Automatic Discrimination in Audio Documents

Chabane Djeraba, Hakim Saadane

IRIN, Nantes University,

2, rue de la Houssinière, BP 92208 - 44322 Nantes Cedex 3, France

☎ *: 33.2.40.37.49.01 -* ***Fax*** *: 33.2.40.37.49.70*

E-mail : *djeraba@irin.univ-nantes.fr*

Abstract

In this paper, we present a content-based classification approach for audio indexing. The classification is based on low level audio features such as temporal sound, energy, fundamental frequency, zero crossing rate, auto correlation curve, and based on transformations such as Short Time Fourier Transform (STFT). The audio classification in classes such as music, noise, silence, and speech is an efficient indexing method, because it permits efficient searches, by limiting the searches in the suitable classes.

Keywords

Audio retrieval, class hierarchy, audio databases, noise, music, speech, silence, indexing, classification.

1. Introduction

Audio invades lots of fields such as video indexing [Min 98], [Liu 97] audio production [Pai 97], broadcast [Gau 00], [Hu 98], etc. In video indexing, the indexed video segments are correlated to music and speech detection. The indexed video segments, when presented on the Video Sound Browser, let users randomly access the video. In audio production, lots of tools exist in the market, such as Lytha Studios-DMD, Making waves and SpyderNet, however few of them use content-based audio retrieval in the production process. In the broadcast domain, indexing and speech transcription techniques have been proposed for content exploitation and power conservation on mobile computers and conventional architectures.

The audio data explosion, like images and videos, is amplified by Internet phenomenon where plethora of audio, are placed at disposal of both business organisms and customers. This explosion is explained [Sti 95] by the fact that humans obtain 11% of their knowledge by hearing, 80% by sight. Furthermore, it puts in evidence that we hold back 20% of what we see, 30% of what we hear and 50% of what we simultaneously see and hear. So, owing to their capabilities to present simultaneously audio, and other medium, to users, multimedia systems will constitute a suitable tool for having access to knowledge.

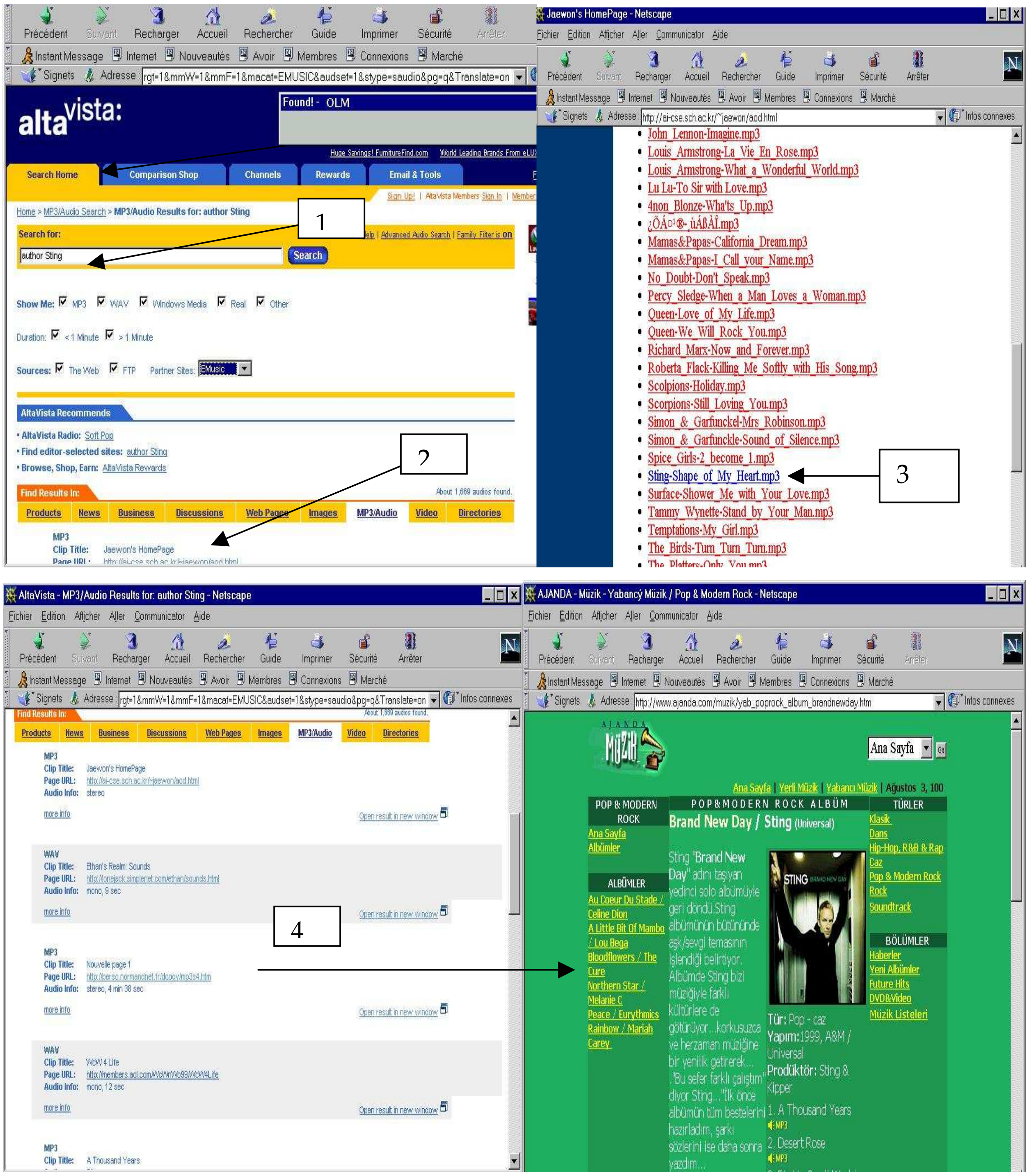

Figure 1: Example of the audio web retrieval engine AltaVista that illustrates the first category of audio retrieval systems (annotation-based retrieval). The arrow 1 points the key word "Sting" of the query. The user asks, "Search audio files that contain Sting songs". The arrow 2 defines the first result of the search engine. The results are contained in web pages. The result is a web page that contains audio files. The arrow 3 defines the name of the audio file. The name is composed of the string "Sting". However, the audio file is not a song of Sting. So the first result returned is not pertinent for the user. The arrow 4 points the fifth result of the query. It is a web page that contains audio files of the singer Sting. It is the first pertinent result.

The developments of content-based retrieval techniques in audio databases are just emerging, and may be grouped in two large categories: annotation-based retrieval and content-based retrieval.

In the first category, the annotation of the content is based on high-level descriptions such as texts and key words supplied by the authors as a basis for indexing and retrieval. The first generation is linked to the well-known concept of "information retrieval" [Moo 51], [Rij 79], [Sal 83] that supports the mental aspect of information description and its specifications for retrieval and whatever techniques that are employed to carry out the operations. They use essentially best-match information retrieval techniques to search a ranked list of relevant documents based, partly, on user's free-text queries.

In this category of systems, we can find Aster [Ram 94] in which text is associated to audio data. We can find, too, IBM-DB2 that is coupled to audio system that allows leaving voice messages on answering machines. DB2 allows audio query based on textual descriptions such as name, description, format, last update, etc. The Internet retrieval engines, such as AltaVista, illustrate well this category of techniques (figure 1).

The principal advantages is that high level annotations are useful, as it describes the secret semantic of the audio data the user looks for. The content description is semantically powerful, because free text is one of the best ways to describe the content naturally and semantically. Their principal disadvantages include the following points: - they do not allow queries based directly on the audio properties, such as melody of music, and - they are dependent on the particular vocabulary used. So the quality of the text-based retrieval depends on the quality of the manual description. Generally, manual descriptions tend to be incomplete and inconsistent, so the text descriptions do not, generally, allow for unanticipated search in subsequent retrievals. Finally, - the annotation creation is time consuming.

In the second category, the descriptions concern low level features, such as pitch, loudness, brightness, bandwidth and harmony. The low level features are extracted and represented automatically. The second category of systems is recent; they started, generally, in the mid-90s. We can find systems such as Muscle Fish [Wol 96], which reduce sounds to perceptual and acoustical features. This lets users find sounds by any one feature or a combination of them, by specifying previously learned classes based on these features, or by selecting reference sounds and asking the engine to retrieve similar or dissimilar sounds. We can find, too, systems, such as in [Fei 94], that use learning mechanisms in order to group together similar audio data. Other systems are limited to music such as in [Fos 82] in which the music is classified in learned classes.

In the second category, we distinguish four basic schemas of indexing:

Indexing based on statistic methods such as variance; zero crossing rate; mean, autocorrelation; etc.

Indexing based on acoustical and perceptual features such as pitch, loudness, sharpness, fluctuation, rugosity, harmony, rough, etc.

Indexing based perceptual features such as harmony, rough, etc. This schema is better to the first one, because it is more resistant to noisy data.

Indexing based on transformation such as Fourier transform, Discrete Cosine Transform (DCT), Wavelet transforms (Haar [Sub 98], Daubechie) and Hamard Transform. It is based, too, on measures linked to such transformations such as fundamental frequency and bandwidth. DCT transform presents good results, however the data indexed are considered as a whole data. Wavelet transforms present better results when searching specific audio features contained in audio data. Wavelet transform replaces the sinusoids of Fourier transformation by a family of wavelet translations and dilations.

In the second category of systems, it is not simple to compare their performance accurately, because there are weak test platforms, compared to speech systems such as speech analysis and synthesis [Ata 71], automatic speech recognition [Bah 83], automatic speaker recognition [Fur 81], [Del 99] which are less recent.

Our approach belongs to the second category of systems, and combine the first and fourth schema of indexing, with the particularity of automatic classification of audio data in semantic classes (silence, noise, speech, music) without a-priori training step of each class. The semantic nature of the indexing contributes to resolve the famous problem of automatic extraction of associations between high level (audio content description, language, instrument, source of music, type of music, interpreter, composer, author, identity of speaker, genre of speaker) and low level (fundamental frequency, direction, source of elevation, distance, tempo, loudness, pitch, sharpness, fluctuation, rugosity, bandwidth) descriptions in MPEG7 [Mpe 99] working groups.

The approach is illustrated by a prototype that is composed of two important parts, indexing and retrieval. If the retrieval process is the most visible part, in user point of view, indexing, based mainly on automatic classification, of audio data is needed to make the retrieval process effective. Our work deals with the indexing part on the basis of the following points:

Automatic extraction and analysis of input audio data.

Automatic classification of input audio data in large semantic categories (Music, Speech, Noise, Silence) with no a priori knowledge and training.

Development of suited data structures (Meta data) for retrieval methods. Ideally, the system should respect Multimedia Content Description Interface (MPEG7) orientations. However, it is too early to satisfy this objective, and it is difficult to say that our descriptors, description schemes, data description languages satisfy the future standard. Our current politic is to contribute and to follow carefully the different points that emerge from the current works on MPEG7.

To describe this work, section 2 presents the hierarchy of audio classes, and the description of the extractions techniques that discriminate the different classes. Section 3 deals with classification and retrieval approaches. Section 4 concludes the paper, summarizes our approach, and presents the perspectives.

2. Framework of semantic classes

There are four basic classes of audio data: Noise, Music, Speech and Silence. Generally, these classes suffice for most of sound classification needs, even if there are times when the basic categories need to be expanded.

We organize the basic categories into a hierarchy logical structure and write functions to manipulate the class instances (create, delete, update, etc.).

The sound database designer defines the Sound class. In the class, there is ability to hide some of its members from the rest of the system. We build a sound class, because we write programs that deal with sounds of one kind to another, and the sound is a basic unit that system must deal with. Any sound entity that the system might process is a candidate to be a sound class. Four basic sound classes are derived from the "sound" class (root): Noise, Music, Speech and Silence classes.

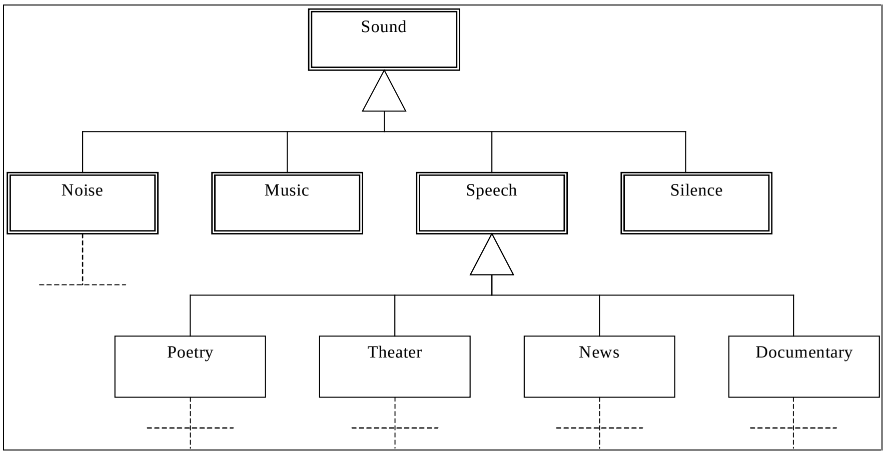

Figure 2: Audio class hierarchy. The underlined rectangles correspond to classes that support the automatic classification. The rest classes support manual classification.

In unified modeling language and prototype development point of views, the class from which we derive is called the base class or super-class and the new classes are called the derived classes or sub-classes. A sound derived class such as Music can be a base class to a lower derived class as well, and may have many derived classes, such as classic, rock, tango, traditional, samba, etc. One derived class (e.g. song) can have many base classes (music, speech). Each of which can have base classes of its own. This forms, a class network. The derived class inherits all the characteristics of its base, e.g. rock, classic, traditional music inherits all the characteristics of Music class. The derived class processes automatically the data members and member functions of the base. The derived classes contain their own members, but the base members become part of the derived class. There are three reasons to derive a class such as Music, Speech, Silence, Noise, etc from Sound. The first is that we modify the extraction and classification process of the existing class. The second is that we are building a well-organized, object-oriented class hierarchy. Finally, there is an inclusion relationship between a base and the derived class, e.g. between speech and News, News are a subset of speech class. These three reasons permit natural representations and design approaches.

We will see, in the following sections, how audio instances are classified automatically?

2.1 Silence Description

A simple method consists of considering the energy measure "em" of the temporal data. If "em" is greater than a certain predefined threshold then it is not a silence. The advantage of such solution is its implementation simplicity; the disadvantage is that in real applications, the measure of energy depends of the sound recording conditions. A bad recording hardware may perturb the silence energy data. This parameter is interesting, but it is not suited alone to silence measure.

A second approach consists of comparing the analyzed data and the reference sample of silence. This comparison is based on distance measure between the input data and the reference sample. The advantage of such approach is its efficiency in a short-term duration. Its disadvantage is that it is not suited for data in which there is an evolution of sound data, for example, there is a variation of the background sound.

Finally, the last approach is based on an adaptive approach. This adaptive approach may be seen as a combination of the previous approaches. In a short period, the index of silence is firstly fixed, after silence detection this level is evolved in the basis of the new measure in the window. The index of silence is then detected dynamically with a strong dependence of the short period analysis. This evolving approach is called Self-Normalizing [Sou 83]. It is a statistical approach to design an adaptive self-normalizing silence index. This approach is implemented in our system.

It is not possible to analyze sounds without analyzing the silence. In our world, there is no a real silence, however we can find some thing that tend towards the silence. This notion of silence depends of the heard capacity of each person. That is why it is more correct to say “low sound level” than “silence”.

2.2 Speech Description

There are two different kinds of voices, “voiced” and “unvoiced” sounds. Unvoiced sound has an energy concentrated in the high frequencies. This energy generates an important zero crossing. However, the “voiced” data have energy limited to 3KHz, they generate a weak rate zero crossing. So, globally, the speech has a high rate of cross. However, music is characterized by a concentration of the energy in the low frequencies, which is why its rate of zero crossing is generally low. This difference makes the automatic classification in speech/music classes possible.

Zero crossing rate is used to estimate the fundamental frequency of periodic sounds. There is a crossing rate by zero, when two successive ranges have opposite signs. The average of zero crossing is illustrated by:

$$Z(n)=\frac{1}{2}\sum_{m=-\infty}^{\infty}[\operatorname{sgn} x(n) - \operatorname{sgn} x(n-1)]w(n-m)$$

Where $w(n) = 1/N$ for $0 \le n \le N$ and $w(n) = 0$ when $n > N$, and $\operatorname{sgn}(x(m))=1$ if $x(m) \ge 0$, $\operatorname{sgn}(x(m))=-1$ when $x(m) < 0$

In the “voiced” class, the vibration of vocal cords creates sounds. Here, we have three basic classes of human sound: man, woman, child. The vocal cord vibration frequency determines these classes. This unit of measures is called fundamental frequency or “pitch”, and varies for :

- Man: between 80 and 150 Hz
- Woman: between 150 and 300 Hz
- Child: between 200 and 500 Hz

2.3 Music Description

Music is the result of man artistic activities and contains a variety of content according to different cultures, as speech. We limit our presentation to music discrimination features, and more particularly to note, harmonics, and octave.

The note describes the height and the duration of the sound. The height corresponds to the frequency of vibration. A sound is generally composed of a main frequency that corresponds to the height and several frequencies that are around the main frequency.

In a musical sound, these secondary frequencies are called harmonics of the main frequency. The harmonics correspond to multiple of the basic frequency. The distance from the basic frequency is proportional to harmonics. Higher is the distance lower is the harmonics.

2.4 Noise Description

Generally, we group inside this class of sounds "Noise" sounds that generate, generally harmful sounds. Furthermore, it is very different from other classes of sounds. Generally, the frontier between the silence and the noise levels is very fuzzy.

After silence extraction, the system extracts the non-silence sound, and more particularly the noise. The noise data is not periodic; the system computes the periodicity of the data based on short Term Autocorrelation measure. The function of Short Term Autocorrelation is used to determine when a data is periodic. At each instant n, it is defined by:

$$Rn(k) = \sum_{m=-\infty,+\infty} x(m)w(n-m)x(m+k)w(n+k-m)$$

"x" is the temporal data and "w" the window chosen. As properties of autocorrelation, Rn(0) is the maximum of the function. When x(m) is periodic with a P period then Rn(k) is periodic too with the same period. Hence, for periodic data with P period, the autocorrelation function tends to the maximum at ranges 0, ±P, ±2P, etc. So, we can estimate the period of a data by measuring the distance between two successive "maxima" of the autocorrelation function. The identification of such maxim determines the periodicity or not of such data. Its period corresponds to the distance between two peaks.

If the periodicity is not detected, because the distance between two peaks is not periodic, we try again with the double of the window 2*w. If there is no periodicity again, we conclude that the data x(n) is a noise in the window w. However, if the periodicity P between the peaks is detected in the second test, we consider that the periodicity is large, and consider that the data is music. When the periodicity in the first test is detected, the data may be music or a speech. We need more information to take a decision. In this case, another step is necessary to classify the sound in speech or music classes. The algorithm of zero cross rate is applied.

3. Classification method

In the following steps, we will see the different transformations of sound data from continuous data to classification that is based on the basic classes.

(a) Data discretization. The sound data is continuous, so the time interval of the data is valued by infinite values. To represent and manipulate this data in the computer, the continuous data has to be transformed into discrete data. The range becomes necessary to discrete the continuous data, and to represent it in a sequence of numeric values. Each value is called « range ». It codes the data magnitude with an instant on the time line. The range is characterized by a period of time. It corresponds to the duration between two ranges. The range frequency is the inverse of the period. If the range period is too long in order to economize data storage, then data between two ranges may be lost. If the range of a period is too short in order to keep all data, then the storage of the data will be very expensive.

Shannon theorem is suited to resolve such range problem. It presents the range frequencies that ensure data conservation when data maximum frequency is known. Here is the principle of Shannon theory. The lost data between continuous data and discrete data is null if and only if range frequency (fr) is at least greater or equal than the double of the highest frequency (fm) contained in the data. So, fm ≤ fr / 2. ». It is important to notice that it exists other methods of range extraction such as comb of Dirac.

(b) Detection of silence. It is a useful step, it contributes indirectly to extract semantic segment. Data segmentation (short term analysis). After data range, the data is segmented. The principle of the segmentation is to analyze fragments of data. The data is decomposed of short segments considered stationary; it means that there are few variations of the data in the duration associated to the segment. For each sound that starts and finishes with a silence, the system extracts a set of window segments w1, w2, …, wn, ... The segmentation process extracts homogeneous windows in which the window size is specified a priori. Mathematically, the short-term analysis is modeled by:

$$\mathrm{Qn}=\sum_{-\infty}^{+\infty} T[x(m)]w(n-m)$$

T[ ] is a transformation and w(n) the used window at the instant n. It exists various windows, rectangular (w(n) = 1 for 0 ≤ n ≤ N-1 and w(n) = 0 for n > N), Hamming (w(n) = 0,54 - 0,46 cos (2n / (N-1)) pour 0 ≤ n ≤ N-1 and w(n)=0 for n > N) Hanning, Blackman, etc. Each one has its advantages and disadvantages.

For each window, our method stores a vector of acoustical attributes. Each window is overlapped with both the precedent one and the next one. So we can see a sound in the object-oriented database as an instance of the Sound class. This instance consists of the sound source (e.g. name of the sound file), a set of windows, and statistical features, such as the average, variation and auto correlation. Each window is represented by an instance of the class Window. It contains the duration of the window, and the feature values associated with it, such as intensity, loudness, pitch, brightness, harmonic and timbre. We distinguish here the statistical features associated to a sound and acoustical (sound data) features associated to a window. If F features are extracted, then an audio data may be considered to be a sequence of M points (M is the number of windows) in a F-dimensional space. So we have statistical and data processing approaches to extract and represent the content of sounds.

(c) Based on the features extracted from the window, the system classifies the sound in the object-oriented model. This classification takes the form of instance creation in one of the semantic classes (Silence, Noise, Music, Speech).

```
{If (the audio data is not periodic)
{it is classified in the Noise class}
            Else  {If (the periodicity is undecidable)
      {the analysis is extended to the double of the current window size;
      If (the signal is not periodic)
      {the signal is classified in the Noise class;}
      Else  {the signal may be classified in Music or Speech class;}
            If (the zero crossing rate is high) {it is a speech;}
                                          Else { it is a music;}}}}}
```

4. Conclusion

We presented an approach that extracts automatically input audio data, and classifies it automatically in large semantic categories (Music, Speech, Noise, Silence) with no a priori knowledge and training. Suited data structures have been developed for Meta sound and retrieval methods. The data structures respects the future orientations of the Multimedia Content Description Interface (MPEG7).

The classification process presents interesting results, however more efforts should be done in specific classifications in semantic classes such as music classes (Rock, classic, techno, oriental, etc.), Noises (violence, water, cars, etc.), Speech (documentary, poetry, dialogues, etc.).

A perspective, we intend to use extensively wavelets in order to ameliorate the results obtained.

References

[Ata 71] B.S. Atal and S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, Journal of the Acoustical Society of America 50, 637-655, 1971.

[Bah 83] L.R. Bahl, F. Jelinek, and R.L. Mercer, A maximum likelihood approach to continuous speech recognition. IEEE Trans. Patt. Anal. Machine Intell. Pami-5, 179-190, 1983.

[Del 99] Delacour P., "Indexation de données audio : reconnaissance de la séquence de locuteur engagés dans une conversation", Thèse de L'ENST, departement Eurécom - 1999.

[Fei 94] Feiten B., S. Guznel “Automatic Indexing of a Sound Database Using Self-Organizing Neural Nets”, Computer Music J., Vol. 18, N° 3, 1994, pages 53-65, 1994.

[Fos 82] S. Foster, W. Schloss, A. J. Rockmore “Towards an intelligent Editor of Digital Audio: Data Processing Methods”, Computer Music Journal, Vol. 6, N° 1, pages 42-51, 1982.

[Fur 81] S. Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. ASSP, ASSP-29, 254-272, 1981.

[Gau 00] Jean-Luc Gauvain, Lori Lamel and Gilles Adda, “Transcribing broadcast news for audio and video indexing, Communications of the ACM”, Volume 43 , Issue 2, 2000.

[Haw 93] Michael Hawley. Structure out of Sound. PhD thesis, MIT Media Laboratory, 1993.

[Hu 98] Qinglong Hu, Dik Lun Lee, Wang-Chien Lee, “A Comparision of Indexing Methods for Data Broadcast on the Air”, Proceedings of the 13th International Conference on Information Networking (ICOIN '98), Tokyo, Japan, January 21-23, 1998.

[Liu 97] Liu Z. & al., “audio feature extraction & Analysis for Scene classification”, IEEE data processing society 1997 workshop on Multimedia data processing, june 23-25, 1997, princeton, new jersey, USA.

[Min 98] Kenichi Minami, Akihito Akutsu, Hiroshi Hamada, and Yoshinobu Tonomura, “Video Handling based on Music and Speech Detection”, IEEE MultiMedia, Vol. 5, No. 3, July-September 1998.

[Moo 51] Moores C. N. “Datacoding applied to mechanical organization of knowledge” AM. Doc. 2 (1951), 20-32.

[Mpe 99] ”MPEG-7 Requirements Document”, Doc. ISO/MPEG N2727, MPEG Seoul Meeting, March 1999

[Pai 97] Wan-Chieh Pai and Peter C. Doerschuk, "Data Processing Using Statistical Nonlinear Speech Production Models", 1997 IEEE Workshop on Nonlinear Data and Image Processing, September 8-10, Grand Hotel on Mackinac Island, Michigan USA, 1997.

[Rag 89] Raghavan, V., Jung, G., and Bollman, P., “A Critical Investigation of Recall and Precision as Measures”, ACM Transactions on Information Systems 7(3), page 205-229, 1989.

[Ram 94] T. V. Raman “Audio system for Technical Readings”. Doctoral dissertation, Cornell University.

[Rij 79] C. J. Keith van Rijsbergen “Information retrieval”, Second edition, London: Butterworths, 1979.

[Sal 83] Salton and McGill, 1983. “Introduction to Modern Information Retrieval”, McGraw-Hill, New York, NY.

[Sau 96] John Saunders. Real time discrimination of broadcast speech/music. In Proc. 1996 IEEE ICASSP, pages 993-996, Mars 1996.

[Sch 97] E. Scheirer et M. Slaney, “ construction and evaluation of a robust Multimedia speech/music discriminator ”, processing of the 1997 international conference on acoustic speech and data processing (ICASSP), Munich, Germany, April 21-24, 1997.

[Sou 83] P. D. Souza « A statistical approach to the design of an adaptative self normalizing silence detector » Proceeding of the IEEE, P678 -P684, vol. ASSP-31, N ° 3, June 1983.

[Sub 98] S.R. Subramanya, A. Youssef, “ Wavelet-based Indexing of Audio Data in audio/Multimedia Database ”, 4th Int. - Workshop on Multimedia DBMS, Dayton, Ohio, August 1998.

[Sti 95] Stiel N., “Multimedia: the new frontier”, Encyclopedia Universalis, 1995, pp. 144-149.

[Wol 96] E. Wold, T. Blum, D. Keislar, J. Wheaton « Content-based classification search and retrieval of audio », Proceeding of the IEEE, P27-P36, 1996.

[Pfe 96] Silvia Pfeiffer, Stephan Fischer, Wolfgang Effelsberg: Automatic Audio Content Analysis. ACM Multimedia 1996: 21-30.