is an important supplementary source of information. It has been observed that audio and speech analysis can sometimes be more effective in determining the contents of audio/visual material than image analysis and recognition.

Generally, while there has been a lot of research done on image signal processing for video management, very little work yet has been done on the audio portion of the multimedia stream. Research on effective management of audio content is expected to gain more attention.

One of our aims in the area of audio analysis is audio signal segmentation and classification. It is highly desirable that the tools being developed would work in the compressed domain, i.e. use data directly from the MPEG encoded bitstream to avoid the computationally difficult decoding-encoding process. This paper describes a proposed algorithm for speech/music discrimination.

2 MPEG-1 Audio Coding

The Moving Pictures Expert Group completed its first standard for video and audio, called MPEG-1, in November 1992. The MPEG-1 audio coding system, specified in ISO/IEC 11172-3 operates in single-channel or two-channel stereo modes at sampling frequencies of 32, 41.1, and 48 kHz, and is well documented [3],[4]. This following paragraph gives a brief explanation and description of the data, which form an MPEG-1 audio encoded bitstream, and may be used for audio analysis.

MPEG-1 audio uses a high performance perceptual coding scheme to compress the audio signal. This means it achieves higher compression by removing acoustically irrelevant parts of the audio signal (those not perceived by the human ear). The standard defines three layers with increasing computational complexity and better compression and performance in the higher layers.

The *Físchlár* system captures TV audio signal and encodes it according to the MPEG-1 Layer-II compression scheme (MP2), which provides very high quality at a data rate of 128 kbit/s per channel. In this system, signals sampled at 44.1 kHz are first divided into 32 uniformly spaced frequency subbands. The signals in each subband are then individually quantized and assigned a bit-allocation according to masking properties of the signal in the given subband (i.e. quantisation noise has to be inaudible). A layer-II frame consists of 1152 samples: 3 groups of 12 samples from each of 32 subbands. A group of 12 samples in each subband gets a bit allocation and, if this is not zero, a scalefactor. Scalefactors are weights that rescale samples so that they fully use the range of the quantiser. The encoder uses a different scalefactor for each of the three groups of 12 samples only if necessary. The quantized signal data, bit allocation information, scalefactors, supplemented with a header, CRC code and ancillary data, form the encoded bitstream.

3 Audio Signal Characteristics

3.1 Speech Signals

There are some properties that distinguish speech from other signals. Roughly, speech exhibits an alternating sequence of 3 kinds of sounds that have different acoustical properties:

1) *Vowels and vowel-like sounds* – longer tonal quasi-periodic segments with high energy, which are concentrated in lower frequencies. The short-term power spectrum always has “peaks” and “valleys”. These peaks correspond to the formants (i.e. resonances) of the vocal tract. The fundamental frequency of the signal corresponds to the pitch of the human voice. The range usually falls between 40-250 Hz for male voices and 120-500 Hz for female voices. The duration of vowels is variable. It depends on speaking style, and usually it is around 100-200 milliseconds long.

2) *Fricative consonants* – noise-like short segments with lower volume. Spectral energy is distributed more toward the high frequencies. The duration of consonants are usually only tens of milliseconds long.

3) *Stop consonants* – short silent segments followed by a very short transition noise pulse.

These three kinds of sounds alternate and form the regular syllabic structure of speech. Therefore, strong temporal variations in the amplitude of speech signals are observed. An average modulation frequency is 4 Hz (it corresponds with the syllabic rate). In addition, short-term spectrum changes over time are observed. Some changes occur very rapidly, such as the spectral transition of stop consonants, whereas other changes are more gradual, like formant motions of semivowels or diphthongs.

3.2 Musical Signals

On the other hand, musical signals have the following unique characteristics:

– Music tends to be composed of a multiplicity of tones, each with an own distribution of higher harmonics
– Usually there are small changes in the short-term spectral envelope over time. On the other hand, fundamental frequencies of tones can change rapidly and over much wider scale than in the case of speech signals.
– The energy contour or envelope 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).

4 Speech/Music Discrimination

Speech/music discrimination is one of the basic problem in audio segmentation and classification. Several different approaches have been reported. Some of them use only a few features derived in the time and/or the frequency domain, followed by a thresholding procedure [5],[6],[7]. Zero-crossing rate, short-time energy and fundamental frequency are the most common features. Another approache uses many more-complicated features, several of which are motivated by perceptual properties of audio, and they apply complicated sophisticated procedures for classification [8]. In [9], ratio of signals obtained by band filtering is used, and a fuzzy feature combiner estimates the probability that the input signal is speech. Artifitial neural networks [10] and Hidden Markov models [11] have also been examined for classification.

4.1 Proposed Approach

Our approach is based on properties of signals in the time domain. Features are derived from the modulation envelope of the frequency-limited audio signal. As mentioned above, *Físchlár* digitizes and stores audio signal in MPEG-1 Layer II format. The aim is to utilize signal analysis that has been done during the MPEG compression process and thus to utilize information that is already stored in the MPEG bitstream.

By definition the scalefactors in the MPEG-1 encoded bitstream carry information about the maximum level of the signal in each subband. In [12], the scalefactors have been used for silence detection for commercial breaks determination in broadcast television programmes. Thus, a time envelope or contour of overall audio signal can be estimated by summation of scalefactors over all subbands. In the case of a frequency-limited signal, summation is done only over given subbands. Time resolution (in the case of sampling rate 44.1 kHz) is 8.7 ms (= 32*12/ 44100 Hz) and cutoff frequency is a multiple of subband width B_S, which is approx. 690 Hz (B_S = 22050 Hz /32). The scalefactors are stored with 6-bit precision so there are 64 different values, which are taken from a look-up table [4].

We measured the time length and rate of high-volume segments or peaks of speech and music. The length of the largest peak and average rate of peaks within a given time frame were selected as features for the classifier. Analysis was done on a frame-by-frame basis with frame length of 4 seconds and 50% overlap.

First, we made several preliminary experiments to find an appropriate threshold for peak edge determination. From histograms, we have found that the values of these features (the lengths of the largest peaks and rates of peaks) have a different distribution for speech and music and could be quite well separable (Fig. 1). The values for speech signals are mostly well bounded. The values of the rate are clustered around 4 frames per sec, which correlates with the average modulation frequency of speech, which is 4 Hz.

We also investigated which subbands are most significant for satisfactory discrimination of speech and musical signals. From these experiments, we found that just lower and middle frequency bands are important. We also assume that by skipping the first subband, the efficiency of the method can be improved. There are two arguments for this: (i) background noise in the presence of speech, which contains mostly low frequencies, can be suppressed; (ii) it reduces sharp energy peaks from instruments like bass or drums in rhythmic music (pop, rock, dance), which often have a similar rate to the syllable rate of speech, and could cause a wrong decision.

By listening tests we confirmed that even though the audio signals were limited to the frequency band 700–4000 Hz, it was easy to distinguish if it was speech or music.

4.2 Feature Extraction Procedure

The scalefactors corresponding to the 2nd-7th subbands of the encoded audio signal are stripped from the bitstream. This corresponds to frequencies from 690 to 4800 Hz. The time envelope $e(n)$ is computed by summation of scalefactors (SCF) in each group followed by smoothing. As mentioned above, the length of the analysed frame is 4 seconds, which corresponding to 459 samples of $e(n)$.

$$\hat{e}(n) = \sum_{i=2}^{7} SCF(n,i) \qquad (1)$$

$$e(n) = \sum_{j=-2}^{2} \hat{e}(n+j) \qquad (2)$$

If the mean value $E\{e(n)\} < 0.003$, the segment is considered as a silence and no further analysis is performed, othervise the threshold value for peak selection is set as

$$H = k \cdot E\{e(n)\}, \text{ where } k = 0.4 \qquad (3)$$

The parts where the envelope curve is above the threshold H are considered as peaks (high-volume segments). This procedure is also explained in the Fig. 2. For a better picture resolution , only a half of analysed frame (i.e. 2 sec.) is depicted.

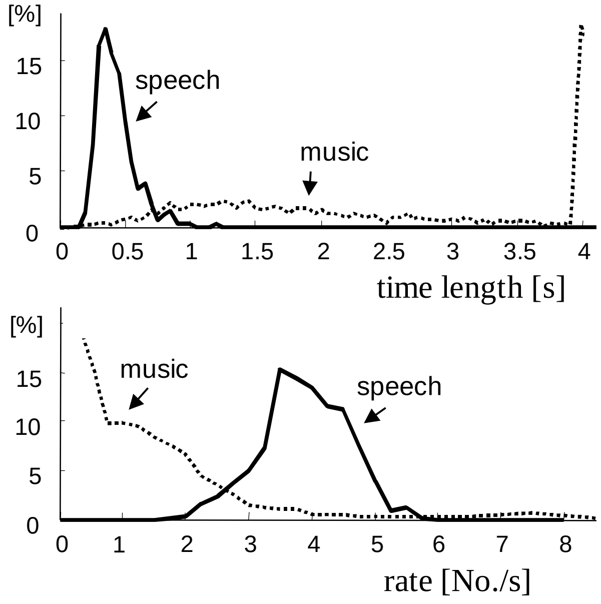

Figure 1. Distribution of values of features L_m and R for speech and musical signals. Modulation envelope is computed from scalefactors for subbands 2-7.

For each frame, the length of the largest peaks L_m and rate of peaks R (number of peaks per second) are computed.

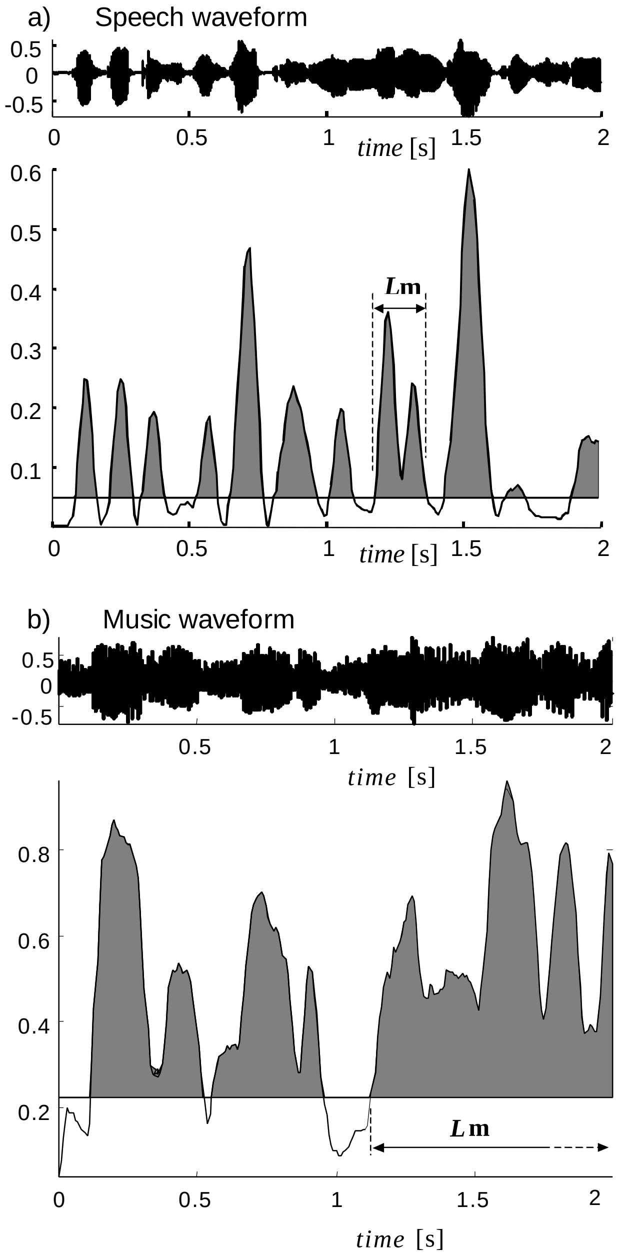

Figure 2. Procedure of feature extraction for a) speech signal; b) music signal ("rock"). Modulation envelope is computed from scale-factors for subbands 2-7.

4.3 Classification

As can be seen from the feature distribution for speech and music depicted in Fig. 1, a simple thresholding method can be used for classification. The thresholds were set at

$$H_L = 0.7\,s,\ H_{R1} = 2.5\,s^{-1},\ H_{R2} = 5.5\,s^{-1}.$$

If $L_\mathrm{m} < H_L$ and $H_{R1} < R < H_{R2}$ then the frame is considered as speech. Otherwise it is considered as music.

5 Experiment

5.1 Test Audio Data

We collected approx. 40 min. of speech from Irish radio and television (RTE) news programs. The first part contains only clean speech (i.e. anchor person, indoor interview). The second part contains clean speech and also speech with high background noise (outdoor reporting, traffic noise, blowing wind, background voices etc.).

The musical recordings (about 1.5 hour) are obtained from several sources (including television broadcast and Internet mp3 files). They contain a variety of instrumental and vocal music (classical, rock, pop, dance, jazz). The music database is divided into three groups. The recordings are stored in PCM and MPEG-1 Layer-II formats. The database is summarised in Table 1.

Table 1: Audio database description.

Name	Description	Duration
s1	clean speech	00:19:10
s2	clean and noisy speech	00:18:32
classic	instrumental music, loose tempo (classic, jazz)	00:43:18
rhythm	instrumental music, strong rhythm (rock, pop, jazz, dance)	00:51:00
vocal	vocal music, songs (classic, rock, pop, rap)	00:53:55

5.2 Experimental Results

We examined the proposed algorithm for three different values of peak threshold (Eq.(3), k=0.3, 0.4, or 0.5). We evaluate each category of speech and music separately. Results are summarised in Table 2.

It seems that the algorithm works best with a peak threshold H (Eq.(3)), where k=0.4.

In the first case (k=0.3 in Table 2), threshold is sometimes below noise level so peaks cannot be separated from noisy speech signals (s2 – only 67.15% correct).

Table 2: Correct recognition rate of the speech/music discriminator.

Name	Frames	Correct recognition %		
		k=0.3	k=0.4	k=0.5
s1	573	94.24	96.34	95.64
s2	554	67.15	83.39	84.84
classic	1279	99.22	97.03	96.40
rhythm	1514	92.54	84.54	82.17
vocal	1598	97.56	93.18	88.99

Performance can be improved by post-processing that corrects single errors (one single speech frame wouldn't normally be between music frames and vice versa). The results for the second case (Eq.(3), k=0.4) are in Table 3. As can be seen, an increase of correct recognition rate after post-processing is about 4% for speech and 1-2% for music.

Table 3: Correct recognition rate of the discriminator with post-processing (k=0.4).

Correct recognition %		
Name	Post-processing	
	NO	YES
s1	96.34	98.08
s2	83.39	87.73
classic	97.03	97.89
rhythm	84.54	86.13
vocal	93.18	94.81

6 Conclusion

It has been shown that simple methods for speech/music discrimination can give satisfactory results for many of the audio signals tested. Since the discriminator works on data taken directly from an MPEG-1 encoded bitstream, the computationally difficult decoding-encoding process is not required. Thus, advantages of this approach are high computational speed (the discriminator is able to work in real-time) and easy implementation.

Future work will be concentrated to two directions: 1) refining discrimination process (speech/music/other discrimination, music classification) by utilizing more information that is contained in the MPEG bitstream; 2) combination of audio and visual analysis.

References:

[1] N.E. O'Connor et al., Físchlár: An On-line System for Indexing and Browsing of Broadcast Television Content, appears in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP'01*, Salt Lake City, UT, 7-11 May 2001.

[2] *Centre for Digital Video Processing /Físchlár Website*, http://www.fischlar.com

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

[4] K. Brandenburg and G. Stoll, The ISO/MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio, *J. Audio Eng. Soc.*, Vol.42, No.10, Oct. 1994, pp. 780-792.

[5] T. Zhang and C.-C. J. Kuo, Content-Based Classification and Retrieval of Audio, *Proc. SPIE on Advance Signal Processing, Algorithms, Architectures and Implementations VIII*, Vol.3461, San Diego, July 1998, pp.432-443.

[6] J. Saunders, Real-Time Discrimination of Broadcast Speech/Music, *Proc ICASSP'96*, Vol.II, Atlanta, May 1996,pp. 993-996.

[7] N. Patel and I. Sethi, Audio Characterization for Video Indexing, *Proc. SPIE in Storage and Retrieval for Still Image and Video Databases*, Vol.2670, San Jose, 1996, pp. 373-384.

[8] E. Schneider and M. Slaney, Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator, *Proc. ICASSP'97*, Vol.II, Munich, Germany, April 1997, pp. 1331-1334.

[9] R. M. Aarts and R.T. Dekkers, A Real-Time Speech-Music Discriminator, *J. Audio Eng. Soc.*, Vol.47, No.9, Sept. 1999, pp. 720-725.

[10] Z. Liu, J. Huang, Y. Wang and T. Chen, Audio Feature Extraction and Analysis for Scene Classification, *Electr. Proc. of IEEE Workshop on Multimedia Signal Processing*, Princeton, NJ, June 1997, pp.1-6.

[11] T. Zhang and C.-C. J. Kuo, Hierarchical classification of audio data for archiving and retrieving, *Proc. ICASSP'99*, Vol. 6, Phoenix, Mar. 1999. pp. 3001-3004.

[12] D. Sadlier, S. Marlow, N. O'Connor and N. Murphy, Automatic TV Advertisement Detection from MPEG Bitstream, appears in *Proc. of Int. Conf. on Enterprise Information Systems ICEIS 2001*, Setubal, Portugal, 7-10 July 2001.