# SOUND CHANNEL VIDEO INDEXING

*Claude Montacié and Marie-José Caraty*

**LIP6** - Université Pierre et Marie Curie - CNRS
4, place Jussieu - 75252 Paris Cedex 5 - France
Tel. (33/0) 1 44 27 62 81, FAX (33/0) 1 44 27 70 00, E-mail: montacie@laforia.ibp.fr

## ABSTRACT

We present in this paper preliminary results using speaker recognition and speech recognition techniques, designed at LIP6, to index audio data of video movies. The assumption that only one person is speaking at the same time is made.
In a first approach, we work on dialogue unsupervised indexing using speaker recognition techniques. For this purpose, we develop Silence/Noise/Music/Speech detection algorithms in order to cut audio data in segments that we hope to be homogeneous in terms of speaker appartenance.
In a second approach, we develop a supervised audio data indexing method knowing the movie script.

## 1. INTRODUCTION

The interest in video databases is increasing fast. Video databases can include video conferencie archives, TV, movies , advertisements, etc...Future Video On Demand (VOD) servers will store hundreds or thousands movies. National or corporate archives store millions of hours of video, as for instance at the french Institut National de l'Audiovisuel (INA). Presently such archives are indexed manually and can only be accessed through the mediation of human experts, whose availability is limited.

In this paper, only the automatic sound channel indexing of video database is studied. This sound channel is the superposition of noise, music and speech, sometimes simultaneously. Superposed voices separation is a very difficult task which can be solved by array processing techniques during the sound recording. Some studies deal with not superposed voices separation [1, 2, 3, 4, 5], but none yet on cinematic movie audio data indexing. We have used speaker and speech recognition techniques for this task. At first, a Silence/Noise/Music/Speech detection algorithm has been developed based on the AR-Vector modeling. This technique is usually used for the speaker recognition. Then a warping technique between the sound channel and a movie script based on the Hidden Markov Model (HMM) is developed. It is an extension of the automatic phonetic labeling of the speech database. The first approach is unsupervised, the second supervised by the knowledge of the video movie script.

## 2. UNSUPERVISED TECHNIQUES

These approaches are based on the assumption that no knowledge on the video movie is available. This knowledge could be speakers number or identities, dialogs, or the script. In this preliminary experiment, a segmentation algorithm based on a speaker recognition modeling is developed and tested.

### 2.1 Speaker Recognition Techniques

Our speaker recognition techniques are based on the Auto-Regressive Vector (ARV) model. ARV model is successfully used in speaker recognition [6].
Let $\{y_n\}$, (n = 1 ,.., N) be a succession of N cepstral p-dimensional vectors. Their evolution is described by an q-dimensional ARV model :

$$y_n = \sum_{i=1}^{q} A_i \, y_{n-i} + e_n$$

where $\{A_i\}$, (i = 1 ,.., q) are p x p matrices and $\{e_n\}$, (n = 1,.., N) is a vectorial white noise which has a covariance matrice D.
The chosen inter-speakers measure $IS_I$ [7] is based on the Itakura measure [8].

### 2.2. Segmentation Method

We propose an automatic segmentation method based on the ARV model. It's an extension of the Forward-Backward Divergence (FBD) method [9] used to detect signal discontinuities. The purpose of the vectorial extension is to detect spectral discontinuities on the analysis vectors evolution. The signal is described by a set of Mel Frequency Cepstral Coefficients (MFCC). Let $w_0$, $w_1$, $w_2$ be three moving windows depicted in figure 1. Three ARV models are computed on these windows and compared so as to detect discontinuities on the spectral evolution. The distance $IS_I$ between the windows $w_1$ and $w_2$, normalized by the distance $IS_I$ between the windows $w_1$ and $w_0$, is computed. When this measure is upper than a given threshold, a Speaker/Music/Noise change occurs.

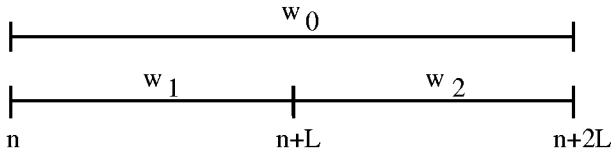


**Figure 1** : Relative locations of the windows $w_0$ {n, ..., n+2L}, $w_1$ {n, ..., n+L}, $w_2$ {n+L, ..., n+2L}

Preliminary experiments are carried out on the concatenation of 200 sentences uttered by 200 speakers. It's a part of the TIMIT speech database [10]. The segmentation method is also tested on a part of a CD-I movie audio data ("Un indien dans la ville" of Hervé Palud.

### 2.3. Database: the CD-I Movie

The first ten minutes of the movie "Un indien dans la ville" are chosen as database. The orignal sampling frequency of the CDI audio channel is 44.1 kHz. The audio format is the first layer of the MPEG1 standard. The signal has been decompressed and undersampled at 16kHz (i.e., the TIMIT sampling frequency). The database labeling is necessary to compare efficiently the algorithms. Without any automatic labeling (i.e., the subject of this paper), it is essential to supervise the labeling. An estimation of the time cost of such an operation is two minutes per second of labeling. We distinguish four kinds of sounds (i.e., noise, music, speech and silence), for which we give the duration, the duration percentage, the number of segments and the number of segments longer than 3 seconds in the whole database.

| Sounds | Duration | Percentage | Seg. | Seg.> 3 s |
|---|---|---|---|---|
| Speech | 368.4 s | 61.4 % | 215 | 29 |
| Musics | 100.8 s. | 16.8 % | 41 | 11 |
| Noises | 105.4 s. | 17.6 % | 104 | 4 |
| Silence | 25.4 s | 4.2 % | 51 | 0 |

**Table 1** : Kinds of sounds characterizing the movie audio data.

Despite the variability of noises, in order to identify these sounds it's necessary to classify the noises. We choose eleven labels to describe the noises.

| Noises | Duration | Percentage | Seg |
|---|---|---|---|
| Rumbling | 23.0 sec. | 21.8 % | 21 |
| Explosion | 0.9 sec. | 0.9 % | 1 |
| Jangling | 2.0 sec. | 1.9 % | 4 |
| Creaking | 6.9 s. | 6.5 % | 10 |
| Friction | 1.3 s. | 1.2 % | 1 |
| Purring | 3.4 s. | 3.2 % | 5 |
| Humming | 8.3 s. | 7.9 % | 5 |
| Pattering | 5.7 s. | 5.4 % | 10 |
| Background | 45.4 s. | 43.1 % | 40 |
| Scraping | 0.2 s. | 0.2 % | 1 |
| Sputtering | 8.4 s. | 7.9 % | 10 |

**Table 2** : Categorization used to label the noises.

Table 3 shows that most of the speech segments corresponds to not superposed voices.

| Speech | Duration | Percentage | Seg |
|---|---|---|---|
| One personne | 315.4 s. | 85.6 % | 190 |
| Several personnes | 8.7 s. | 2.4 % | 14 |
| Song | 44.3 s. | 12.0 % | 11 |

**Table 3** : Categorization used to label the speech.

The speaker recognition techniques need three consecutive seconds of speech to be efficient. We note in the following table that we don't have this necessary duration.

| Speakers | Duration | Perc. | Seg. | Seg.>3 s. |
|---|---|---|---|---|
| Kourou staff | 3.4 s. | 1.0 % | 3 | 0 |
| Mimi-Siku | 4.8 s. | 1.5 % | 2 | 1 |
| Stéphane | 149.4 s. | 47.4 % | 98 | 9 |
| Computer | 2.5 s. | 0.8 % | 2 | 0 |
| Passenger | 0.5 s. | 0.2 % | 1 | 0 |
| Counsel | 53.2 s. | 16.9 % | 28 | 5 |
| Charlotte | 2.7 s. | 4.0 % | 7 | 1 |
| Employee | 4.1 s. | 1.3 % | 5 | 0 |
| Richard | 10.3 s. | 3.3 % | 4 | 1 |
| Boatman | 1.5 s. | 0.5 % | 1 | 0 |
| Indian | 5.6 s. | 2.1 % | 2 | 0 |
| Indian child. | 20.9 s. | 6.6 % | 1 | 2 |
| Patricia | 45.6 s. | 14.5 % | 26 | 3 |

**Table 4** : Speakers of the database.

We have defined sixteen kinds of boundaries. Two of them are presented in the following figure (i.e., Music/Noise, Noise/Speech).

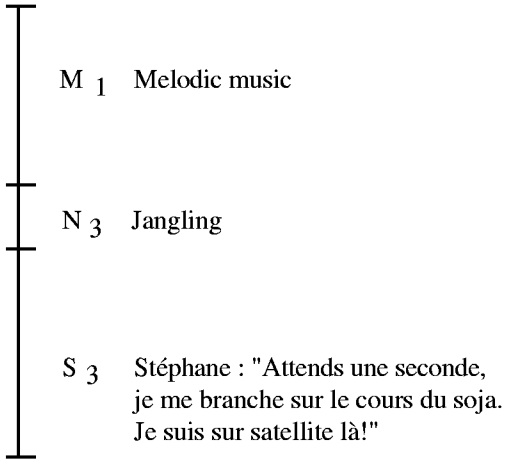


**Figure 2** : Example of scenario : Music/Noise/Speech

### 2.4. Results

The recognition boundary rate (R) is equal to the identification boundary rate (Id) minus the insertion boundary rate. The segmentation method is tested on the TIMIT speech database. For the 200 speakers separations, the error rate is about 9 %.

| | TIMIT database | | | |
|---|---|---|---|---|
| tolerance location | Id. | Del. | Ins. | Rec. |
| 0.3 sec. | 91.7 % | 8.3 % | 10.8 % | 80.9 % |
| 0.4 sec. | 96.1 % | 3.9 % | 6.5 % | 89.6 % |
| 0.5 sec. | 96.7 % | 3.3 % | 5.9 % | 90.8 % |

**Table 5** : Segmentation results on the TIMIT database.

The same experiment is carried out on the movie database, without any modification of the segmentation method.

| | Video movie database | | | |
|---|---|---|---|---|
| tolerance location | Id. | Del. | Ins. | Rec. |
| 0.3 sec. | 36.3 % | 63.7 % | 28.8 % | 7.5 % |
| 0.4 sec. | 40.1 % | 59.9 % | 25.0 % | 15.1 % |
| 0.5 sec. | 43.0 % | 57.0 % | 22.0 % | 21.0 % |

**Table 6** : Segmentation results on the video movie.

The high variability of the sounds and the low segments duration explain the increase of the error rate comparatively to the previous experiment. We detailed this result for Non-Speech/Speech, Speech/Non-Speech, Speech/Speech and Non-Speech/Non-Speech boundaries.

| | Movie database | | | |
|---|---|---|---|---|
| tolerance location | Id. | Del. | Ins. | Rec. |
| NS/S | 35.3 % | 64.7 % | 26.7 % | 8.6 % |
| S/NS | 53.0 % | 47.0 % | 17.5 % | 35.5 % |
| S/S | 44.8 % | 55.2 % | 20.9 % | 23.9 % |
| NS/NS | 44.0 % | 66.0 % | 24.0 % | 20.0 % |

**Table 7** : Segmentation results for different boundaries (0.5 sec. tolerance location)

The Speech/Non-Speech boundaries are better recognized than the Non-Speech/Speech boundaries. This difference results from the low recognition of the Music/Speech boundaries. A possible method to solve this problem would be to use a backward spectral changes detection or another unsupervised segmentation techniques. But the complexity of the movie description (Table 1-4) shows that it is essential to use knowledge on the movie.

## 3. SUPERVISED TECHNIQUES

When the script mapping is available, it is possible to use the methods developed for the automatic phonetic labeling of the speech database [11]. The difficulties to adapt these methods to the video movie audio data indexing are the duration and the superposition of the noise and music. In this preliminary experiment, the music and specific noises (cf. Table 2) superposition is not tested.

### 3.1. Script Warping

The warping technique is based on the HMM technique. A network warping is made from the phonetic transcription of the words script. The phonetic dictionary [12] allows word phonetic variants as missing elisions and liaisons (cf. fig. 3). 37 phonetic models including silence represented by 3-states Bakis models are used. Gaussian mixtures represent the HMM states distributions. These mixtures are first trained [13, 14] on a reference database [15], then the audio channel is segmented by the Viterbi decoding algorithm. For each phonetic segment, a segment likelihood coefficient is computed from the segment duration and the segment probability. The mixtures are trained again on the well warped phonetic segment (i.e., segment likelihood coefficient lower than a given threshold). This procedure is iterated until no segmentation difference is observed. Word segmentation is built from the phonetic segmentation using dynamic programming algorithm.

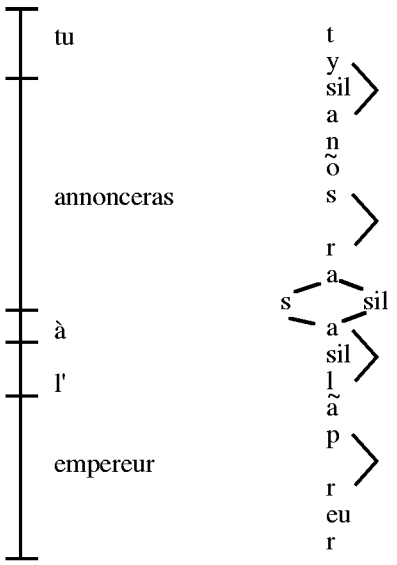


**Figure 3** : Example of warping network

### 3.2. DataBase and Experiments

Twenty minutes of a cartoon movie "Ivanhoe" are chosen as database. There are about 15 speakers and 2,608 occurrences of words (i.e., 747 different words). The warping network have 12232 nodes and 16,705 arcs. Three decoding/training sessions have been necessary. The segmentation computational cost of one iteration is about one day (Pentium Pro). The segmentation assessment have been performed by listening the result of each segmentation. Four kinds of errors have been defined for each word segment: Phone Shifting segmentation, Word Shifting segmentation (excluding one phone word) , Multiple Words Shifting segmentation and Sentence Shifting segmentation.

| Cartoon movie database | | | | |
|---|---|---|---|---|
| PS Err. | WS Err. | MWS Err. | SS Err. | Rec. |
| 4.1% | 3.2% | 1.5% | 0% | 91.2% |

**Table 8** : Segmentation results on the cartoon movie database.

The segmentation results of script warping are very good. There is no Sentence Shifting segmentation. But

the segmentation error should be increased with the duration. This technique will be used to index an 2 hours long entire movie film "Contes de Printemps" of Rohmer .Contextual phonetic models will be used to improve segmentation quality.

## 4. CONCLUSIONS

Two techniques for indexing audio data of the video movie are presented. The first one is based on speaker recognition, the second one on the speech recognition. Preliminary results on two video movies have been presented. On the first video movie, the high variability of noises and the poor sound quality of the CD-I, increase the difficulty of the audio channel indexing by speaker recognition separation technique. The second video movie is characterized by caricatural voices very unusual in the well known speech databases. On this video, the knowledge of the audio data script has allowed an audio time relocation using a warping network with good results, making the supervised technique quite usable for indexing.
Using this word relocation and speakers separation to help the segmentation of the image channel video movie in scenes and shots [16] will be studied.

## ACKNOWLEDGMENTS

We are grateful to the french Institut National de l'Audiovisuel (INA) for allowing us to use the cartoon movie "Ivanhoe" in our experiments.

## 5. REFERENCES

[1] Man-Hung Siu, George Yu, and Herbert Gish, "An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers", ICASSP, vol. II, pp. 189-192, 1992.
[2] George Yu and Herbert Gish, "Identification of speakers engaged in dialog", ICASSP, vol. II, pp. 383-386, 1993.
[3] M. Sugiyama, J. Murakami, and H. Watanabe, "Speech Segmentation and Clustering Based on Speaker Features", ICASSP, vol. II, pp. 395-398, 1993.
[4] Lynn Wilcox, Francine Chen, Don Kimber, and Vijay Balasubramanian, "Segmentation of speech using speaker identification", ICASSP, vol. I, pp. 161-164, 1994.
[5] Jesper Ø. Olsen, "Separation of speakers in audio data", Eurospeech 95, pp. 355-358, 1995.
[6] Claude Montacié, Jean-Luc Le Floch, and Marie-José Caraty, "Procédé et dispositif d'un contrôle d'accès par la voix". European patent application, 1996.
[7] Claude Montacié, Paul Deléglise, Frédéric Bimbot, and Marie-José Caraty, "Cinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction", ICASSP, 1992.
[8] F. Itakura : Minimum Prediction Residual Principle Applied to Speech Recognition. IEEE Trans. ASSP, Vol. 23, pp. 67-72, 1975.
[9] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals", IEEE Trans. ASSP, Vol. 36, no. 1, pp. 29-40, 1988.
[10] W. Fisher, V. Zue, J. Bernstein, D. Pallet "An Acoustic-Phonetic Data Base" J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.
[11] B. Weatley, G. Doddington, C. Hemphill, J. Godfrey, E. Holliman, J. McDaniel, and D. Fisher, "Robust Automatic Time Alignement of Ortographic Transcription with Unconstrainded Speech", ICASSP, 1992.
[12] Marie-José Caraty, Claude Montacié, Fabrice Lefèvre, "Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation", Eurospeech , 1997.
[13] S.J. Young, "HTK Version 1.4: Reference Manual and User Manual", Cambridge University Engineering Department - Speech Group, 1992.
[14] Claude Barras, Marie-José Caraty and Claude Montacié, "Temporal Control and Training Selection for HMM-based System", Eurospeech, 1995.
[15] Jean-Luc Gauvain, Lori F. Lamel and Maxine Eskénazi, "Design Considerations and Text Selection for BREF, a large French read-speech corpus", ICSLP, 1990.
[16] Pascal Faudemay, Liming Chen, Claude Montacié, Marie-José Caraty, Christine Maloigne, XiaoWei Tu, Mohsen Ardebilian ,Jean-Luc Le Floch, "Multi Channel Video Segmentation", SPIE, 1996.