To index efficiently the soundtrack of multimedia documents, it is necessary to extract elementary and homogeneous acoustic segments. In this paper, we explore such a prior partitioning which consists in detect the two basic components, which are speech and music components. The originality of this work is that music and speech are not considered as two classes and two classification systems are independently defined, a speech/non-speech one and a music/non-music one. This approach permits to better characterize and discriminate each component: in particular, two different feature spaces are necessary as two pairs of Gaussian mixture models. More, the acoustic signal is divided into four types of segments: speech, music, speech-music and other. The experiments are performed on the soundtracks of audio video documents (films, TV sport broadcasts). The performance proves the interest of this approach, so called the Differentiated Modeling Approach.

1. INTRODUCTION

With the fast growth of audio and multimedia information, the number of documents, such as broadcast radio and television increases greatly and the development of technologies for spoken document indexing and retrieval is in full expansion. Commonly, to describe a sound document, key words, key sounds (jingles) or melodies are semi-automatically extracted, 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 [2] (respectively music [3]) indexing, speech (respectively music) segments are selected; the other segments are rejected. Of course, the two detections are not studied with the same attention. In this paper, we explore a prior partitioning which consists in detect the two basic components, which are speech and music components, with an equal performance. In that purpose, music and speech are not considered as two classes (two classification systems are independently defined, a speech/non-speech one and a music/non-music one) and more, two observation spaces are used. We called this approach, the Differentiated Modeling Approach to emphasize the fact it is necessary to better characterize and discriminate each component through its own feature space and its own statistical modeling. The first part of this paper points out the benefit of the Differentiated Modeling; we precise the two different feature spaces and the using of the Gaussian mixture models in each case. In the second part, we describe some experiments which are performed on the soundtracks of audio video documents (TV movies, sport reports). The performance proves the interest of this approach.

2. THE SEPARATION OF SPEECH AND MUSIC

As we have said previously, in a classic approach of discrimination between speech and music components, the choice is binary. Among the classic methods, some authors who belong to the musician community, have given greater importance to features which increase this binary discrimination: for example, the zero crossing rate and the spectral centroid are used to separate voiced speech from noisy sounds [4], the variation of the spectrum magnitude (the spectral "Flux") attempts to detect harmonic continuity [5]. Authors who study automatic speech processing, have preferred cepstral features [2]. Two concurrent classification frameworks are usually investigated, the Gaussian Mixture Model framework and the k-nearest-neighbor one [6].
In a classification problem, the Differentiated Modeling approach [7] may be introduced when specific parameter observations or specific statistical models must be defined to characterize each class. The detection of each class is performed by comparing a Class model and a Non-Class model estimated on the same representation space. Then the classification problem is reduced to precise for each class the set C defined as follow:

- \( C = \{\text{Representation space, Class model, Non-Class model}\} \)

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. The differentiated modeling is thus completely adapted to our problem. We have defined two sets:

- Speech set, \( S = \{\text{Cepstral space, Speech model, Non-Speech model}\} \)
- Music set, \( M = \{\text{Spectral space, Music model, Non-Music model}\} \)

These sets will be specified in the following paragraph.

### 3. THE SYSTEM

The extraction of speech and music parts being made in a separate way and on the general schema (cf. figure 1), the system is divided in two sub systems, each one consists of two modules: Acoustic preprocessing and Classification. Due to the Differentiated Modeling Approach, these two modules are very different for each of both classes. They are followed by a possible module of fusion which must be developed according to the aimed real application.

#### 3.1. Acoustic preprocessing

The speech preprocessing consists of a cepstral analysis according to Mel scale. The soundtrack is decomposed in frames of 10ms. For each frame, 18 parameters are used: 8 MFCC plus energy and their associated derivatives. The cepstral features are normalized by a cepstral subtraction [8]. For music, a simple 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 set, we chose to model the Class and the Non-class by a Gaussian Mixture Model (GMM) [9]. The classification by GMMs is made by calculation of the log-likelihood for each model of Class and of Non-Class. Following this classification phase, a phase of merging allows to concatenate neighboring frames which have obtained the same index during classification. A smoothing function is necessary to delete insignificant size segments. This function processes in two successive phases. A meadow smoothing deletes the segments of length lower than 20ms, i.e. not significant for speech and for music. The second step consists in keeping the important zones (in size) of speech (respectively of music). This smoothing is about 400ms for the Speech system and about 2000ms for the Music one.

### 3.3. Training

The modeling is based on a Gaussian Mixture Model (GMM). The training consists in an initialization step then in an optimization step (cf. figure 2). Initialization step is performed using Vector Quantization (VQ) based on the algorithm of Lloyd [10]. Optimization of parameters is made by the classic algorithm Expectation-Maximization (EM) [11]. After experiments, the number of Gaussian laws in the mixture has been fixed to 32 for the models of Speech and Non-Speech and in 16 for the models of Music and Non-Music.

The training requires the use of a manually labeled corpus.

### 4. EXPERIMENTAL RESULTS

For our experiments, we used a subset of a database provided by the French National Institute of Audiovisual (INA) in the framework of the French national research project AGIR. This subset contains various extracts of television programs including sport broadcasts and a TV movie.

#### 4.1. Data base

The corpus consists of three television documents: the first one is a comment of a championship of figure skating of 30mn, the second is a sport report of 34mn and the third is a TV movie of 50 mn. This database 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...) with 9 main speakers (3 women and 6 men). For the needs of the experiment, the corpus is divided in two parts. The first one represents 74mn (35mn of TV movie, 14mn of sport report and 25 mn of championship) and is used for the training. The second one of 40mn (15mn of TV movie, 20mn of sport report and 5mn of championship) is used for recognition.
4.2. Modeling

The Speech model is trained from signal containing audible speech (i.e. pure or mixed with music when it does not damage too much speech or mixed with noise such as noise generated by the phone canal or a stadium). 32 (respectively 16) Gaussian laws are necessary to model Speech (respectively Music) and Non-Speech (respectively Non-Music).

4.3. Results of indexation

The figure 3 presents the results obtained during the Speech/Non-speech classification and the Music/Non-Music classification. Evaluation of the automatic classification has been made in comparison with the manual indexation. Having aligned both segmentations, we have measured various delays between automatic frontiers and corresponding manual frontiers if they exist. We have noted the number of insertions and deletions.

We observe there is always a manual frontier corresponding to every automatic frontier. The delays (1) are good: 96% are lower than 20 cs. There are very few deletions and insertions, and no substitution is made. The precision of the speech classification (accuracy) (2) is excellent: 99.5%.

For the music, the rate of delays lower than 20 cs is 91 % and the accuracy is 93 %. The music insertions correspond to car noise, rustlings of tires, explosions and elevator noise which are not represented in the training. The non-music insertions correspond to segments containing some very weak music.

<table>
<thead>
<tr>
<th>Results</th>
<th>Speech</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay &lt; 20cs</td>
<td>96%</td>
<td>91%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99.5%</td>
<td>93%</td>
</tr>
</tbody>
</table>

Figure 3: Evaluation of Speech and Music indexation. Percentage of delays lower than 20 cs.

The figure 4 presents an example of indexation of about 14 seconds of signal extracted from TV movie.
This figure shows the results of speech and music indexations compared with manual indexing; we notice that the frontiers are correct. There is only little shift between the manual and the automatic indexing. The music segment at time 2:08 does not appear because of the smoothing procedure (2000 ms).

### 5. CONCLUSION

We have presented the first experiments of a system of speech/music classification based on a Differentiated Modeling Approach. The idea of this method is to associate to each of the classes its own feature space of representation and its own statistical modeling. This approach is implemented from GMM based on a cepstral analysis for speech and of a linear spectral analysis for music. The results of the indexation, for speech and for music are excellent considering the nature of the corpus and the limited quantity of training data. The indexation is reliable: for every automatic frontier, there is always a manual frontier corresponding to it and associated with the same category of transition. Further experiments must complete the first experiments and confirm these results.

### 6. ACKNOWLEDGEMENTS

This work was supported by RNRT (“Réseau National de Recherche en Télécommunications”) under the AGIR (“Architecture Globale pour l’Indexation et la Recherche par le contenu de données multimedia”) project.

### 7. REFERENCES


