ROBUST SPEECH / MUSIC CLASSIFICATION IN AUDIO DOCUMENTS

Julien PINQUIER, Jean-Luc ROUAS and Régine ANDRÉ-OBRECHT

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
{pinquier, rouas, obrecht}@irit.fr

ABSTRACT

This paper deals with a novel approach to speech / music segmentation based on four original features: 4 Hz modulation energy, modulation entropy and two segmental parameters. The relevance of these features is evaluated in a first experiment based on a development corpus composed of collected samples of speech and music. Another corpus is employed to verify the robustness of the employed algorithm. This experience is made on a TV movie soundtrack and shows performances reaching a correct identification rate of 90 %.

1. MOTIVATIONS

To describe and index an audio document, key words or melodies are semi-automatically extracted and speakers are detected. More recently, the problem of topics retrieval has been studied [1]. Nevertheless all these detection systems presuppose the extraction of elementary and homogeneous acoustic components. When the study addresses speech indexing [2] (respectively music indexing[3]), only speech segments (respectively music segments) are considered. In this paper, we explore a prior partitioning which consists in detecting speech and music components. The two original points of our study is to merge unusual features (4Hz modulation energy, entropy modulation and duration) and to propose a robust decision algorithm for which no training phase is necessary to process any new audio document.

This paper is divided into three parts: a definition of original features, the evaluation of the relevance of these features on a development corpus, and a description of test experiments performed on the soundtracks of audio video documents.

2. FEATURES

Many approaches to speech music discrimination have been described in the literature. On one hand, the musician community has given more importance to features which increase the choice between music / non-music. For example, the zero crossing rate and the spectral centroïd are used to separate voiced speech from noisy sounds [4], [5] whereas the variation of the spectrum magnitude (the spectral "Flux") attempts to detect harmonic continuity [6]. On the other hand the automatic speech processing community has focused on cepstral features [2]. Three concurrent classification frameworks are usually investigated, the Gaussian Mixture Model framework, the k-nearest-neighbor one [7] and the Hidden Markov Models.

In a previous paper [8], we used a Differentiated Modeling approach: two different classification systems were defined (a speech / non-speech one and a music / non-music one). We used spectral and cepstral coefficients and the modeling was based on a Gaussian Mixture Model (GMM). In this paper, we present four features: 4 Hz modulation energy, entropy modulation, number of "stationary" segments and segment duration for a more robust discrimination.

2.1. 4 Hz modulation energy

Speech signal has a characteristic energy modulation peak around the 4 Hz syllabic rate [9]. In order to model this property, the classical procedure is applied: the signal is segmented in 16 ms frames. Mel Frequency Spectrum Coefficients are extracted and energy is computed in 40 perceptual channels. This energy is then filtered with a FIR band pass filter, centered on 4 Hz. Energy is summed for all channels, and normalized by the mean energy on the frame. The modulation is obtained by computing the variance of filtered energy in dB on one second of signal. Speech carries more modulation energy than music (Figure 1).

2.2. Entropy modulation

Music appears to be more "ordered" than speech considering observations of both signals and spectrograms. To measure this "disorder", we evaluate a feature based on signal entropy. The signal is segmented in 16 ms frames, the entropy is computed on every frame. This measure is then used to compute the modulation of entropy on one second of signal. The modulation of entropy is higher for speech than for music (Figure 1).

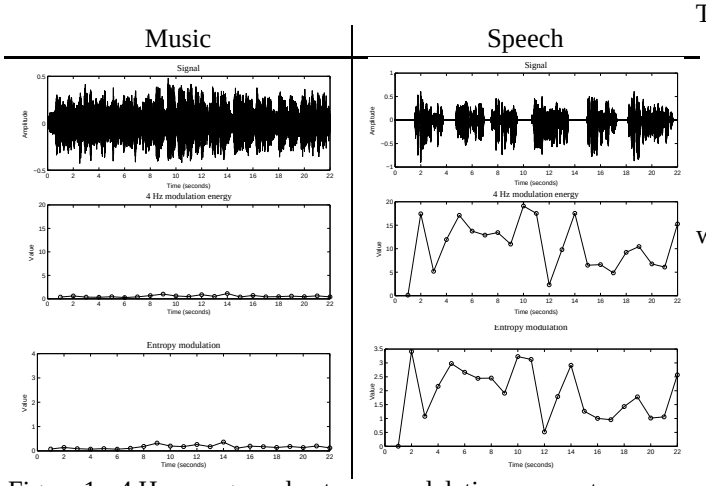

Figure 1 - 4 Hz energy and entropy modulation parameters for music and speech.

2.3. Segmental duration

2.3.1. Segmentation and speech activity detection

The segmentation is provided by the "Forward-Backward Divergence algorithm" [10] which is based on a statistical study of the acoustic signal. Assuming that the speech signal is described by a string of quasi stationary units, each one is characterized by an auto regressive Gaussian model. The method consists in performing a detection of changes in the auto regressive parameters. The use of an *a priori* segmentation partially removes redundancy for long sounds, and a segment analysis is relevant to locate coarse features. This approach have given interesting results in automatic speech recognition: experiments have shown that segmental duration carry pertinent information [11].

2.3.2. Duration

- Number of segments

The duration feature is the consequence of the application of the segmentation algorithm described above. The speech signal is composed of alternate periods of transient and steady parts (steady parts are mainly vowels). Meanwhile, music is more constant, that is to say the number of changes (segments) will be greater for speech (Figure 2a) than for music (Figure 2b). To estimate this feature, we compute the number of segments on 1 second of signal.

- Segment duration

The segments are generally longer for music (Figure 2b) than for speech (Figure 2a). We have decided to model the segment duration by a Gaussian Inverse law (Wald law).

The probability density function (pdf) is given by [12]:

$$p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} * e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}} \qquad g \geq 0$$

$$p(g) = 0 \qquad g \leq 0$$

with μ = mean value of g and $\frac{\mu^3}{\lambda}$ variance of g.

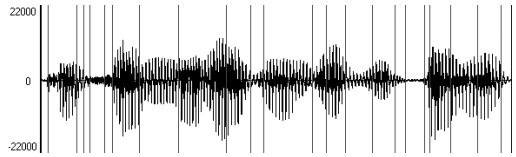

Figure 2a - Segmentation on about 1 second of speech.

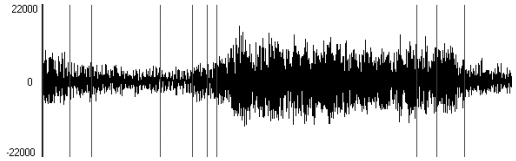

Figure 2b - Segmentation on about 1 second of music.

3. STUDY OF THE FEATURE DISTRIBUTIONS

In order to evaluate the relevance of all features, we used a corpus based on read speech excerpts (MULTEXT corpus [13]: 20kHz sampling rate) and a corpus based on various musical excerpts composed of different kinds of music, from classical music to rock (16kHz sampling rate). The total duration for each corpus (music and speech) was about 2000 seconds.

3.1. 4 Hz modulation energy

The Figure 3 shows the histogram of the 4 Hz modulation energy for speech and music components.

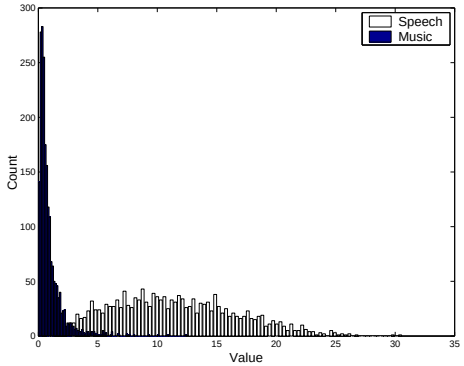

Figure 3 - Distribution of the 4 Hz modulation energy per second.

We observe that speech and music are clearly discriminated. The intersection of the two histograms (modulation energy = 2.5) can be used as a threshold. The error probabilities are:

$$Pr(music|speech) = Pr(music > threshold) = 6.4\%.$$
$$Pr(speech|music) = Pr(speech < threshold) = 3.2\%.$$

3.2. Entropy modulation

The same experiment was re-conducted for the entropy modulation feature, the results are shown on Figure 4.

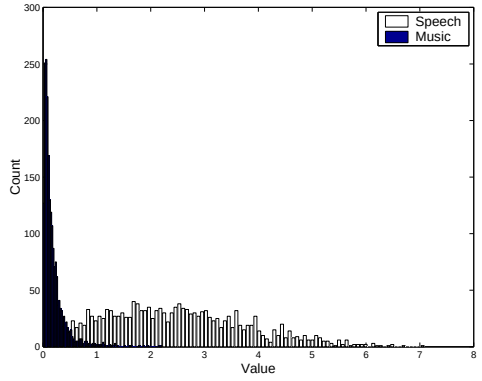

Figure 4 - Distribution of the entropy modulation per second.

This feature is also relevant in the speech/music discrimination task. Each histogram is clearly separated, and we can also determine an experimental threshold (modulation entropy = 0.5). The error probabilities are given by:

$$Pr(music|speech) = Pr(music > threshold) = 7.2\%.$$
$$Pr(speech|music) = Pr(speech < threshold) = 3.4\%.$$

3.3. Duration

3.3.1. Number of segments

The distribution of the number of stationary segments obtained automatically are represented in Figure 5.

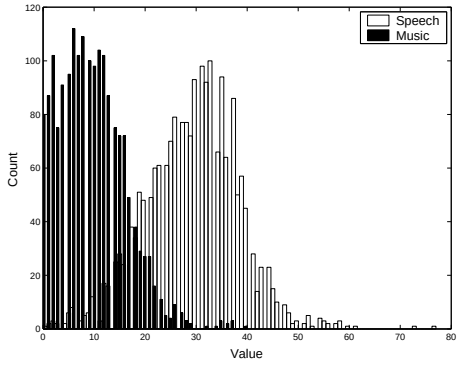

Figure 5 - Number of segments on 1s of signal.

The two separate histograms show that this feature is relevant, speech and music can be discriminated with a simple threshold (number of segments = 17). We expressed the error probabilities:

$$Pr(music|speech) = Pr(music > threshold) = 11.6\%.$$
$$Pr(speech|music) = Pr(speech < threshold) = 3.6\%.$$

3.3.2. Segment duration

The Figure 6 below describes the distributions of speech and music segments duration.

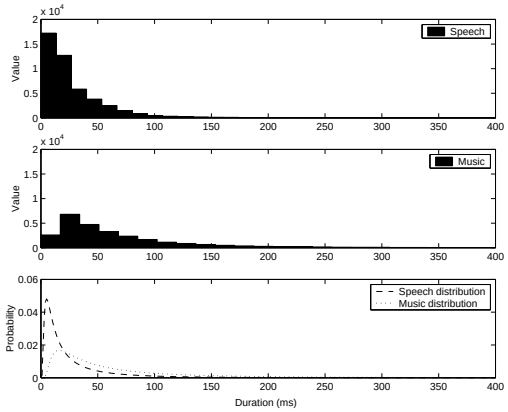

Figure 6 - Distribution of speech and music segment duration.

The parameters λ and μ of the Gaussian Inverse Model have been estimated to:

$Speech$	$Music$
$\lambda = 15.2753$	$\lambda = 50.6069$
$\mu = 30.1865$	$\mu = 74.9350$

For this feature, we propose to use a Bayesian decision to detect speech and music: a simple maximum likelihood procedure is performed.

4. EXPERIMENTS AND EVALUATION

4.1. Corpus

The experiments are made on a different corpus than the one used to study the feature distribution and to determinate the thresholds. Thus we can evaluate the robustness of our features. The experimental corpus is the audio part sampled at 16 KHz of a 20 min TV movie (“*the avengers*”). In this database, we find long periods of speech, music and “mixed” zones containing speech, music and/or noise. The speech parts are recorded in several conditions (phone calls, outdoor recordings, crowd noises) with 5 main speakers (1 woman and 4 men).

4.2. Evaluation

In the first time, we have assessed separately the three discrimination functions based on the 4 Hz energy modulation, the entropy modulation and the number of segments. The experiments (Table 1) provide similar classification rates (about 84 %). The Bayesian approach with the segment duration and the Gaussian Inverse law gives a lower classification rate (76.1%).

To improve the performance, we have proposed a hierarchical classification algorithm: we have merged the 4 Hz energy modulation and the entropy modulation criterions.

- If both classifiers agree and provide a speech (respectively non-speech) decision, the segment is labeled speech (respectively non-speech).
- If they do not agree, the decision is taken by the segment number criterion.

The final performance of this algorithm is 90.1 % of correct classification (Table 1).

Features	Performance (Correct Identification Rate)
(1) 4 Hz Modulation energy	84.1 %
(2) Entropy modulation	84.3 %
(3) Number of segments	83.2 %
(4) Segments duration	76.1 %
(1) + (2)	85.2 %
(1) + (2) + (3)	**90.1 %**

Table 1 - Best results obtained with the different features and merging of the approaches.

5. DISCUSSION

We propose in this paper four original features based on different properties of the signal. All those features considered separately are relevant in a speech / music classification task, and the correct classification rates vary from 76 to 84 %. Then, we proposed an algorithm based on a combination of those features. This approach permits to raise the correct classification rate up to 90 %.

Furthermore, we have demonstrated the robustness of our features in spite of a noisy corpus. Thus, this preprocessing is efficient and can be used for the high-level description of audio documents, which is essential for indexing the speech segments in speakers, key words, topics and the music segments in melodies or key sounds.

6. ACKNOWLEDGMENTS

This research is supported by the RAIVES project of the *"société de l'information" SHS-STIC* from the CNRS (french center for scientific research).

7. REFERENCES

[1] M. Franz, J. Scott McCarley, T. Ward and W. Zhu, *Topics styles in IR and TDT: Effect on System Behavior*, EUROSPEECH'2001, Scandinavia, pp. 287, September 2001.

[2] J.L. Gauvain, L. Lamel and G. Adda, *Audio partitioning and transcription for broadcast data indexation*, CBMI'99, Toulouse, pp. 67-73.

[3] S. Rossignol, X. Rodet, J. Soumagne, J.L. Collette and P. Depalle, *Automatic characterization of musical signals: feature extraction and temporal segmentation*, Journal of New Music Research, 2000, pp. 1-16.

[4] J. Saunders, *Real-time discrimination of broadcast Speech/Music*, ICASSP'96, pp. 993-996.

[5] T. Zhang, C.-C.J. Kuo, *Hierarchical System for Content-Based Audio Classification and Retrieval*, Conf. on Multimedia storage and Archiving Systems III, SPIE Vol. 3527, pp. 398-409, November 1998.

[6] E. Scheirer and M. Slaney, *Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator*, ICASSP'97, Munich, Vol. II, pp. 1331-1334.

[7] M.J. Carey, E.S. Parris and H. Lloyd-Thomas, *A comparison of features for speech, music discrimination*, ICASSP'99, Phoenix.

[8] J. Pinquier, C. Sénac et R. André-Obrecht, *Audio Indexing: Speech and Music components retrieval*, RFIA'2002, Angers, France, volume 1, pp. 163-170, January 2002.

[9] T. Houtgast and H. J. M. Steeneken, *A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria*, J. Acoust. Soc. Am., vol.77, No.3, pp. 1069-1077, March 1985.

[10] R. André-Obrecht, *A New Statistical Approach for Automatic Speech Segmentation.*, IEEE Trans. on ASSP, January 1988, vol. 36, n 1.

[11] R. André-Obrecht, B. Jacob, *Direct Identification vs. Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech Recognition*, Proc. of ICASSP'97, Munich, pp. 989-992.

[12] Johnson, Kotz, *Continuous Univariate Distributions*, Willey interscience publication, 1970.

[13] E. Campione and J. Véronis, *A Multilingual Prosodic Database*, Proc. of ICSLP'98, Sidney, 1998.