RHYTHM DETECTION FOR SPEECH-MUSIC DISCRIMINATION IN MPEG COMPRESSED DOMAIN

Roman Jarina, Noel O'Connor, Seán Marlow, and Noel Murphy*

Centre for Digital Video Processing
Dublin City University, Dublin 9, IRELAND
jarinar@eeng.dcu.ie, oconnorn@eeng.dcu.ie

Abstract: A novel approach to speech-music discrimination based on rhythm (or beat) detection is introduced. Rhythmic pulses are detected by applying a long-term autocorrelation method on band-passed signals. This approach is combined with another, in which the features describe the energy peaks of the signal. The discriminator uses just three features that are computed from data directly taken from an MPEG-1 bitstream. The discriminator was tested on more than 3 hours of audio data. Average recognition rate is 97.7%.

1. INTRODUCTION

Audio content classification is a very important task for the browsing, indexing and retrieval of audio/video databases. The development of standards for high-quality audio and video compression such as the family of MPEG standards, coupled with increased computing performance enables easy recording, storage and manipulation of multimedia content. The aim of the research presented in this paper is to extend the capabilities of the web-based digital video system developed by the Centre for Digital Video Processing at Dublin City University [1], [2].

The problem of distinguishing speech signals from music signals is a first step in audio segmentation and classification. Several different approaches to speech/music discrimination 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 [3], [4]. Zero-crossing rate (ZCR), short-time energy, and fundamental frequency are the most commonly used features. Other approaches use many more complicated features, several of which are motivated by perceptual properties of audio, and they apply more sophisticated procedures for classification (e.g. Gaussian mixture model, k-nearest neighbour) [5], [6]. Hidden Markov models (HMM) [7] and Artificial Neural Networks [4], [8] have also been investigated for classification.

For example, Bugatti et al. [4] compared two different algorithms. They used 30 minutes of labelled audio signal for their experiments, and obtained a performance of 82% using the simple method (ZCR + Bayesian Classifier), and 94% for ANN classifier (8 features + MLP). Scheirer and Slaney [5] evaluated 13 features using various classification frameworks (multi-dimensional Gaussian MAP estimator, GMM, k-d tree, and k-NN). They used 40 minutes of audio for their experiments and they obtained similar results for all the methods. The best classifier achieved 94.2% performance on a frame-by-frame basis, and 98.6% when integrating 2.4 seconds segments of sound.

All the above approaches [3]-[8] have been conducted for PCM data. In [9] [10], audio classification on the MPEG-1 subband domain has been proposed. Both used short–term energy based features, which were computed from subband encoded audio samples. This approach has the following advantages:

- It can deal with long audio and video files, which are mostly stored in MPEG 1 format [11];
- It has small storage and computational requirements, because the computationally difficult decoding process is not required;
- The audio signal analysis carried out during the encoding process can be utilized (e.g. subband filtering, volume estimation).

In this paper, we introduce the new feature *Rhythm metric* that quantifies the strength of rhythm in audio signals. We believe that this kind of feature could be very useful for the speech-music discrimination task. Results presented in this paper prove this hypothesis and show that the rhythm metric has good discriminative properties. The proposed discriminator processes the signal in the MPEG-1 compressed domain.

2. SPEECH-MUSIC DISCRIMINATION

In [12], we presented a simple method for distinguishing speech signals from music signals. The method is based on the observation of the modulation envelope (or contour) of the band-limited signal. From the envelope, high-volume peaks are extracted. The width of the widest peak and average rate of peaks within a time interval of 4 seconds are chosen as features for the discriminator.

* also with the University of Žilina, Slovakia

0-7803-7503-3/02/$17.00 ©2002 IEEE

This approach gives satisfactory results for many of the audio signals tested. However, poor results were obtained for music signals with strong rhythms, where the average rate of peaks was often confused with the syllabic rate of speech, which is around 4 Hz. The same problem arose in [10] where the authors reported only 88% correct detection of music signals due to the fact that segments with intermittent sound (e.g. drums) were often classified as speech. Therefore, we decided to improve the performance of the discriminator by incorporating rhythm detection.

The proposed discriminator does not use an audio signal waveform as the input data, rather it utilizes information taken directly from an MPEG-1 encoded bitstream. Unlike the methods introduced in [9] [10], we use only scalefactors for audio signal analysis. Advantages of using scalefactors instead of coded samples are as follow:

- It is much easier to find the position of the scalefactors in the MPEG frame and decode them than to find and decode audio samples.
- The scalefactors are a very small part of the MPEG bitstream.
- Detecting and processing the scalefactors is very straightforward and fast.

3. RHYTHM DETECTION AND FEATURE EXTRACTION

The approach presented in this paper is based on the assumption that the majority of musical signals have periodic regular structure or "rhythm". Therefore, the strength of rhythm could be a good discriminative feature. A similar kind of feature (called *Pulse metric*) was also used in [5]. We use the expression "rhythm detection" in the sense of the automatic extraction of rhythmic pulses from an audio signal. Other more complex approaches to beat and rhythm analysis for musical signals are presented in [13], [14].

Our proposed approach is based on the autocorrelation method. An example of rhythm analysis based on long-term autocorrelation of band-passed signals is shown in Figure 1. As can be seen from Figure 1a, musical signals have periodic structure in some frequency subbands. The music character determines which subband exhibits periodicity. Beats are mainly detected in low frequencies (bass instruments) or high frequencies (percussions). Speech signals have no long-time periodic structure (Figure 1b).

3.1 Feature extraction procedure

Tested audio signals are sampled with sample frequency f_s= 44100 Hz, and stored in MPEG-1 Layer II audio format (MP2) [11]. 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 receives 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 only uses a different scalefactor for each of the three groups of 12 samples if it is necessary. Subbands are uniformly spaced and their width is approx. 690 Hz. By definition, the scalefactors in the MPEG-1 encoded bitstream carry information about the maximum level of the signal in each subband.

We estimate a time envelope or contour of the frequency-limited audio signal from the scalefactors for each subband. For analysis, a sliding window is used with a window length of 4 seconds and 50% overlap. The window length corresponds to 460 scalefactors in each subband (time resolution is $32*12/f_s$= 8.7 ms). Only the first 15 subbands are used for computation (up to 10 kHz).

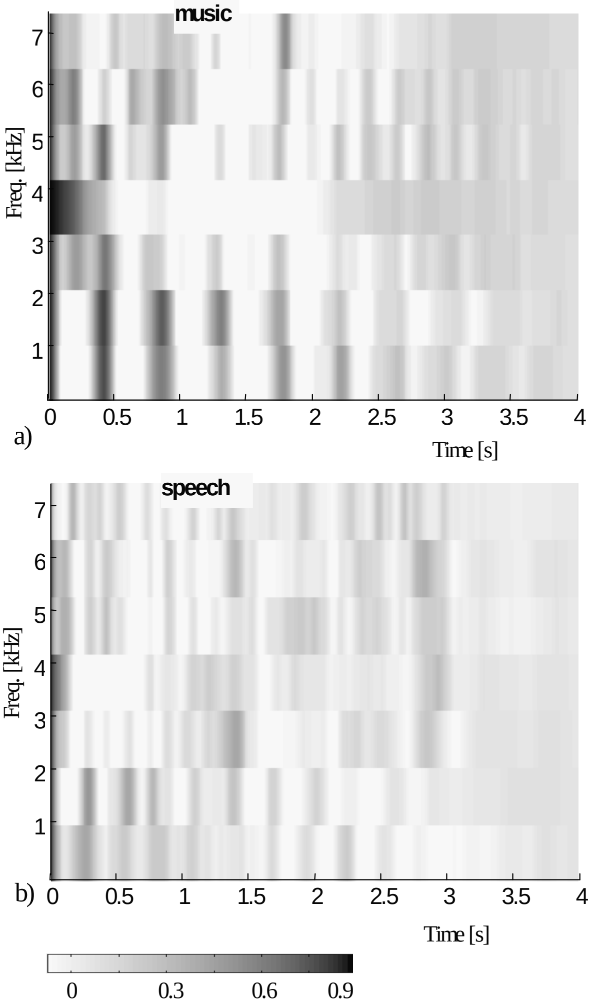

Figure 1. Examples of long-term autocorrelation of band-passed audio signals

The mean value of the set of scalefactors $s(t,k)$, where $s(t,k)$ is a value of the t-th scalefactor in the k-th subband, is calculated as:

$$E = \frac{1}{460} \sum_{t=1}^{160} \sum_{k=1}^{15} s(t,k)\,, \qquad (1)$$

If E is less than determined threshold, the frame is considered as a silence and no further analysis is performed, otherwise the envelope curve for each subband is smoothed by following method:

$$e(t,k) = \sum_{j=-2}^{2} s(t+j,k), \quad k = 1,2,...15 \tag{2}$$

Since such a high frequency resolution is not necessary in the higher subbands, the subbands 11-15 are merged together by summing the relevant scalefactors.

For each subband, the normalized autocorrelation function R_n is computed.

$$R(t,k) = \sum_{n=1}^{460-t} \hat{e}(n,k) \cdot \hat{e}(n+t,k), \;\; t = 0,1,...200 \tag{3}$$

$$R_n(t,k) = R(t,k) / R(0,k) \tag{4}$$

where $\hat{e}$ means envelope *e* without its DC component.

We search R_n over the interval corresponding to time 0.2s - 1.75s to find peaks. We set *p*(*k*) to the value of the highest peak in the *k*-th subband. Then we define the feature *rhythm metric* $\boldsymbol{P_m}$ as

$$P_m = \max\{p(k)\}, \;\; k = 1,2,...11 \tag{5}$$

where $0<P_m<1$. The higher the value of P_m , the greater amount of rhythmicality in the signal.

4. EXPERIMENT

4.1 Test Audio Database

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

The musical recordings (about 2.5 hours) 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 sample frequency is 44.1 kHz. 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
`cla`	instrumental music, loose tempo (classic, jazz)	00:43:18
`rhy`	instrumental music, strong rhythm (rock, pop, jazz, dance)	00:51:00
`voc`	vocal music, songs (classic, rock, pop, rap)	00:53:55

4.2 Rhythm analysis results

We evaluated the rhythm metric P_m (Eq. 5) for all signals from the database. We found the following important properties of P_m:

(i) For 99.73% of speech signals (`s1 + s2`), $P_m< 0.6$;

(ii) for 97.43 % of speech, $P_m< 0.5$;

(iii) in the case of music signals (`cla, voc, rhy`), values of P_m are not so well bounded, $P_\mathrm{m}\geq0.5$ for only 42.33% of music signals, and $P_\mathrm{m}<0.5$ for 57.67% of music signals.

Thus, we can say that if $P_\mathrm{m}\geq0.5$, we detect music signal (if an error of 2.5% is tolerated), but if $P_\mathrm{m}< 0.5$, it is not clear if the signal is speech or music.

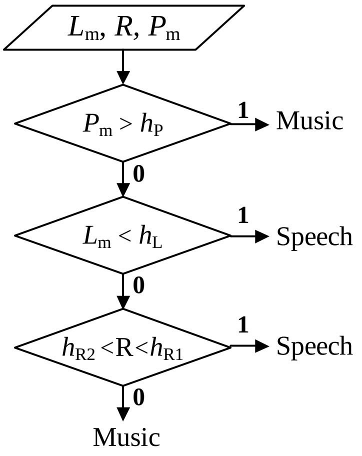
L_m, R, P_m
$P_\mathrm{m} > h_\mathrm{P}$
1
Music
0
$L_\mathrm{m} < h_\mathrm{L}$
1
Speech
0
h_R2 <R< h_R1
1
Speech
0
Music

Figure 2. Flowchart of the classification procedure

4.3 Classification

We incorporated the rhythm metric feature into the model described in [12]. As mentioned in the chapter 2, the discriminator now uses three features:

- Width of the widest peak ($\boldsymbol{L_m}$) in the modulation envelope of the band-limited signal 690-4800 Hz. The modulation envelope was estimated by summing scalefactors corresponding to the 2nd-7th subbands.
- Rate of peaks (***R***) in this envelope (i.e. the number of peaks per second)
- Rhythm metric ($\boldsymbol{P_m}$)

For evaluation, a simple rule-based threshold classifier is used. The procedure is depicted in Figure 2. The thresholds were set at:

$$h_P = 0.5, \;\; h_L = 0.7\,s, \;\; h_{R1} = 2.5\,s^{-1}, \;\; h_{R2} = 5.5\,s^{-1}.$$

5. TEST RESULTS

The discriminator was tested separately for all signal groups from the audio database. The results are shown in Table 2. The results taken from [12] are in the third column in Table 2, where just two features L_m (max. peak width) and R (rate of peaks) were used for discrimination. The results obtained by also incorporating the rhythm metric P are shown in the fourth column. We can see improvement of discrimination in all five groups of audio, but the most significant improvement is for rhythmic music (`rhy`) – 12%.

Finally, performance is improved by post-processing. The output from the discriminator is a sequence of ‘0’ and ‘1’, where each digit represents a character of the sound in one frame (frame length is 2 sec). Digits ‘0’ and ‘1’ are used for music and speech signals respectively (only non-silent sections are considered). Median filtering of the 3-rd order is applied on this output sequence. This corrects single errors due to the fact that a single speech frame wouldn’t normally be between music frames and vice versa. From the last column of the table, we can see a further increase in recognition (around 2%).

Table 2. Recognition rate of the speech/music discriminator

Name		Frames	Correct recognition %		
			Features		
			L_m+R	L_m+R+ $\boldsymbol{P_m}$	L_m+R+ $\boldsymbol{P_m}$ + Postprocess.
Speech	`s1`	573	96.34	98.25	99.30
	`s2`	554	83.39	90.75	94.40
Music	`cla`	1279	97.03	97.53	100.0
	`rhy`	1514	84.54	95.89	97.15
	`voc`	1598	93.18	95.18	97.02
Total				95.81%	**97.71%**

6. CONCLUSIONS

In the paper, a novel feature for speech-music discrimination, which determines the strength of rhythm, is introduced. By including this feature into the model developed by our research group [12], the performance of the discriminator increases significantly for certain types of audio signals.

The discriminator was tested on more than 3 hours of audio data. Average recognition rate is 97.71%. Although the discriminator uses just three features and a simple classification procedure, results are comparable to the results of other more complicated systems. However, to prove this, further evaluation of the system on a much larger amount of audio data must be carried out. A full comparison between the approach presented in this paper and other approaches is to be completed in future work.

REFERENCES

[1] N. O’Connor, et al., “Físchlár: An On-line System for Indexing and Browsing of Broadcast Television Content”, *Proc. ICASSP’01,* Salt Lake City, UT, pp. 418-421, May 2001.

[2] http://www.cdvp.dcu.ie

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

[4] A. Bugatti, et al., “Audio Classification in Speech and Music: A Comparison of Different Approaches”, *Proc. COST Workshop on Image Analysis for Multimedia Services, WIAMIS ’01,* Tampere, Finland, pp. 153-158, May 2001.

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

[6] J. Panttilä, J. Peltola, and T. Seppänen, “A Speech/Music Discriminator – Based Audio Browser with a degree of certainity Measure”, *Proc. Int. Workshop on Inform. Retrieval, IR’01*, Oulu, Finland, pp. 125-131, Sept. 2001.

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

[8] 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, pp.1-6, June 1997.

[9] 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.

[10] Nakajima Y. et al: “A Fast Audio Classification from MPEG Coded Data”, *Proc. ICASSP’99*, Phoenix, Arizona, May 1999.

[11] 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.

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

[13] E. Scheirer, “Tempo and Beat Analysis of Acoustic Musical Signals”, *J. Acoust. Soc. Am.* 103(1), pp. 588-601, Jan. 1998.

[14] J. Foote, and S. Uchihasshi, “The Beat Spectrum: A New Approach to Rhythm Analysis”, *Electr. Proc. Int. Conf. on Multimedia & Expo, ICME 2001*, Tokyo, Aug. 2001.

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.