A FUSION STUDY IN SPEECH / MUSIC CLASSIFICATION

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

In this paper, we present and merge two speech / music classification approaches of that we have developed. The first one is a differentiated modeling approach based on a spectral analysis, which is implemented with GMM. The other one is based on three original features: entropy modulation, stationary segment duration and number of segments. They are merged with the classical 4 Hertz modulation energy. Our classification system is a fusion of the two approaches. It is divided in two classifications (speech/non-speech and music/non-music) and provides 94 % of accuracy for speech detection and 90 % for music detection, with one second of input signal. Beside the spectral information and GMM, classically used in speech / music discrimination, simple parameters bring complementary and efficient information.

1. INTRODUCTION

Commonly, to describe a sound document, key words, key sounds (jingles) or melodies are semi-automatically extracted and speakers are detected. Nevertheless all these detection systems presuppose the extraction of elementary and homogeneous acoustic components. When the study addresses speech indexing [1] (respectively music indexing [2]), speech (respectively music) segments are selected; the other segments are rejected.

Of course, the two detections are not studied with the same care. We observe two tendencies:

- authors who belong to the musician community, have given greater importance to features which increase a binary discrimination: for example, the zero crossing rate and the spectral centroïd are used to separate voiced speech from noisy sounds [3], the variation of the spectrum magnitude attempts to detect harmonic continuity [4].
- authors who study automatic speech processing, have preferred cepstral parameters [1]. Two concurrent classification frameworks are usually investigated, the Gaussian Mixture Model (GMM) framework and the k-nearest-neighbors one [5].

In this paper, we describe a system able to detect the two basic components (speech and music) with an equal performance. The system is divided in two classifications: a speech/non-speech one and a music/non-music one. Inside each classification, binary features and spectral parameters are processed at the same time. Speech and music are not considered as two classes.

This paper is divided into four parts: a presentation of our classification system, a differentiated modeling approach, a description of original features and test experiments performed on radio documents.

2. CLASSIFICATION SYSTEM

As we say above, a speech / music detection system is studied and it results of the study of two classification subsystems:

- the differentiated modeling approach (based on a spectral analysis) [6].
- the extraction of original features [7] (entropy modulation, number of segments, segment duration and 4 Hz modulation energy) provides a complementary classification.

For the speech detection, we used cepstral coefficients (speech and non-speech GMM), entropy modulation and 4 Hz modulation energy. For the music detection we use the other parameters: spectral coefficients (music and non-music GMM), number of segments and segment duration (these parameters are better for the music/non-music classification). For each classifier, we propose a statistical model, and the decision is made regarding to the maximum likelihood criterion (scores). Finally, we have two classifications for each second of input signal: the speech/non-speech one and the music/non-music one) (Figure 1).

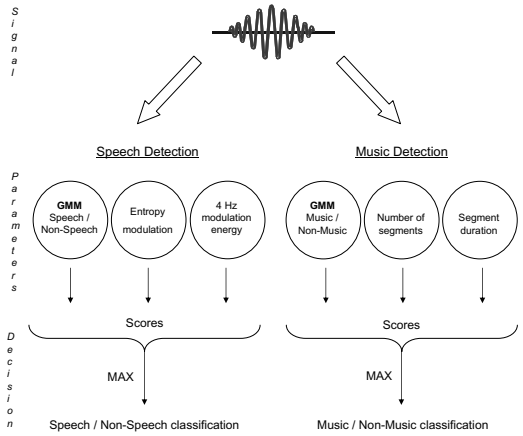

Figure 1 - Classification system.

3. DIFFERENTIATED MODELING APPROACH

In a classification problem, the differentiated modeling approach [8] can be used when specific parameter observations or specific statistical models must be defined to characterize each class. The class detection is performed by comparing a Class model and a Non-Class model estimated on the same representation space. With such an approach, the classification system has as many classifiers as classes. When studying speech and music, significant differences of production may be observed: speech is characterized by a formantic structure, whereas music is characterized by a harmonic structure. We have defined two classification systems:

- Speech classifier, S = {Cepstral space, Speech model, Non-Speech model}
- Music classifier, M = {Spectral space, Music model, Non-Music model}

3.1. Acoustic preprocessing

The speech preprocessing consists of a cepstral analysis according to the Mel scale. The soundtrack is decomposed in 10 ms frames. For each frame, 18 parameters are used: 8 MFCC plus energy and their associated derivatives. The cepstral features are normalized by cepstral subtraction. For music, a spectral analysis is made on the same frames. So, an acoustic feature vector of 29 parameters is computed: 28 filters outputs and the energy. The distribution of filters is placed on a linear frequency scale.

3.2. Classification

For each classifier, we chose to model the Class and the Non-class by GMM [9]. The classification by GMM is made by computing the log-likelihood for each model on 10 ms frames. Following this centisecond classification, a decision phase is taken on one second of signal: we choose the class, which is the more representative.

3.3. Training

The GMM training consists in an initialization step followed by an optimization step. The initialization step is performed using Vector Quantization (VQ) based on the algorithm of Lloyd [10]. The Expectation-Maximization (EM) algorithm [11], makes an optimization of parameters. After experiments, the number of Gaussian laws in the mixture has been fixed to 128 for all the models: Speech, Non-Speech, Music and Non-Music.

4. ORIGINAL FEATURES APPROACH

4.1. Speech features

- 4 Hz modulation energy

Speech signal has a characteristic energy modulation peak around the 4 Hz syllabic rate [12]. 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 filtered energy variance in dB on one second of signal. Speech carries more modulation energy than music.

- 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 ($H = \sum_{i=1}^{k} -p_i log_2 p_i$, with p_i=proba. of event i). The signal is segmented in 16 ms frames, the entropy is computed on every frame. This measure is used to compute the entropy modulation on one second of signal. Entropy modulation is higher for speech than for music.

- Classification

We use a classical Gaussian Model to describe the class-conditional probability density function (pdf) of each feature.

4.2. Music features: duration

The segmentation is provided by the "Forward-Backward Divergence algorithm" [13] 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.

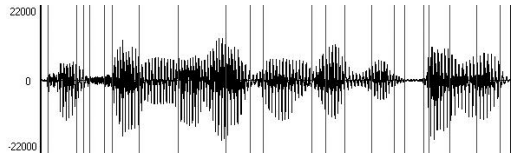

Figure 2a - Segmentation on about 1 second of speech.

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

- Number of segments

The speech signal is composed of alternate periods of transient and steady parts. 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 one second of signal, and we model it by Gaussian laws.

- Segment duration

The segments are generally longer for music (Figure 2b) than for speech (Figure 2a). We use the segment duration as feature and we decide to model it by a Gaussian Inverse law. The pdf is given by [14]:

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

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

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

5. EXPERIMENTAL RESULTS

5.1. Training

For the original features training (4 Hz modulation energy, entropy modulation, number of segments and segment duration), we used a personal database. One part consists on read speech excerpts (MULTEXT [15]) and the other on various musical excerpts composed of different kinds of music (including songs), from classical to rock music. The total duration for each corpus (music and speech) was about 30 mn.

The main database consists in multilingual radio broadcast (interviews, reports, information...) of RFI (Radio France International). The RFI database (8 hours and 20 mn) has the advantage to present long periods of speech, music and 'mixed' zones containing speech and music and/or noise. The corpus contains speech recorded in different conditions (phone call, outdoor recordings, crowd noise...) with many speakers and many languages. For the needs of the experiment, the corpus is divided in two parts. The first part (6 hours and 45 mn length) is used for the training of the GMM.

5.2. Evaluation

The second part of the RFI database (1 hour and 35 mn) is used for evaluating the relevance of each parameter and the efficiency of the system.

First time, we have tested separately all the parameters. The experiments (Table 1) provide similar accuracy (about 87 %) for entropy modulation and 4 Hz modulation energy. The number of segments gives about the same accuracy for music detection (Table 2). Only the Bayesian approach with segment duration and Gaussian Inverse law gives a lower accuracy rate (78 %). The GMM approach gives the best identification rate (about 91 % for speech detection and 87 % for music detection).

The final performance of our system is 93.9 % of accuracy for speech detection and 89.8 % of accuracy for music detection.

Features	Accuracy
(1) Cepstral coefficients (GMM)	90.9 %
(2) 4 Hz Modulation energy	87.3 %
(3) Entropy modulation	87.5 %
(1) + (2) + (3)	**93.9 %**

Table 1 - Speech / Non-Speech Classification.

Features	Accuracy
(1) Spectral coefficients (GMM)	87 %
(2) Number of segments	86.4 %
(3) Segments duration	78.1 %
(1) + (2) + (3)	**89.8 %**

Table 2 - Music / Non-Music classification.

6. DISCUSSION

We present a speech / music classification. A Differentiated Modeling approach is implemented from GMM

based on a cepstral analysis for speech and of a linear spectral analysis for music. We process four features simultaneously (entropy modulation, number of segments, segment duration and 4 Hz modulation energy) to exploit different properties of the signal. All those features considered separately are relevant in a speech / music classification task. The combination of those approaches allows to raise the accuracy rate up to 94 % for speech detection and 90 % for music detection.

It appears that we can complete a classical ("big") classification system based on a spectral analysis and GMM, with simple and robust features. Four features and four pdfs are sufficient to improve the global performance. Note that training of these models was performed on personal database (different of the RFI database), so this part of the system is perfectly robust and task-independent.

Thus, this preprocessing is efficient and can be used for the high-level description of audio documents, which is essential to index the speech segments in speakers, keywords, topics and the music segments in melodies or key sounds.

7. ACKNOWLEDGEMENTS

We would like to thank the CNRS (French national center for scientific research) for its support to this work under the RAIVES project.

8. REFERENCES

[1] J. L. Gauvain, L. Lamel, and G. Adda, "Systèmes de processus légers : concepts et exemples," in *International Workshop on Content-Based Multimedia Indexing*, Toulouse, France, Oct. 1999, pp. 67–73, GDR-PRC ISIS.

[2] 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*, vol. 28, no. 4, pp. 281–295, Dec. 1999.

[3] J. Saunders, "Real-time discrimination of broadcast speech/music," in *International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, May 1996, pp. 993–996, IEEE.

[4] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *International Conference on Audio, Speech and Signal Processing*, Munich, Germany, Apr. 1997, pp. 1331–1334, IEEE.

[5] M. J. Carey, E. J. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discriminatio," in *International Conference on Audio, Speech and Signal Processing*, Phoenix, USA, Mar. 1999, pp. 149–152, IEEE.

[6] J. Pinquier, C. Sénac, and R. André-Obrecht, "Indexation de la bande sonore : recherche des composantes parole et musique," in *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Angers, France, Jan. 2002, pp. 163–170.

[7] J. Pinquier, Jean-Luc Rouas, and R. André-Obrecht, "Robust speech / music classification in audio documents," in *International Conference on Spoken Language Processing*, Denver, USA, Sept. 2002, vol. 3, pp. 2005–2008.

[8] F. Pellegrino, J. Farinas, and R. André-Obrecht, "Comparaison of two phonetic approaches to language identification," in *European Conference on Speech Communication and Technology*, Budapest, Hongrie, Sept. 1999, pp. 399–402, 5-9 sep.

[9] M. Seck, I. Magrin-Chagnolleau, and F. Bimbot, "Experiments on speech tracking in audio documents using gaussian mixture modeling," in *International Conference on Audio, Speech and Signal Processing*. May 2001, vol. 1, IEEE.

[10] J. Rissanen, "An universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, pp. 416–431, Nov. 1982.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39 (Series B), pp. 1–38, 1977.

[12] T. Houtgast and J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[13] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 36, no. 1, Jan. 1988.

[14] *Continuous Univariate Distributions*, Willey interscience publication, New-York, USA, 1970.

[15] E. Campione and J. Véronis, "A multilingual prosodic database," in *International Conference on Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 3163–3166.