Detection and Speech/Sound Segmentation in a Smart Room Environment

Michel Vacher, Dan Istrate, Jean-François Serignat, Nicolas Gac

DETECTION AND SPEECH/SOUND SEGMENTATION IN A SMART ROOM ENVIRONMENT

Michel VACHER, Dan ISTRATE, Jean-François SERIGNAT, Nicolas GAC

CLIPS - IMAG , Team GEOD
UMR CNRS-INPG-UJF 5524
385, rue de la Bibliothèque - BP 53, 38041 Grenoble cedex 9
France (Europe)
phone: +33 4 7663 5795, fax: +33 4 7663 5552
email: Michel.Vacher@imag.fr

Because of cost or convenience reasons, patients or elderly people would be hospitalized at home and smart information system would be needed in order to assist human operators. In this case, physiologic and position sensors give already numerous informations, but there are few studies for sound use in patient's habitation. However, sound classification and speech recognition may greatly increase the versatility of such a system: this will be provided by detecting specific sounds or short sentences which could characterize a distress situation for the patient. Sounds emitted in patient's habitation may be useful for patient's activity monitoring. The proposed sound analysis system is made of four modules: the first module in charge of sound and speech extraction is the detection module, it is followed by a segmentation module needed to transmit the extracted wave to the Sound Classification module or to the Speech Recognition module. The first two modules -Detection and Segmentation- are presented and evaluated in this paper in experimental recorded noise conditions. The detection method uses transient models, based upon dyadic trees of wavelet coefficients to insure short detection delay. The segmentation step is a classical Gaussian Mixture Model classifier based on acoustical parameters like MFCC.

Key words: Gaussian Mixture Model, Noise, Segmentation, Smart Room, Sound Extraction, Sound Classification, Wavelet Transform.

1 INTRODUCTION

In this paper a sound detection and speech/sound segmentation method is presented. This method has been developed as part of a medical telesurvey system intended for home hospitalization. The aim of this system is to detect a distress situation of the patient using sound analysis. In distress case a medical center is automatically called with the aim to give assistance to the patient. The decision of calling is taken by a data fusion system from smart sensors and particularly a sound system as explained in [1]. Others sensors give information about patient position (infrared sensors and door contacts) and state of health (oxymeter, tensiometer, thermometer and actimeter).

Each sound produced in the apartment is characteristic of:

- **a patient's activity:** the patient is locking the door, or he is walking in the bedroom,
- **the patient's physiology:** he his having a cough,
- **a possible distress situation for the patient:** a scream or a glass breaking are suddenly appearing.

If the system has a good ability of classification for such sounds [2], it will be feasible to know if the patient is needing help. Several usual sound classes needed for this application have been defined and a corpus has been recorded in our laboratory.

In the same way, the speech said by the patient may give precious information on the patient:

- **a distress case :** “Help me!”, “Doctor!”,
- **a normal state :** “Coffee is cold!”, “The door is open!”.

An adapted corpus in French has been defined and recorded in our laboratory. Before sound/speech segmentation, it is necessary in a first step to establish the start and the stop time of the sound to be classified in the environmental noise. The precision of these two times must be sufficient to allow the segmentation step good performances. In the context of audio signal encoding, the input signal can be decomposed into ``tonal", ``transient" and ``stochastic" components as described by Daudet in [3]; our problem is restricted to transient detection for which large wavelet coefficients are more easily interpreted as transients. The proposed method is based on wavelet tree analysis. In case of "transient", a significant coefficient is likely coming with additional significant coefficients of lower scale occurring at the same time.

Segmentation is frequently studied for speaker, speech/music or speech/music/singing segmentation in various noise recording conditions. A summary of obtained errorless segmentation rates is shown in Table 1.

Table 1 - Segmentation Results for various Class Type

Segmentation	Document Type	Segmentation Rate	References
Musical Genre	Music Recording	60-70%	[13][14]
Speech, Music, Singing	Music Recording	80-90%	[15]
Speech, Music, Silent	Radio Broadcast News	95-98%	[16][17]
Speech, Music, Silent, Speaker	Radio Broadcast News	80-90%	[18][19]
Speaker	Meeting Recording	80-90%	[18][19]
Speech, Music, Silent	Motion Picture	95-98%	[4]
Speech, Music, Noise, Silent	Motion Picture	80-90%	[6]
Whistle, Crowd (crapping, laughing), Speaker	Football Matchs	80-90%	[20]
Man, Woman, Music Instrument, Machine, Water, Animals	Various Kind of Records	85-90%	[26]
Life Sound Classification	Home Recording	60-90%	[2][5]

Segmentation methods are various, including Hidden Markov Model, KullBack-Leibler distance, Bayesian Information Criterion, Artificial Neural Network, Decision Rules and Gaussian Mixture Models. Results in case of speech/music discrimination are very good (98%), but they decrease of 10% if noise has to be discriminated. For musical genre results are poor and below 70%.

The proposed speech/sound segmentation method is a classification method using GMM models [4][5][6], evaluation is done in noisy conditions with our corpus. We also present in this paper the results of sound detection and speech/sound segmentation system in noisy conditions on audio recordings[22].

2 THE SOUND SURVEILLANCE SYSTEM

2.1 The Telemonitoring System

The aim of our study is to obtain useful sound informations and to transmit them through network to a medical supervising application in a medical center. The habitat we used for experiments is a $30m^2$ apartment situated at the TIMC laboratory inside the Faculty of Medicine of Grenoble. It is equipped with various sensors, especially microphones in every room (hall, toilet, shower-room, living-room)[1]. The entire tele-monitoring system is composed of three computers which exchange information through local network (see Figure 1).

This system is designed for the surveillance of the elderly, convalescent or pregnant women. Its main goal is to detect serious accidents or falls or faintness at any place of the apartment. It was noted that the elderly have difficulties in accepting video camera monitoring, considering it a violation of their privacy. Thus, the originality of our approach consists in replacing video camera by a multi-channel sound acquisition system.

Each time a sound event is analysed, a message is sent to the Data Fusion PC, notifying occurrence time of detection, most probable sound class or recognized sentence, localization of the emitting source. From this and from other data obtained from localisation and physical sensors, the Data Fusion PC could send an alarm if necessary.

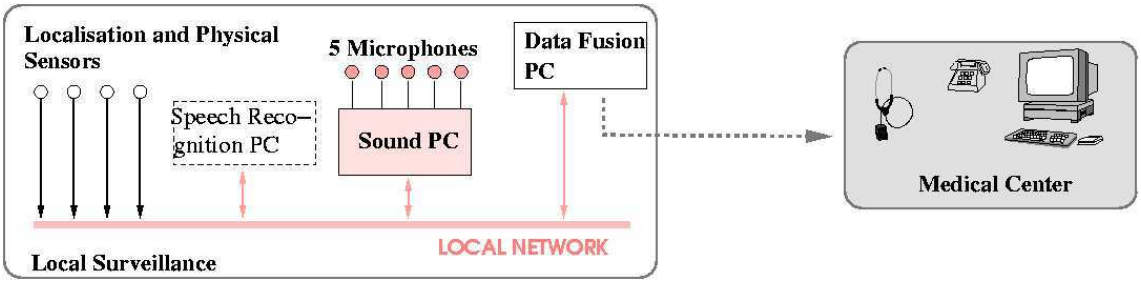

Figure 1 - Medical Telemonitoring System

2.2 The Sound Surveillance Architecture

The sound analysis system has been divided in four modules as shown in Figure 2. The first module is the detection module in charge of extraction of audio events from the signal flow. Extracted signals are then transmitted to the segmentation stage, which switches it to the classification module in case of life sounds or to the recognition module in case of speech. At the end, the obtained information will be send to the data fusion system, which will respond at the question: "Is it a normal case or a distress situation?"

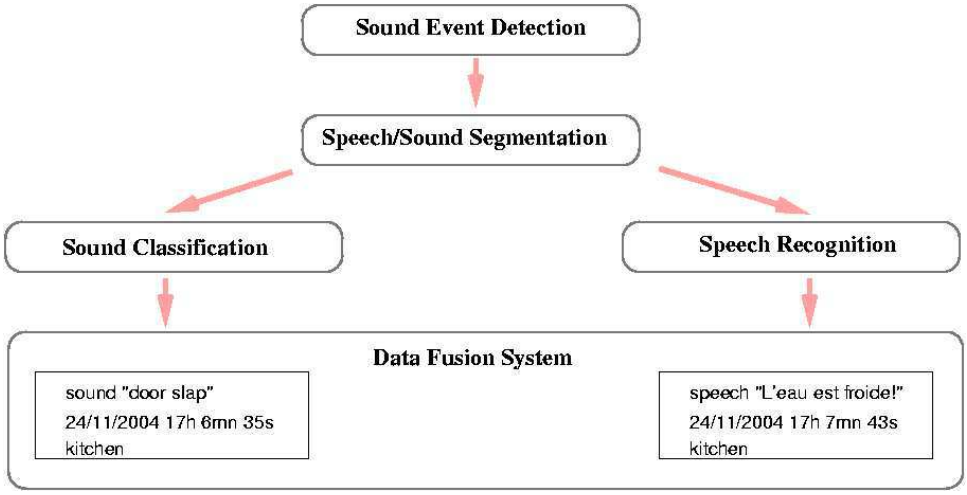

Figure 2 - Sound analysis system

3 SOUND AND SPEECH CORPUS

3.1 The sound Corpus

The everyday sounds are divided into 7 classes. The criteria used for this repartition were: statistical probability of occurrence in everyday life, sounds significant for a distress situation (scream, person fall) and duration of the sound (significant sounds are considered to be short and impulsive). The 7 sound classes are related to 2 categories:

- **normal** sounds related to a usual activity of the patient (door clapping, phone ringing, step sound, human sounds like cough or sneeze, dishes sound, door lock),
- **abnormal** sounds that generate an alarm (breaking glasses, screams, fall sounds).

As no everyday sound database was available in the scientific area, we have recorded a sound corpus. This corpus contains recordings made at the CLIPS laboratory, files of "Sound Scene Database in Real

Acoustical Environment" (RCWP Japan) and files from a commercial CD: door slap, chair, step, electric shaver, hairdryer, door lock, dishes, glass breaking, object fall, water, ringing. 20 types of sounds were selected with 10 to 300 repetitions per type.

Table 2 - Everyday sound corpus

Class of Sound	Number of Files	Duration Average	Class Length	% of the Corpus
Door Slap	523	737ms	385s	33%
Glass Breaking	88	861ms	76s	6%
Ringing Phone	517	928ms	480s	40%
Step Sound	13	2257ms	29s	2%
Screams	73	1930ms	141s	12%
Dishes Sounds	163	402ms	65s	6%
Door Lock	200	36ms	7s	1%
Entire Corpus	1577	751ms	20mn	100%

The sound classes of our corpus are described in Table 2. This corpus is not yet complete, 2 classes very useful for this application are remaining to record: Human Sounds and Fall Sounds.

3.2 The speech Corpus

This corpus has been recorded in the CLIPS laboratory by 21 speakers (11 men and 10 women) between 20 and 65 years old. It is composed of 126 sentences in French: 64 are characteristic of a normal situation for the patient: "*Bonjour!*" (Hello), "*Où est le sel*?" (Where is the salt)… and 64 are distress sentences: "*Au secours!*" (Help), "*Un médecin vite!*" (A doctor quick)… This corpus has a total duration of 38 minutes and is constituted by 2646 audio files.

3.3 The Noised Corpus

First investigations showed that white noise performances are not sufficient to insure satisfactory performances in real conditions. For this reason, we use audio noise recorded in our test apartment: **HIS noise**. It results of all noises in the building, it is a transient noise similar to usual sounds to detect, but transients are partially reduced by propagation inside the structure of the building. This kind of noise is not a stationary noise.

With this corpus, a noised corpus has been generated for 4 signal to noise ratios: SNR=0dB, SNR=+10dB, SNR=+20dB and SNR=+40dB. Used noise was recorded in our experimental apartment [1]. For each SNR, the noised corpus is made of 1577 life sound files (20mn) and 2646 speech files (38mn).

4 THE DETECTION MODULE

4.1 Transient Modeling

Methods based on wavelet transforms are often used for singularity characterization and for transient detection, because of the compact support of wavelets in conjunction of the dyadic properties of these transforms.. These two properties are allowing the analysis of reduced parts of the processing window. The Figure 3 shows a wavelet tree with 3 level depth beginning at the highest hierarchical level. Each node is corresponding to a wavelet whose support is drawn in frequency and time domain. For wavelets of highest level the support in time is twice the sampling period.

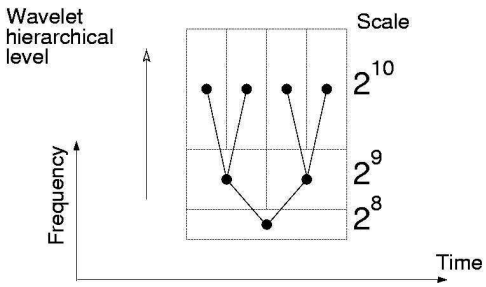

Figure 3 - Tree of wavelet coefficients for N=2048 sample window (tree depth of 3 levels)

For our purpose it is not necessary to determine the full tree corresponding to the transient, we limit our study to these 3 levels and we characterize each tree by his energy e, the sum of the energy of all nodes. We have chosen Daubechies wavelets Ψ with 6 vanishing moments to compute DWT on 2048 sample windows (128 ms), the wavelet basis is generated by translation ($-2^{j}n$ term) and scaling ($\frac{1}{2^{j}}$ and $\frac{1}{\sqrt{2^{j}}}$ factors) of the mother wavelet Ψ :

$$\left\{\Psi_{j,n}(t)=\frac{1}{\sqrt{2^{j}}}\Psi\left(\frac{t-2^{j}n}{2^{j}}\right)\right\}_{(j,n)\in\Re}$$

The mother wavelet is a function with finite energy and fast decay. The Figure 3 illustrates the variation of the support of wavelets of highest hierarchical level: the higher the coefficient level is, the more the support of the wavelet function is compact in time and large in frequency. This is valid for all levels.

As we consider the energy e of the tree, the non-significant nodes are implicitly not taken into account because they are negligible in the summation. With this approach the tree is not pruned and we don't eliminate nodes at scale 2^{10} if their mother node at scale 2^{9} is not significant, but this might not be very harmful because of the low depth of the tree.

A signal of a falling chair with HIS noise is drawn on the left of Figure 4. The sound appears at time t=10s. The right sub-figure displays tree energy evolution across the time. Energy corresponding to useful signal is surrounded by isolated noise pulses, which are sometimes greater, but useful signal is associated with numerous adjacent trees and in this way could be detected.

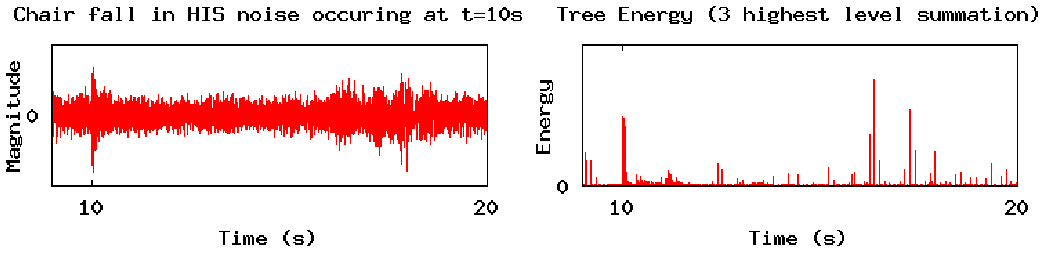

Figure 4 - Signal in HIS noise with corresponding tree energy over the time (tree depth of 3 levels)

4.2 Detection Algorithm

Evaluation of the detection algorithm was done from Receiver Operating Curves (ROC) giving *missed detection rate* (MDR) as a function of *false detection rate* (FDR), the Equal Error Rate (EER) being achieved when MDR=FDR. Results for the proposed algorithm are given in Table 3 in noisy conditions.

4.2.1 Start of sound

This algorithm is based on several wavelet tree means. DWT is calculated on N = 2048 sample windows (128ms) as shown in Figure 5. From this DWT the energy e_i of each tree is obtained by $500\mu s$ time translation across the transform. The processing window is cut into 4 consecutive frames containing 64 trees.

Thus at 16 kHz sampling rate, corresponding width for these 64 values is 32 ms. Energy $e=(\sum_{i=1}^{64} e_i)$ of each frame is calculated in order to suppress noise influence. A transient is characterized by a large increase of e.

The detection threshold th, which is applied on e, is adaptive: $th=\kappa+1.2\mu_{emeans}$, with μ_{emeans} referring to the mean of the thirty last values of e and κ to an adjusting parameter (see Table 3). The coefficient 1.2 was introduced because of remaining oscillations on e.

As soon as the threshold is overrun, the detection time or start of detected sound and the signal energy are memorized. This threshold will be used for end of signal detection.

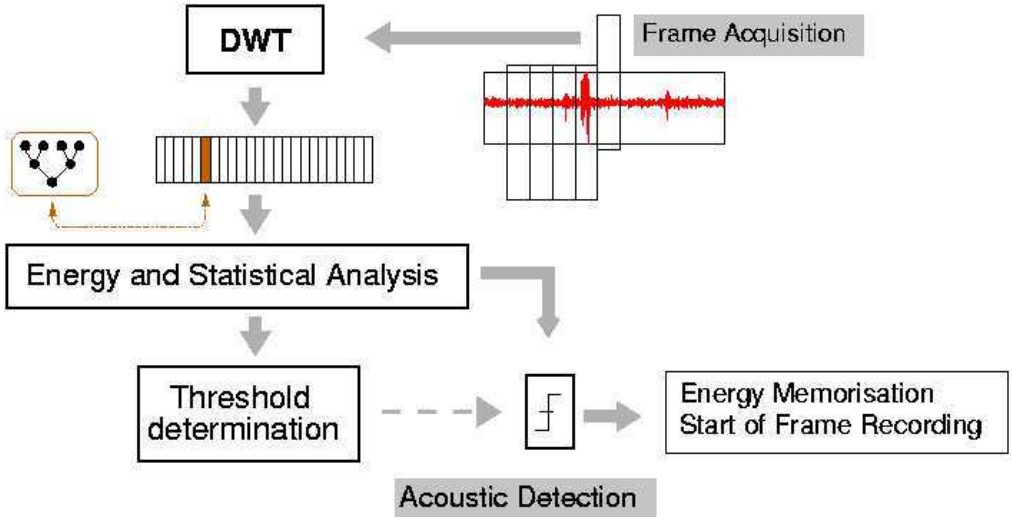

Figure 5 - Detection algorithm using energy tree evaluation

4.2.2 End of sound

As soon as the beginning of a sound is detected in previous step, incoming signal is recorded until the end of sound is detected. The energy value at detection time, corrected by a constant, is used as end threshold on the same wavelet signal energy. It is necessary to allow silence sequences until 384ms length (12 frames of 64 trees) to take into account word separation in case of speech: 12 consecutives frames below the threshold must be counted as shown in Figure 6. These 12 frames are not considered as a part of extracted signal unless tree energy of the following frame is above the threshold: in this case these frames are a silent part.

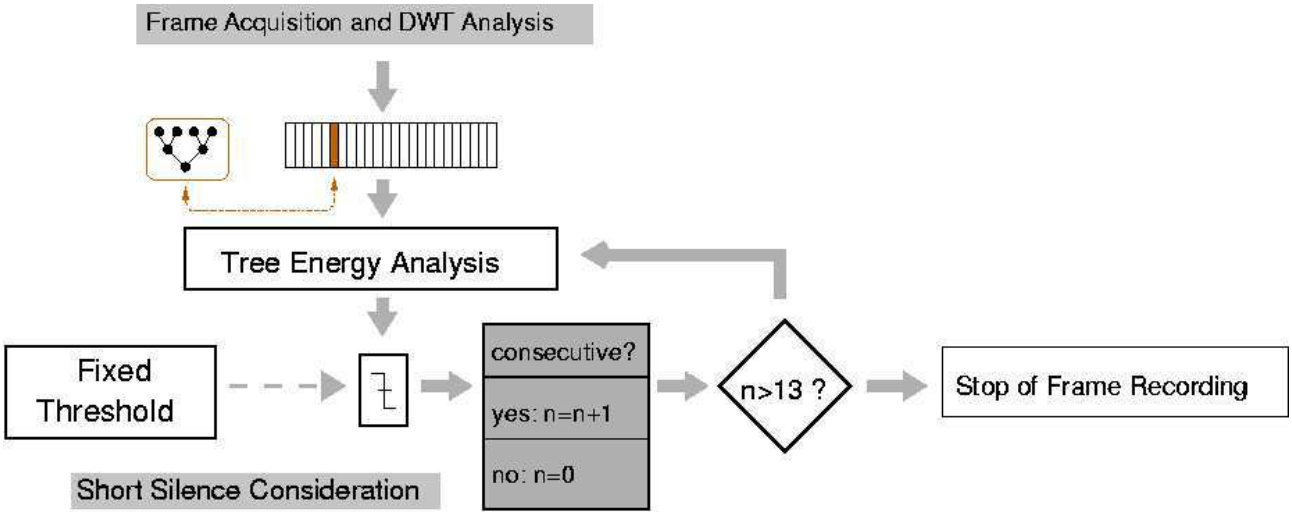

Figure 6 - End of signal determination

4.2.3 Sound extraction example

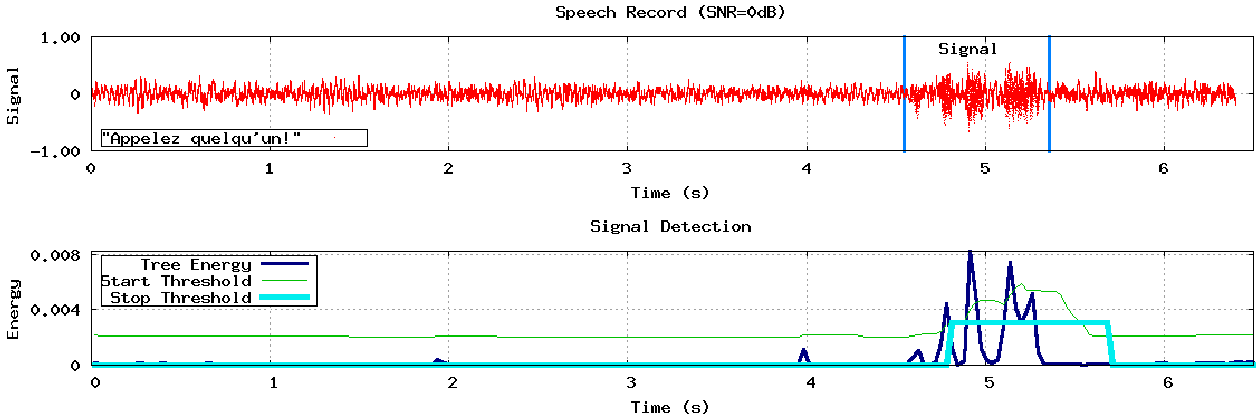

Figure 7- Extraction of a Door Slap

The top window of Figure 7 presents the sentence "Appelez quelqu'un" mixed with HIS noise at 0 dB of SNR. In the bottom window wavelet tree energy is represented in dark colour; the adaptative start threshold in grey color with finest line and, respectively, the end threshold with the largest line. We can observe that the start and end signal are detected with precision. According to Table 4 the precision of signal end is almost independent of SNR.

4.3 Detection Results

Evaluation of the algorithm was done from Receiver Operating Curves (ROC) giving *missed detection rate* (MDR) as function of *false detection rate* (FDR), the Equal Error Rate being achieved when MDR=FDR. Results are given in Table 3, κ is the coefficient used to fix detection threshold. Worst results (6.5%) are obtained for 0dB SNR and even in this case the number of false or missed detection will be low for a value of . κ=0.002.

Table 3 - Detection Results in noisy environment (HIS noise)

SNR [dB]	**EER** [%]	**κ**	**MDR** [%]	**FDR** [%]
0	6.5	0.002	2.8	12.9
10	0		0	0
20	0		0	0
40	0		0	0

In order to insure the best classification results, a short detection delay is very important. The precision of start and end signal detection, for short signals (signals with a length <2s),are given in Table 4 for each SNR in the previous conditions (threshold fixed to 0.002). We can observe that beginning and ending detection time precision is approximately independent of SNR.

Table 4 - Mean of signal extraction precision

SNR [dB]	**Global** [ms]		**Sounds** [ms]		**Speech** [ms]	
	Start	**Stop**	**Start**	**Stop**	**Start**	**Stop**
0	119	222	106	190	130	248
10	95	242	116	192	77	280
20	99	258	117	228	84	280
40	103	301	118	241	90	346

5 THE SEGMENTATION MODULE

We have used a Gaussian Mixture Model (GMM) in order to segment the sounds into speech and usual sounds [8][9]. There are other possibilities for classification: HMM, Bayesian method, etc. GMM has been chosen because it procures comparable performances and requires low processing time.

5.1 GMM Method

The classification with a GMM method supposes that the acoustical parameters repartition for a sound class may be modelled with a sum of Gaussian distributions. This method evolves in two steps: a training step and a classification step. The GMM evaluation has been done on the Elisa platform.

In the training step for the sound class and the speech class a Gaussian model is estimated, each model contains the characteristics of each Gaussian distribution. The number of distributions will be discussed later. The training step start with a K-Means algorithms followed by EM algorithm (Expectation-Maximization) in 20 steps.

In the classification step the likelihood for each sound class is calculated for each acoustical vector (or 16ms frame) of the detected signal. The global likelihood for each class is the geometrical average of all acoustical vector likelihood. The signal belongs to the sound class for which likelihood is maximum.
Since identification decision is made by comparison between average of all vector likelihood, a signal truncation is less important than an addition of noise vectors at the end of signal. This addition will alter average with noise likelihood in the same ratio of number of added vectors to number of original vectors.

5.2 Acoustical Parameters

GMM classification is not done directly on signal but uses extracted acoustical parameters before the training step and the classification step. Acoustical parameters are a synthetic representation of time signal.

Acoustical parameters classically used in speech/speaker recognition are: **MFCC**(Mel Frequencies Cepstral Coefficients), **LFCC** (Linear Frequencies Cepstral Coefficients), **LPC**(Linear Predictive Coefficients). MFCC were chosen because of their characteristics which are very similar to human hearing. MFCC parameters are cepstral coefficient frequently used un speech recognition. They allow deconvolution of exciting signal and conduit contribution. Obtained results are very good [11] and these features are used as reference in case of new parameter study [12][13].

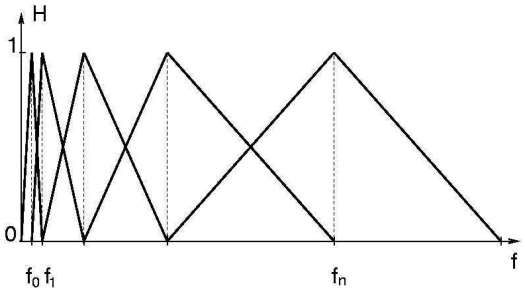

Figure 8 - Triangular MEL filter response

The calculus steps for the MFCC parameters are: pre-accentuation and windowing; Fast Fourier Transform of the analysis frame signal; Mel triangular filtering (see Figure 8); logarithm calculus of the filtered coefficients and inverse cosines transform. The Mel frequency scale is logarithmic: $f_{Mel}=2595.\log\left(1+\frac{f}{100}\right)$. The inverse cosinus transform is obtained according to: $c[n]=\sum_{m=0}^{M-1} E[m]\cos\left(\frac{\pi n\left(m-\frac{1}{2}\right)}{M}\right),\ 0\le n<M$.

Acoustical parameters used in speech/music/noise segmentation are : ZCR (zero crossing rate), RF (roll-off point), centroid. **Zero Crossing Rate (ZCR)** is the number of crossings on time-domain through zero-voltage within an analysis frame (see Figure 10(a)). **Roll-off Point (RF)** is the frequency, which is above 95% of the power spectrum (see Figure 10(b)). **Centroid** represents the balancing point of the spectral power distribution within a frame (see Figure 10(c)).

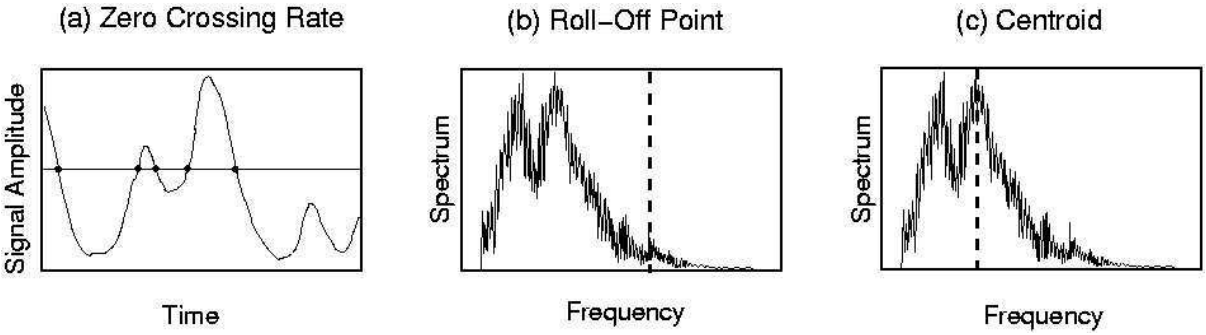

Figure 9 - Zero Crossing Rate (a), Roll Off Point (b) and Centroid (c)

Normalised energy of the frame is used as additional parameter. For each frame, the energy is normalised with the average of energies of the frames of the complete signal. This parameter is in this way less dependent of experimental recording conditions.

5.3 Model Selection

The Bayesian Information Criterion (BIC) is used in this paper in order to determinate the optimal number of Gaussian models. BIC criterion selects the model trough the maximization of integrated likelihood:

$BIC_{m,K} = -2L_{m,K} + \nu_{m,K}\ln(n)$. Where $L_{m,K}$ is logarithmic maximum of likelihood, equal to $\log f(x|m,K,\tilde{\theta})$ (f is integrated likelihood), m is the model and K the component number of the model, $\nu_{m,K}$ is the number of free parameters of model m and n is the number of frames. The minimum value of BIC indicates the best model.

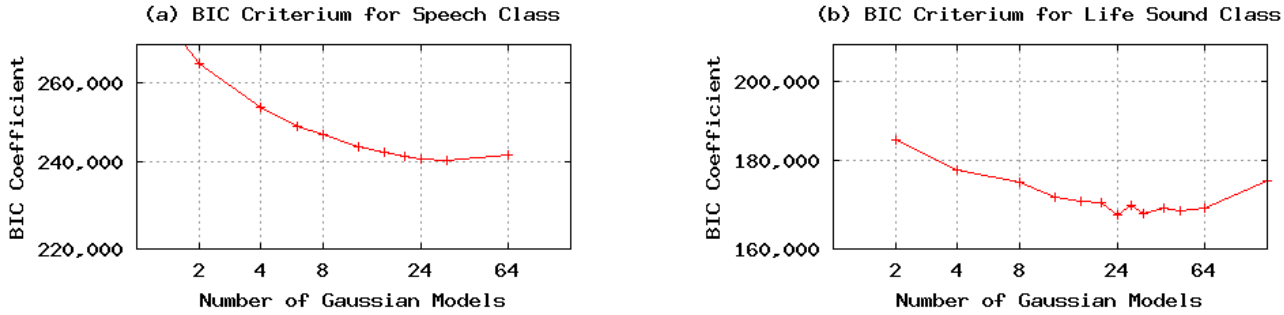

Figure 10 - BIC Coefficient Evolution for Speech (a) and Sounds (b) as a Function of Gaussian Model Number

The BIC criterion has been calculated for the sound class and for the speech class in noiseless conditions, for 2, 4… and 64 Gaussian models in case of 16 MFCC parameters in conjunction with Zero Crossing Rate, Roll-Off Point and Centroid. The results of the figure 11 are given for a number of Gaussian models between 2 and 64 in case of speech class (subfigure a) and sound class (subfigure b). Performances will be optimal when the BIC criterion is minimal. As it appears on these 2 curves, a number of Gaussian models between 20 and 32 seems to correspond to the best sound modelling. We have decided to use 24 Gaussian models, which may be a good compromise between segmentation performances and calculus consumption (real time constraints).

5.4 Segmentation Results in Noisy Conditions

The analysis window was set to 16ms with an overlap of 8ms. Usual length for speech analysis is 20 ms. With used sampling rate (16 kHz) the nearest value is 16 ms (2^8 samples). The GMM model is made of 24 Gaussian distributions for each class: life sound and speech. Training is made with pure sounds and testing with sounds mixed with HIS noise at 0, +10, +20 and +40dB. The test uses a "cross-validation" protocol: training is achieved with 80% of the corpus, and each file of the 20% remaining is tested according to the models. Firstly the full corpus is cut into five parts of same size (speech, each class of life sounds) and secondly the speech of each part is still divided in two parts to insure that no test will be done on a model trained with the same speaker or the same sentence (see Figure 11).

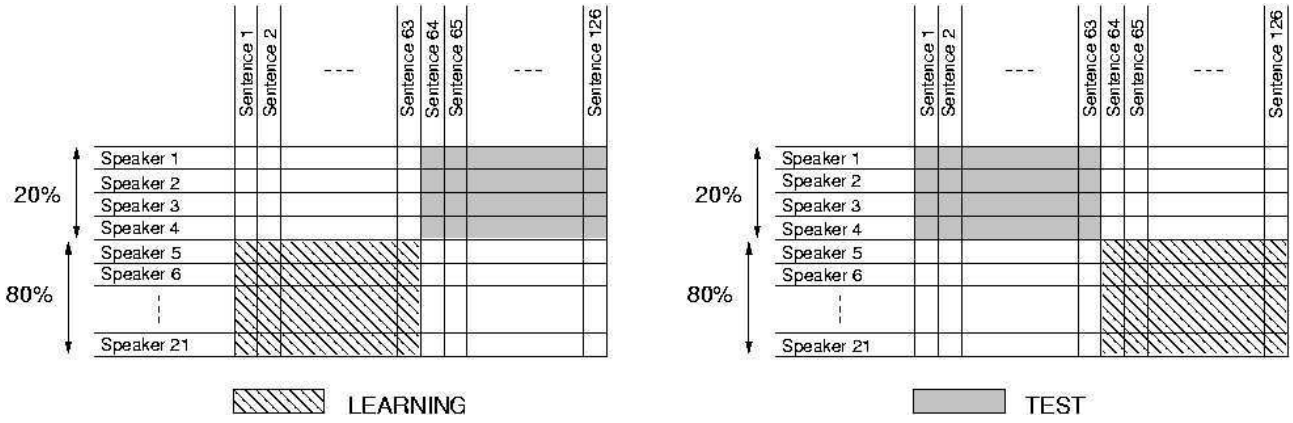

Figure 11 - Speech Corpus Partition in order to insure test independence from speaker and from sentence

The sound classification performances are evaluated through the error segmentation rate (ESR) which represents the ratio between the bad classified sounds and the total number of sounds to be classified. In Table 5, the classification results are presented for 16 MFCC acoustical parameters coupled or not with Zero-Crossing-Rate, Roll-Off-Point, Centroid and normalized energy. We can observe that in low noise conditions (+40dB), best results are achieved for MFCC coupled with normalized energy: ESR=4.5%. These results are remaining stable for $SNR \geq +10dB$, but they decrease below, ESR=22% at 0dB.

Table 5 - Segmentation Error Rate (24 Gaussian models)

SNR [dB]	**16 MFCC [%]**			**16MFCC + normalised energy [%]**			**16MFCC+ZCR+RF+Centroid [%]**		
	Global	**Speech**	**Sounds**	**Global**	**Speech**	**Sounds**	**Global**	**Speech**	**Sounds**
0	23.6	33.6	7.0	22.0	29,.9	8.6	58.1	90.5	3.8
10	6.0	3.3	10.6	4.2	5.9	3.1	7.5	10.3	2.7
20	5.1	1.5	11.3	4.4	9.1	1.6	5	2.8	8.5
40	5.0	1.4	11.2	4.5	9.6	1.4	6.1	2.7	11.9

ZCR and RF parameters are dependent of high frequency components of the signal and thus very noise sensitive. Results are poor at low SNR for that reason.

Normalised energy is processed on the complete sound file. In case of speech, normalisation is affected by silences between words. We can notice that except for 0 dB, results are better for 16 MFCC without normalised energy in case of speech: +8% at +20 and +40 dB. It's the other way round in case of sounds: +10% at +20 dB and +40 dB.

6 RESULTS WITH CRESSON CORPUS

It should be interesting to test this system with sounds recorded in real conditions. The CRESSON laboratory of the Architecture School of Grenoble has made such sounds [22] available for our study. These sounds have been recorded in yards or apartments of buildings located in the historical centre or new districts of the Grenoble town. Some extracts relevant with our study have been selected: speech, dialog, cough, fall, laugh, slap, steps, barking and birds. These sounds could be related to an activity inside of the apartment or outside if the door or a window is open. It should be pointed out that these sounds have been recorded with important noise and reverberation conditions.

Classification has been made in the previous conditions using GMM models obtained with 80% of our corpus (speech and sounds). Results with 32 selected files are shown in Table 6.

Table 6- Segmentation Error Rate for Cresson Corpus (24 Gaussian)

Type of Sound	16 MFCC	**16 MFCC + normalised energy**	16 MFCC + ZCR, RF, Centroid
Speech (14 files)	0	1 file	2 files
Sound (18 files)	6 files	3 files	7 files
Global (32 files)	19%	**16**%	28%

Zero Crossing Rate, Roll off Point and Centroid are not suitable for our application. Best results are obtained with MFCC and MFCC in conjunction with normalised energy. Speech is well classified in these two cases. Sounds like barking, bird singing, fall, laugh, street noise and tyres crunching are not represented in the training corpus. Classification errors occur in case of bird singing, tyres crunching and street noise, which are often classified as speech. All of these sounds present some similarities with speech: for example street noise comports some speech distorted by reverberation. Using a wider corpus may solve this problem.

7 CONCLUSIONS

Extraction method presented in this paper is allowing us to detect and segment between sounds and speech acoustical events recorded in a nursing home. An evaluation of the proposed detection method has been made on an adapted corpus in an experimental noisy environment. This method introduces a low delay after signal beginning –100ms typical - and acceptable end of signal truncation so that link to classification step is not disturbed. In the same way word of a same sentence are not separated if the silent duration is shorter than 385ms.

Detection is error-less for 10dB and upper and segmentation error below 5% is reached in the same conditions: according to these two results we can conclude that this detection/segmentation system may be used under realistic conditions with moderate noise.

We are working to implement these algorithms coupled with life sound classification and speech recognition in order to develop a complete acoustical analysis system.

ACKNOWLEDGEMENT

This study is a part of the DESDHIS-ACI "Technologies pour la Santé" project of the French Research Ministry. This project is a collaboration between the CLIPS ("Communication Langagière et Interaction Personne-Système") laboratory, in charge of the sound analysis , and the TIMC ("Techniques de l'Imagerie, de la Modélisation et de la Cognition") laboratory, in charge of the medical sensor analysis and data fusion. CLIPS and TIMC are 2 laboratories of IMAG ("Informatique et Mathématiques Appliquées de Grenoble") institute. IMAG has funded the implementation of the habitat used for our studies through the RESIDE-HIS project.

REFERENCES

1. VIRONE G., ISTRATE D., VACHER M., et all, *First Steps in Data Fusion between a Multichannel Audio Acquisition and Information System and an Information System for Home Healthcare*, IEEE Engineering in Medicine and Biology Society, IMBS'2003 Proceedings, pp. 1364-1367, Sept. 2003.
2. VACHER M., ISTRATE D., BESACIER L., SERIGNAT J.-F., CASTELLI E., *Sound Detection and Classification for Medical Telesurveillance*, 2nd International Conference on Biomedical Engineering, BIOMED'2004 Proceedings, ACTA PRESS, pp 395-398, Feb. 2004.
3. DAUDET L., TORRESANI B., *Hybrid representations for audiophonic signal encoding*, Journal of Signal Processing, Special issue on Image and Video Coding Beyond Standards, **Vol. 82(11)**, pp. 1595-1617, Nov. 2002.
4. PINQUIER J., SENAC C., ANDRE-OBRECHT R., *Speech and Music Classification in Audio Documents*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'02 Proceedings, Vol. 4, p. 4164, May 2002.
5. COWLING M., SITTE R., *Analysis of speech recognition techniques for use in a non-speech sound recognition system*, IEEE Transactions on Speech and Audio Processing, **Vol. 10**, pp. 504-516, Oct. 2002.
6. LU L., ZHANG H. J., JIANG H., *Content analysis for audio classification and segmentation*, IEEE Transactions on Acoustics, Speech and Signal Processing, **Vol. 10(7)**, pp. 504-516, Jan. 2002.
7. MALLAT S., *Une exploration des signaux en ondelette*, Les Editions de l'Ecole Polytechnique, Palaiseau, France, ISBN 2-7302-0733-3, 2000.
8. EZZAIDI H., ROUAT J., *Speech, musics and songs discrimination in the context of handsets variability*, International Conference on Spoken Language Processing, ICLSP'2002 Proceedings, Sept. 2002.
9. SECK M., MAGRIN-CHAGNOLLEAU I., BIMBOT F., *Experiments on speech tracking in audio documents using Gaussian Mixture Modeling*, IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'2001 Proceedings, **Vol.1**, May 2001.
10. SCHWARTZ G., *Estimating the dimension of a model*, Annals of Statistics, pp. 461-464, 1978.
11. CAREY M. J., PARRIS E. S., LLOYD-THOMAS H., *A comparison features for speech, music discrimination*, IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'99 Proceedings, March 1999.
12. KARNEBACK S., *Expanded examination of a low frequency modulation feature for speech/music discrimination*, International Conference on Spoken Language Processing, ICSLP'02 Proceedings,. Sept. 2002.
13. MCKINNEY M. F., BREEBAART J., *Features for Audio and Music Classification*, International Symposium on Music Information Retrieval, ISMIR'2002 Proceedings, Sept. 2003.
14. TZANETAKIS G., COOK P., *Musical Genre Classification of Audio Signals*, IEEE Transactions on Speech and Audio Processing, **Vol. 10**, No. 5, July 2002.
15. CHOU W., GU L., *Robust Singing Detection in Speech/Music Discriminator Design*, IEEE International Conference on Acoustics Speech and Signal Processing, ICASP'2001 Proceedings, pp. 865-868, May 2001.
16. SAUNDERS J., *Real-Time Discrimination of Broadcast Speech/Music*, IEEE International Conference on Acoustics Speech and Signal Processing, ICASP'96 Proceedings, **Vol.2**, p. 99, May 1996.
17. AJMERA J., MCCOWAN L., BOURLARD H., *Speech/music segmentation using entropy and dynamism features in a HMM classification framework*, Speech Communication 2003, **Vol. 40**, pp.351-363, 2003.
18. MORARU D., BESACIER L., *Toward Conversational Model for Speaker Segmentation*, Speech Technology and Human Dialog, SPED2003 Proceedings, pp. 69-78, April 2003.
19. CETTOLO M., VESCOVI M., *Efficient audio segmentation algorithms based on the BIC*, IEEE International Conference on Acoustics Speech and Signal Processing, ICASP'2003, **Vol. 6**, pp. 537-540, April 2003.
20. LEFEVRE S., MAILLARD B., VINCENT N., *A two level classifier for audio segmentation*, IEEE International Conference on Pattern Recognition, ICPR'02 Proceedings, **Vol. 3**, Aug. 2002.
21. REYES-GOMEZ M. J., ELLIS D. P., *Selection Parameter Estimation and Discriminative Training of Hidden Markov Models for General Audio Modeling*, IEEE International Conference on Multimedia and Expo, ICME'03 Proceedings, **Vol. 1**, pp. 73-76, July 2003.
22. AUGOYARD J.-F., *Sonorité, sociabilité, urbanité : méthode pour l'établissement d'un répertoire des effets sonores en milieu urbain*, Centre de Recherche sur l'Espace Sonore et l'Environnement Urbain (CRESSON), Ecole d'Architecture de Grenoble, Recherche N°80.471 du Ministère de l'Urbanisme et du Logement, 1982.