Audio Indexing on the Web: a Preliminary Study of Some Audio Descriptors

Nathalie Parlangeau-Vallès, Jérôme Farinas, Dominique Fohr, Irina Illina, Ivan Magrin-Chagnolleau, Odile Mella, Julien Pinquier, Jean-Luc Rouas, Christine Sénac

AUDIO INDEXING ON THE WEB : A PRELIMINARY STUDY OF SOME AUDIO DESCRIPTORS

Nathalie Parlangeau-Vallès[(1)], Jérôme Farinas[(2)], Dominique Fohr[(1)], Irina Illina[(1)], Ivan Magrin-Chagnolleau[(3)], Odile Mella[(1)], Julien Pinquier[(2)], Jean-Luc Rouas[(2)], Christine Sénac[(2)]

[(1)] LORIA – Campus Scientifique – BP 239 – 54506 Vandoeuvre-les-Nancy - France
[(2)]IRIT – Université Toulouse III – 118, route de Narbonne – 31062 Toulouse – France
[(3)] Laboratoire Dynamique Du Langage – CNRS & Université Lyon 2 – 14, avenue Berthelot – 69363 Lyon Cedex 07 - France

ABSTRACT

The "Invisible Web" is composed of documents which can not be currently accessed by Web search engines, because they have a dynamic URL or are not textual, like video or audio documents. For audio documents, one solution is automatic indexing. It consists in finding good descriptors of audio documents which can be used as indexes for archiving and search. This paper presents an overview and recent results of the RAIVES project, a French research project on audio indexing. We present speech/music segmentation, speaker tracking, and keywords detection. We also give a few perspectives of the RAIVES project.

Keywords : audio-content indexing, speech/music detection, speaker tracking, keyword detection.

1. INTRODUCTION

Internet has become a very important medium of communication during the last few years. Most of the search engines currently access mainly the HTML pages (or equivalent textual data). But there is an important part of the data which is not accessible, because the data is not indexed, has a dynamic content, or belongs to a category which is not easily accessible. All this data belongs to what is called the *invisible web*, including audio and video documents.

In this paper, we describe some techniques used to structure and index audio documents. Audio indexing systems can be based on a complete transcription but it is not the only meaning-full information which can be extracted from an audio document. Non-verbal information is also formative for an audio document, and can lead to the extraction of pertinent descriptors. We focus on this kind of information extraction.

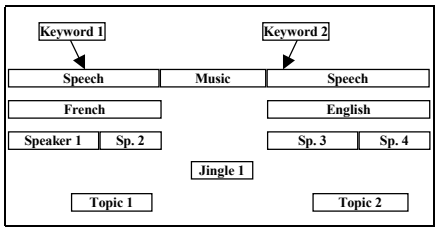

Figure 1: Example of audio descriptors.

For instance, as shown on Figure 1, we can separate speech segments from music segments, detect key sounds (like jingles), identify the language of a segment, segment by speakers, detect some keywords, or extract the main topics.

In this paper, we present some preliminary work done in the framework of the RAIVES project, a French research project on audio indexing. Then we present three audio descriptors, namely speech/music segmentation, speaker tracking and keyword detection, and we give some results for these three audio descriptors. We finally conclude this paper and give some perspectives of the RAIVES project.

2. DESCRIPTION OF THE RAIVES PROJECT

The Database

In the RAIVES project, we want to be able to search and index radio data on the web. We contacted the French public radio station RFI (Radio France International) to get some good quality radio data. The database that we got is composed of 10 hours of programs in 18 languages (French, English, Spanish, Portuguese, Mandarin, etc.) for a total of 180 hours of stereo 44.1 kHz data. Programs are broadcast news as well as interviews and musical programs. In the first phase of this project, we work directly on this good quality data. In a second phase, we will code this data using the most commonly used coders on the Web and study the influence of the coders on the performance.

System Architecture

As shown on Figure 2, six modules make our indexing system, each of them dedicated to a specific cue detection.

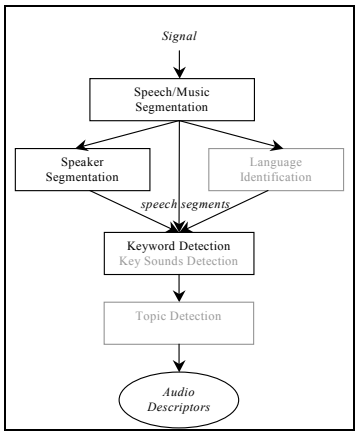

Figure 2: Synoptic schema of the RAIVES indexing system.

The first module is devoted to the separation of speech and music segments. The speaker segmentation module permits to segment the signal into speaker turns or to identify some of them when possible. Prior to keyword detection, a language identification module permits to determine the spoken language and to choose the adequate keyword system. Then a topic detection module allows us to select events related to the same topic. Additionally, a key sounds detection module is used to identify sounds characteristic of a particular structure in audio documents (like jingles or applauds for instance).

3. RESULTS ON SOME AUDIO DESCRIPTORS

Experimental Database

The experimental database for this work was a subset of the database consisting in 8 French programs re-sampled from 44.1 kHz stereo to 16 kHz mono. These programs were very diverse in term of speech and musical contents, speaking styles, speakers, noise conditions and channels. Only the speaker tracking system has been tested on different data because there were not enough labeled material for this task.

Speech/Music Segmentation

Principle

The first task of interest is the tracking of speech and/or music segments in order to segment an audio document into speech and music portions. Most of broadcast news transcription systems use this kind of separation of speech/music segments in order to confine the application of speech transcription systems to speech segments [1]. This segmentation is often related to a specific structure of the document (advertisements, jingles, etc.), and it seems important to keep it as a descriptor in an audio indexing system. When studying speech and music, significant differences of production may be observed: speech is characterized by formant structure whereas music is often characterized by harmonic structure. Recently, several approaches have been investigated for the discrimination of speech and music signals. They use mainly features permitting to capture the temporal and spectral structures of the signal. Besides the classical cepstral coefficients, these features include zero crossing rate and the spectral centroid that are used to separate voiced speech and noisy sounds, the variation of the spectrum magnitude (the spectral "flux") which attempts to detect harmonic continuity, the 4Hz modulation energy, the entropy modulation, the number of stationary segments and segment duration which have been used by [10] for the speech/music segmentation. Moreover, another key point is the choice of a good classifier. Various classifiers are commonly used: k-Nearest-Neighbors, Hidden Markov Models [1], Gaussian Mixture Models (GMM) [14], [10].

Implementation

What seems important in speech/music separation is the notion of independency between the two tracked cues. Also, as in [10] we use a "differentiated modeling" which permits to exploit the structural differences between speech and music. The problem is reduced to the specific for each class (speech or music) the set C defined as follows:

C = {Representation space, Class model, Non-Class model}

The extraction of speech and music parts being made in a separate way. The system is divided in two sub-systems (one for speech and one for music). Each one consists of two modules: the acoustic preprocessing and the classification. Speech preprocessing consists of a cepstral analysis (8 MFCC plus energy and their derivatives) followed by a cepstral mean substraction. For music, a simple spectral analysis is made (28 filters outputs and the energy). For each set, the Class and Non-Class are modeled by a GMM: after experiments, the number of Gaussian mixtures has been fixed to 128 for all the models. The classification has been made by calculation of the log-likelihood for each model of Class and Non-Class. Following this classification phase, a phase of merging allows to concatenate neighboring frames with the same class index. Then special smoothing functions are applied to keep only significant speech (respectively music) segments.

Evaluation

For the training, programs 1 to 7 excluding program 4 were used providing approximately a total duration of 3 hours, 1h20mn for music, 1h21mn for speech and 19mn of noise. The system was evaluated using the program 4

in a first time. This program contains interviews recorded in very noisy environment; it mainly contains speech from both male and female speakers with different types of music in the background. Then a Spanish program was tested containing interviews and music where the environment is much less noisy than for the program 4. Results obtained during the speech/non speech and the music/non music decisions are shown in the table 1. The evaluation of the automatic decision has been made in comparison with the manual labeling and the accuracy was computed with:

$$\text{Accuracy} = (\text{length}_{\text{test corpus}} - \text{length}_{\text{insertions}} - \text{length}_{\text{deletions}} - \text{length}_{\text{substitutions}}) / (\text{length}_{\text{test corpus}})$$

	Music	Speech
Program 4	72.73%	87.08%
Spanish program+ Program 4	84.24%	88.24%

Table 1: Results in term of accuracy for both programs

For Speech/NonSpeech segmentation, major errors occur in case of speech classified as non speech: speech superimposed with music and noise, rap music, very hardly audible speech, and some clean speech. For Music/NonMusic segmentation, major errors occur in case of non music segments classified as music segments : very noisy segments and false errors. In fact, these errors are induced by a too simple modeling but the lack of segments containing "singing speech", *a capella* singing and rap has prevent us to train a new class for these three classes that are now included in the music class.

Speaker Tracking

Principle

There are several ways to extract information about speakers from an audio document. One of them is called speaker tracking and consists in looking for all the segments which have been uttered by a particular speaker. That implies that we already have a model for that particular speaker. We also suggest that the speech/music segmentation has been done accurately, and we do the speaker tracking only on the speech portions of data.

The first phase of a speaker tracking system consists in learning statistical models for each speaker that we want to track. We first apply a cepstral analysis to several utterances that have been pronounced by a target speaker. We use feature vectors composed of 16 MFCC (without the first one), 16 Δ MFCC and the Δ log-energy. Therefore, feature vectors are 33-dimensional vectors. These feature vectors are calculated on signal frames of 20 ms every 10 ms. The Δ calculation is done with a time span of 5 vectors. Then, we train a Gaussian mixture model (GMM) composed of 128 Gaussian distributions with diagonal covariance matrices. The GMMs are trained using an expectation-maximization (EM) algorithm initialized by a vector quantization (VQ) algorithm. We finally obtain a GMM for each speaker that we want to track. We also learned a GMM corresponding to the data of several female and male speakers pooled together. This model, called a world model or a background model, is used to normalize the likelihood scores during the tracking phase.

Once all the GMMs have been learned, the tracking of one or several speakers can be done on the speech portions of any audio document. First, the audio document is converted into feature vectors as described previously. Then a decision is made for each feature vector in the following way: a log-likelihood ratio is calculated using the target speaker model and the world/background model. Then a smoothed log-likelihood ratio is calculated by averaging the log-likelihood ratio values of a block of 31 feature vectors around the current vector. Before the average calculation, a Hamming window is applied to the block of vectors. Finally, the smoothed value is compared to a threshold, and the feature vector is labeled with the target speaker identity if this smoothed value is higher than the chosen threshold.

Implementation

For our experiments, we could not use the data of the RAIVES project yet because there were not enough data labeled by speakers with common speakers over several programs. So we decided to test our system on a subset of the HUB4 database as a preliminary study. The subset that we used was composed of 15 news programs of about 30 minutes, sampled at 16 kHz, coded by 16 bits. 7 programs were used to train a world/background model and three target speaker models (1 female and 2 males). The 8 other programs were used for testing.

Evaluation

The results of these preliminary experiments are shown on Figure 3 under the form of a DET curve. This curve represents the miss probability as a function of the false alarm probability, that is, all the couples (miss probability, false alarm probability) for all the possible values of the threshold.

The equal error rate (EER), which is the point corresponding to the equality between the two types of errors, is 10.2%.

Perspectives

The next step, for this module, will be to tune the various parameters of the system on the subset of the HUB4 database and then to test it on the RAIVES data when more data is labeled. We will also develop a gender detection system in order to segment speech data into male and female speakers. This will be a first step to the speaker segmentation module. We will finally develop a

complete speaker segmentation system, which consists of determining the set of speakers presented within a given audio document as well as the boundaries of each intervention without using any *a priori* information on speakers.

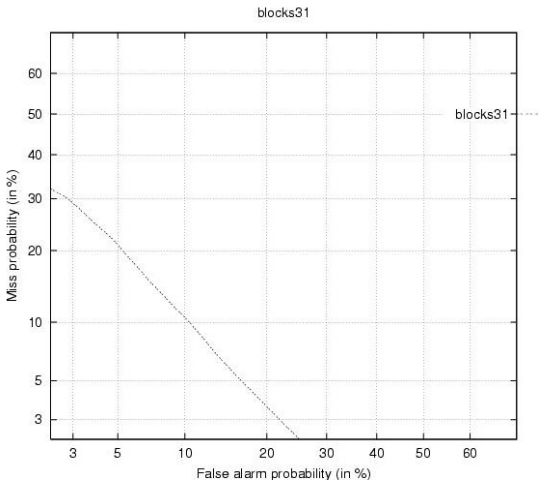

Figure 3: DET curve for the speaker tracking system on a subset of the HUB4 database.

Keyword Detection

Principle

Word-spotting systems based on hidden Markov models are considered more efficient at modeling arbitrary speech than template based systems [11]. The most obvious approach is to use a large vocabulary continuous speech recognition system (LVCSR) to produce a word string. Then, search algorithms are applied for keyword detection in that string. This approach is considered as giving very good results [15] with the drawback of a high computational cost and the need of a large training database. Another common approach is based on the use of keyword and filler models. These latest represents the non-keyword of the utterance [7]. Models can be the sub-word keyword models like phonetic models or can be the whole-word models.

Implementation

Our approach is based on the use of phonetic models for the keywords and the filler. This method do not need a large amount of training data. The filler model is a union of all possible phonetic models. In order to favor the keyword detection, a weighting factor proportional to the number of phones of the word is used: the longer the keyword, the more important the weighting factor.

The acoustico-phonetic decoding system is based on phonetic models. These latest are 3-states speaker-independent context-free models. Since only a small part of the RAIVES database is already transcribed, these models are trained on Bref80 which is a corpus of read-speech in very clean conditions. Therefore, in order to reduce the mismatch between the training database and the broadcast database conditions (non native speakers, noisy and musical backgrounds,..), these models have been adapted using a transcribed part of the RAIVES corpus. This supervised adaptation uses a Maximum Likelihood Linear Regression algorithm (MLLR).

The keyword detection step begins with a channel detection based on the spectral shape of the signal to separate broadcast quality speech and telephone quality speech. So as to adapt the phonetic models to the testing conditions, an unsupervised SMLLR (Structural MLLR) adaptation is performed on the test data. SMLLR adjusts the number of linear regression matrices that will be applied to the estimation of Gaussian mean vectors. It uses a binary tree structure that cluster the gaussians of HMM models, according to the available amount of adaptation data [6].

Evaluation

We used 31 phonetic models (including pause) and a cepstral parameterization: 12 MFCC, 12 Δ MFCC, 12ΔΔ MFCC removing C_0. Program2 has been chosen among transcribed programs in order to implement the supervised adaptation.

We evaluated the performances of our system on four different programs: Bref80 corpus (read-speech in clean conditions), program7 (broadcast program characterized by a very important number of speakers and quite few musical segments), program4 (female and infant speakers, few music) and program8 (a lot of music, male speakers).

We first evaluated the acoustico-phonetic decoding system. Results are shown in table 2. Results show that adaptation is efficient in spite of the weak amount of adaptation data.

	Bref80	Program 8	Program 4	Program 7
No Adaptation	72%	56.7%	48.9%	37.9%
Adaptation		63%	54%	45%

Table 2 : acoustico-phonetic decoding accuracy.

For the evaluation of the keyword detection system, a set of 20 keywords has been defined for each program (12 keywords are in the file, 8 are not). For this experiment, telephone speech segments have been ignored. For different number of false alarms rates, the probability of keyword detection is computed (ROC). Figure 4 shows the results for the four programs.

These preliminary results are very encouraging considering our small training corpus. With less than two keywords inserted per hour, the detection probability is about 80% for clean programs. In spite of the 10 % phonetic recognition rate difference between Bref80 and program8, the keyword detection rate is equivalent. Detection rate declines very quickly with degraded

conditions like noise and music surimposed with speech. Our future investigations will focus on this problem. The integration of speech/music detector and speaker segmentation gives us hope to have a better model adaptation. Moreover, the method has a lower computational cost than LVCSR. This 20-keyword detection task is real-time on a PC. This criterion is important according to the envisaged application for the web.

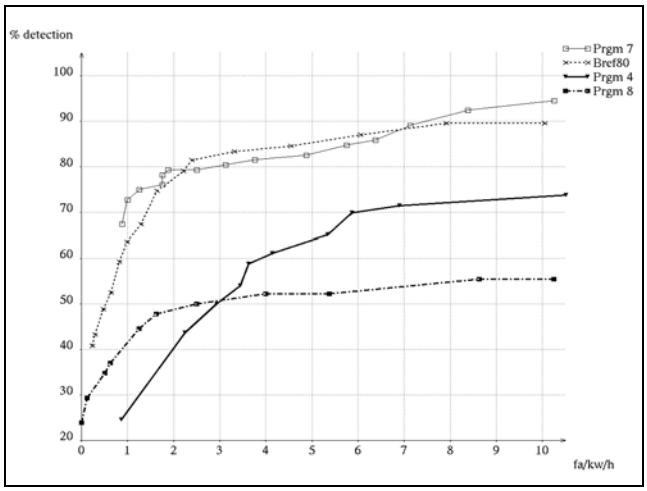

Figure 4: Results for keyword detection for the four programs.

4. CONCLUSION AND PERSPECTIVES

In this paper, we have presented the RAIVES project, a French project about audio indexing on the web. After describing briefly the database and the architecture of our system, we presented in details three modules of audio descriptor extraction, namely speech/music segmentation, speaker tracking, and keywords detection.

Important issues will emerge on with the fusion of these three first descriptors. We will also investigate other audio descriptors: language identification and topic detection in order to have a complete indexation system. The last step of this project will be to evaluate the performance of our algorithms on compressed versions of the audio documents, using the most common compressions found on the Web nowadays.

5. AKNOWLEDGEMENTS

We would like to thank the CNRS (French national center for scientific research) for its support of the RAIVES project through the program “Information Society”. Great thanks to Régine André-Obrecht who helped this work to get started.

6. REFERENCES

[1] Ajmera, J., McCowan, I.A., and Bourlard, H., “Robust HMM-Based speech/music segmentation,” *Proceedings of ICASSP 2002*.

[2] Brun, A., Smaïli, K., and Haton, J.P, "Contribution to Topic Identification By Using Word Similarity," *Proceeding of ICSLP 2002*.

[3] Chen, S. and Gopalakrishnan, P., “Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion,” *Proceedings of the Broadcast News Transcription & Understanding Workshop*, 1998.

[4] Farinas, J., Pellegrino, F., Rouas, J.L., and André-Obrecht, R., “Merging Segmental And Rhythmic Features For Automatic Language Identification,” *Proceedings of ICASSP 2002*.

[5] Gauvain, J.L., Lamel, L., and Adda, G., “Audio partitioning and transcription for broadcast data indexation,” *Proceedings of CBMI 1999*.

[6] Lauri, F., Illina, I., and Fohr D., "Comparaison de SMLLR et de SMAP pour une adaptation au locuteur en utilisant des modèles acoustiques markoviens," *Proceedings of the XXIVèmes Journées d'Etude sur la Parole*, 2002.

[7] Manos, A. and Zue, V., “A segment-based wordspotter using phonetic filler models,” *Proceedings of ICASSP 1997*.

[8] Meignier, S., Bonastre, J.-F., and Igounet, S., “E-HMM approach for learning and adapting sound models for speaker indexing,” *2001: A Speaker Odyssey*.

[9] Petrucci, G., El-Maleh, K., Klein, M., and Kabal, P., “Speech/Music discrimination for multimedia application,” *Proceedings of ICASSP 2000*.

[10] Pinquier, J., Rouas, J.L., and André-Obrecht, R., “Robust speech/Music classification in audio documents,” *Proceedings of ICSLP 2002*.

[11] Rose, R.C. and Paul, D.B., “A hidden Markov model based keywords recognition system,” *Proceedings of ICASSP 1990*.

[12] Saunders, J., “Real-time discrimination of broadcast Speech/Music,” *Proceedings of ICASSP 1996*.

[13] Scheirer, E. and Slaney, M., “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator,” *Proceedings of ICASSP 1997*.

[14] Seck, M., Magrin-Chagnolleau, I., and Bimbot, F., “Experiments on speech tracking in audio documents using Gaussian Mixture Modeling,” *Proceedings of ICASSP 2001*.

[15] Weintraub, M., “Keyword-spotting using SRI's DECIPHER large vocabulary speech recognition system,” *Proceedings of EUROSPEECH 1993*.