# Sound Detection and Classification for Medical Telesurvey

M Vacher, D Istrate, Laurent Besacier, Jean-François Serignat, E Castelli

# Sound Detection and Classification for Medical Telesurvey

M. Vacher, D. Istrate, L. Besacier, J.F.Serignat and E. Castelli
CLIPS - IMAG (UMR CNRS-INPG-UJF 5524) Team GEOD
Grenoble , France
Michel.Vacher@imag.fr, Dan.Istrate@imag.fr

**ABSTRACT**
Medical Telesurvey needs human operator assistance by smart information systems. This paper deals with the sound event detection in a noisy environment and presents a first classification approach. Detection is the first step of our sound analysis system and is necessary to extract the significant sounds before initiating the classification step. An algorithm based on the Wavelet Transform is evaluated in noisy environment. Then Wavelet based cepstral coefficients are proposed and their results are compared with more classical parameters. Detection algorithm and sound classification methods are applied to medical telemonitoring. In our opinion, microphones surveying life sounds are better preserving patient privacy than video cameras.

**KEY WORDS**
Acoustical Signal Processing, Noise, Sound Detection, Sound Classification, Wavelet Transform

## 1 Introduction

In this paper, we present an application of sound information extraction for medical telesurvey which is more and more frequently used in order to reduce hospitalization costs. Few of the studies related to this subject are involved in sound analysis. Detected sounds may be classified in normal or abnormal type, and according to this type an information or an alarm may be transmitted.

In order to reduce calculation time needed by a multichannel real time system, our sound extraction process is divided in two steps: detection and classification. The classification stage is only initiated if a sound event is detected. Sound event detection and classification are complex tasks because audio signals occur in noisy environment. In recognition step using a statistical study applied to acoustical parameters, we can choose the appropriate parameters that give the best classification results with a GMM system.

Detection and classification proposed methods are evaluated in noisy environment. The aim of our study is to obtain useful sound informations and transmit them through network to a medical supervising application running in a medical center.

## 2 Application

The habitat we used for experiments is a 30 $m^2$ apartment situated in the TIMC laboratory inside the Faculty of Medicine of Grenoble, equipped with various sensors, especially microphones [1]. The entire telemonitoring system is composed of three computers which exchange information through a Control Area Network bus (see *Figure 1*). The master computer is in charge of data fusion and analyzes both data coming from fixed and moving sensors and information coming from the sound computer, which is continuously surveying the microphones.

The sound analysis system is working as follow: each time a sound event is analyzed, a message is sent to the master computer, notifying occurrence time of detection, type of event (speech or other sound), localization of the emitting source ; it also should indicate either the most probable sound classes. The sound or speech source can be localized by comparing the sound levels of the microphones.

From this the master computer could send an alarm if necessary. At the moment, the recognition system is only in test and the detected events are classified by a human operator.

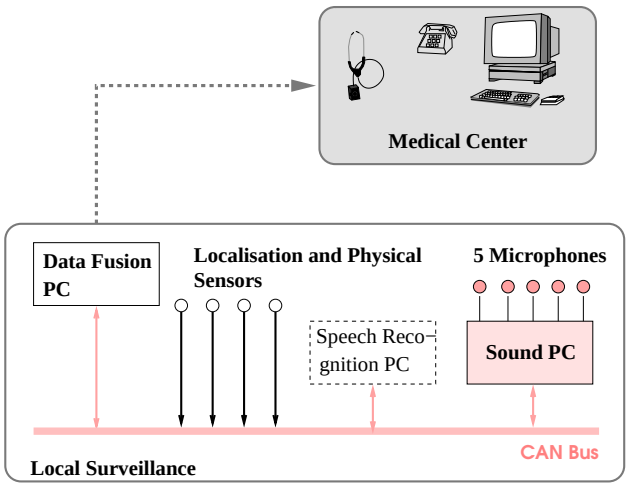


Figure 1. Acquisition and analysis system

## 3 Sound classes for medical telesurvey

The system we work on is designed for the surveillance of the elderly, convalescent persons or pregnant women. Its main goal is to detect serious accidents as falls or faintness (which can be characterized by a long idle period of the signals) at any place in the apartment. It was noted that the elderly had difficulties in accepting the video camera monitoring, considering it a violation of their privacy.

Thus, the originality of our approach consists in replacing the video camera by a system of multichannel sound acquisition. A microphone is placed in every room (hall, toilet, shower-room, living-room). Each of the 5 microphones is connected to the system which is analyzing in real time sound environment of the apartment. Detection of abnormal sounds (objects or patient's falls) could indicate a distress situation in the habitat.

To respect again privacy, no continuous recording or storage of the sound is made, since only the last 10s of the audio signal are kept in a buffer and sent to the alarm monitor if necessary. It can be used by the human operator to take the decision of a medical intervention.

The everyday sounds are divided into 7 classes. The criteria used for this repartition were : statistical probability of occurrence in everyday life, possible alarm sounds (scream, person fall) are priority, the 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 (cough, sneeze,...), dishes sound, door lock

- abnormal sounds that generate an alarm (breaking glasses, screams, fall sounds).

If an abnormal sound is recognized, the sound analysis system transmits an alarm to the medical supervising application. The decision to call the emergency is taken by this data fusion system [2].

As no everyday life sound database was available in the scientific area, we have recorded a sound corpus. This corpus contains recording made in 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, screams, water, ringing, etc. The corpus contains 20 types of sounds with 10 to 300 repetitions per type. The test signal database has a duration of 3 hours and consists of 2376 files.

The sound classes of our corpus are described in the following table; the frame number for each class is given too. Each frame has a duration of 16ms.

This corpus is not yet complete, 2 classes very usefull for this application are remaining to record: Human Sounds and Fall Sounds.

| Sound Class | Frame Number | Alarm |
|---|---|---|
| Door Slap | 47 398 | No |
| Breaking Glasses | 9 338 | YES |
| Ringing Phone | 59 188 | No |
| Step Sound | 3 648 | No |
| Scream | 17 509 | YES |
| Dishes Sounds | 7943 | YES |
| Door Lock | 605 | No |

## 4 Sound classification

### 4.1 Sound extraction

The detection of a signal (useful sound) is very important because if an event is lost, it is lost forever. On the other hand, start and stop time of sound must be accurately established to use classification-step with the best conditions.

Unlike Fast Fourier Transform, Wavelet Transform is well adapted to signals that have more localized features than time independent wave-like signals: door slap, breaking glasses, step sound, etc... They are more and more used for signal detection [3] and audio processing. We have chosen Daubechies wavelets with 6 vanishing moments to compute DWT [4]. A complete, orthonormal wavelet basis consists of scalings ($s$ factor) and translations ($u$ delay) of the mother wavelet function $\psi(t)$, a function with finite energy and fast decay:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\,\psi\left(\frac{t-u}{s}\right). \qquad (1)$$

Two Daubechies-wavelet are shown in *Figure 2*, they are in different hierarchical levels of scale, and also at different spatial positions. As illustrated in this figure, the higher the coefficient level is, the more the Wavelet function support is compact; in the case of Fourier Transform, Sine and Cosine keep the same support, only their zero crossing rate will be higher. Moreover, the frequency spectrum of the Wavelet function will widen, whereas Sine and Cosine spectrum are always Dirac pulses. Therefore Wavelet Transform on a 512 sample frame corresponding to a 32ms window allows good signal enhancement in HIS

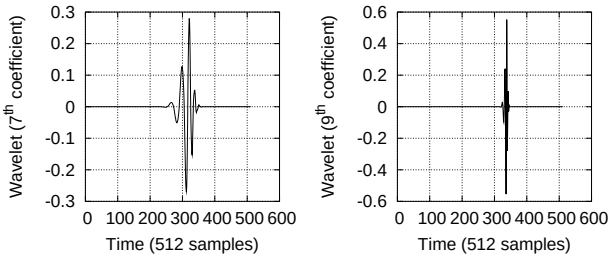


Figure 2. Daubechies-wavelet time variation for 2 different scaling factors

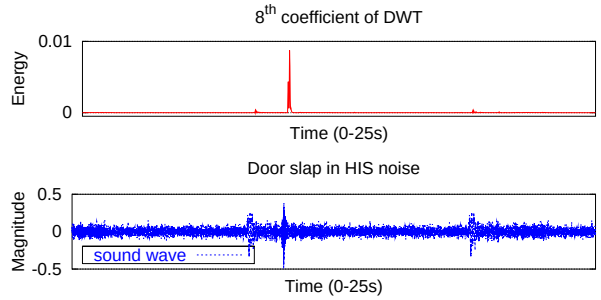


Figure 3. Time evolution of $8^{th}$ DWT coefficient's energy

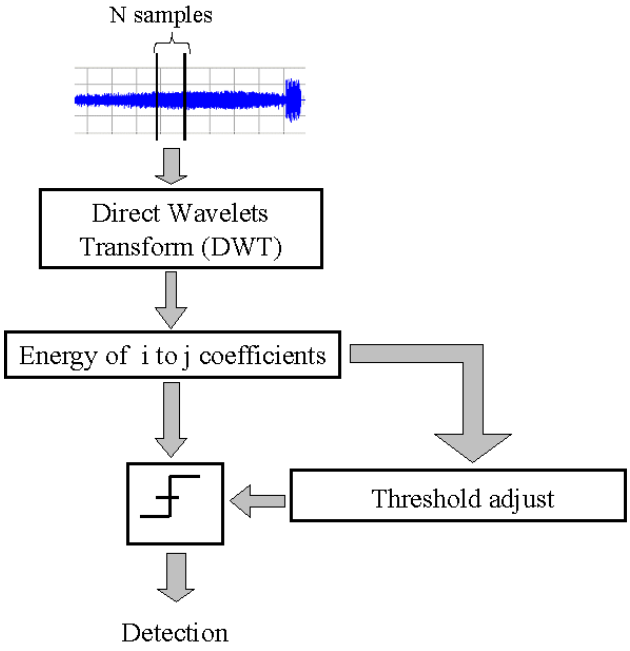


Figure 4. Flowchart of Wavelet based detection

and white noise. Discrete Wavelet Transform result is a matrix of same size than the signal (512); this matrix contains 10 vectors or *wavelet coefficients* but the two first coefficients should strictly be called *mother-function coefficients*. Each coefficient component number is double of preceding hierarchical level coefficient.

The algorithm (see flowchart in *Figure 4*) is computing the energy of the 8, 9 and 10 wavelet coefficients, because most significant wavelet coefficients for sounds to be detected are rather high order, as shown in *Figure 3*: two parasitic noises which are flanking the sound are nearly cleared and useful sound is appearing at time 10s.

The detection is achieved by applying a threshold on the sum of energies. The threshold is self-adjustable and depends on the average of the 10 last energy values: $Th = \kappa + 1.2 \cdot E_{Average}$ . Overlap between two consecutive analysis windows is 50%.

## 4.2 Gaussian Mixture Model: GMM

We have used a **G**aussian **M**ixture **M**odel (GMM) method in order to classify the sounds [5]. There are other possibilities for the classification: HMM, Bayesian method and other [6]. This method evolves in two steps: a training step and a recognition step. We have chosen to use a model with only 4 Gaussian components, since preliminary experiments have shown no improvement with more components.

**The Training Step.** The GMM training has been done on the ELISA [7] platform. The training is initiated for each class $\omega_k$ of signals of our corpus and gives a model containing the characteristics of each Gaussian distribution ($1 \leq m \leq 4$) of the class: the likelihood $\pi_{k,m}$, the mean vector $\mu_{k,m}$, the covariance matrix and the inverse matrix $\Sigma_{k,m}^{-1}$. These values are achieved after 20 iterations of an "EM" algorithm (Expectation Maximization). The matrices are diagonal.

**The Recognition Step.** Each extracted signal, $X$, is a series of $n$ acoustical vectors, $x_i$, of $p$ components. The parameters $\pi$, $\mu$ and $\Sigma$ have been estimated during the training step for each of the 4 Gaussians. The likelihood of membership of a class $\omega_k$ for each acoustical vector $x_i$ is calculated for each class according to estimated Gaussian parameters. The signal $X$ belongs to the class $\omega_l$ for which $p(X \mid \omega_l)$ is maximum: $p(X \mid \omega_k) = \prod_{i=1}^{n} p(x_i \mid \omega_k)$.

## 4.3 Sound parameters

**MFCC.** The calculus steps for the MFCC coefficients are: Fast Fourier Transform of the analysis signal window; Mel triangular filtering; logarithm calculus of the filtered coefficients and inverse cosines transform. The Mel frequency scale $f_{Mel}$ is logarithmic (see *Formula 2*). The response of Mel triangular filters are shown in the figure 5.

$$f_{Mel} = 2595.log\left(1 + \frac{f}{700}\right) \quad (2)$$

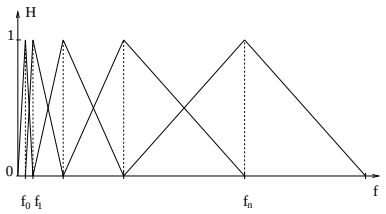


Figure 5. Triangular Mel filters

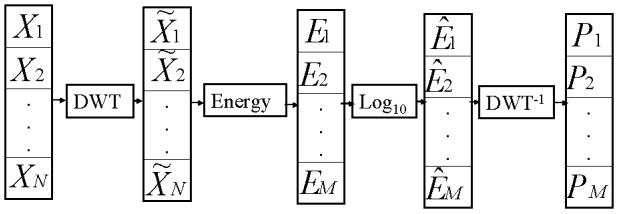


Figure 6. Wavelet based coefficient determination ($N = 256$ samples, $M = 6$ high order coefficients)

The inverse cosinus transform is obtained according to:

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n(m-\frac{1}{2})}{M}\right) \quad 0 \leq n < M \quad (3)$$

In order to model the signal time variation, we have tested also the $\Delta$ (first derivative) and $\Delta\Delta$ (second derivative) of MFCC Coefficients.

**Wavelet based coefficient (DWTC).** This type of acoustical parameters is based on Wavelet Transform as shown in *Figure 6*. Firstly Discrete Wavelet Transform is applied on a 256 sample window. Secondly, the energy of the last 6 Wavelet Transform coefficients is calculated followed by an amplitude logarithmic transformation. The acoustical vector contains DWT$^{-1}$ of logarithmic energy coefficients. The total number of parameters is 6.

**ZCR, RF and centroïd.** The value of the zero-crossing rate is given by the number of crossings on time-domain through zero-voltage within an analysis frame. In order to eliminate the noise influence, we have introduced a symmetric clipping threshold. The value of clipping threshold represents 0.03% of signal amplitude. In fact, the zero-crossing rate indicates the dominant frequency during the time period of the frame.

Roll-off Point (RF) and Centroïd are used to measure the frequency which delimits 95% and 50% of the power spectrum. The roll-off point can be viewed as a measure of the "skewness" of the spectral shape. Their values are solutions $\Phi_i$ of *Equation 4* (RF: $\alpha_1 = 0.95$, Centroïd: $\alpha_2 = 0.5$).

$$\sum_{k<\Phi_i} X[k] = \alpha_i \sum_{k} X[k] \quad (4)$$

# 5 Results

## 5.1 Extraction in noisy environment

*Wavelet filtering* algorithm results are given in *Figure 7*. Best results are for HIS noise: EER=0% for SNR$\geq$+10dB and EER=7.6% for SNR=0dB. The results are roughly less for white noise (EER=4% for SNR=10dB), but they are enough to allow good performances for similar noises like water flow.

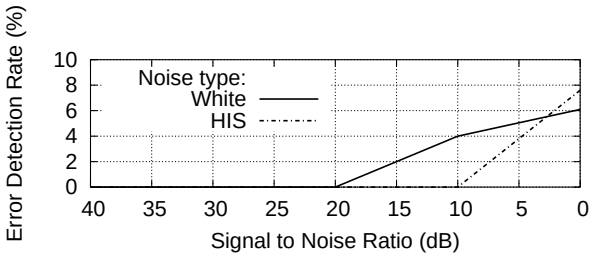


Figure 7. Detection results in noisy environment

In order to validate the results obtained on the simulation test set, we have recorded 60 files inside test apartement (real conditions) at various SNR (minimum 2 dB, maximum 30 dB, average 15 dB). We have used the same sounds as in the simulation test set, played with a loudspeaker.

There was no false alarme and no missed detection. This confirmed simulated results.

## 5.2 Classification in noiseless environment

The analysis window was set to 16 ms with an overlap of 8 ms. The GMM model is made of 4 Gaussian distributions. The training/test protocol is a "leave one out" protocol: the model of each class is trained on all the signals of the class, excepting one. Next, each model is tested on the remaining sounds of all classes. The whole process is iterated for all files (1577 tests).

The experimental results are in *Table 1*. The average of **E**rror **C**lassification **R**ate (ECR: number of recognition error divided by the number of tests) and the correspondent number of parameters (PN) are given. For each parameter, we calculate the average of the error value of all the classes. This first sound classification results are encouraging.

We can observe that the best results are obtained with the MFCC parameters (speech specific parameters) coupled with new parameters like zero crossing rate, roll-off point, centroïd.

Proposed DWTC parameters have not as good performances as classical MFCC parameters; however only 6 parameters are needed for classification against a minimum of 17 in the other case. However, this could reduce time calculation if needed.

| Parameters | PN | ECR [%] |
|---|---|---|
| $\Delta$,$\Delta\Delta$(16MFCC+Energy +ZCR+RF+Centroïd) | 60 | 8.71 |
| **16 MFCC + Energy+ ZCR+RF+Centroïd** | 20 | **11.47** |
| 16MFCC+Energy | 17 | 15.21 |
| DWTC | 6 | 18.78 |

Table 1. Results of sound classification methods

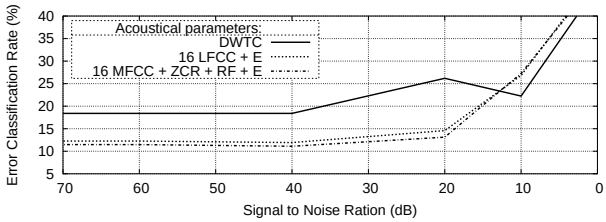


Figure 8. Classification error in HIS noise

### 5.3 Classification in noisy environment

We have tested our classification system in HIS noise. For 16MFCC + ZCR + RF + Centroïd, results are roughly constant for SNR$\geq$+20dB, but they decay beyond: error classification is 27% for SNR=+10dB (see *Figure 8*). For DWTC, better results are obtained for SNR=+10dB but they are remaining worse for SNR$\geq$+20dB (ECR$\leq$26%). Real conditions are between 10 and 20dB of SNR and these first results are not sufficient. We are actually working to improve performances by signal enhancement.

## 6 Conclusions

This system is developed for medical supervision application in the framework of DESDHIS project, but possible applications of our sound extraction process are numerous: multimedia documents classification, security sound surveillance, medical telemonitoring etc.

We have presented detection and classification methods which allow us to detect and classify a sound event in the habitat. Firstly, we have used classical parameters of speech recognition; secondly we have tested new parameters like Wavelet based coefficients. First results are encouraging but sound classification in noisy environment must be improved.

We are working on speech recognition techniques to allow the call for help by the patient in case of distress situation: recognition of specific alarm keyword may be very useful for the data fusion system.

## Acknowledgment

This system is a part of the DESDHIS-ACI "Technologies for Health" 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, charged with the medical sensors analysis and data fusion.

## References

[1] G.Virone, N.Noury, and J.Demongeot, "A system for automatic measurement of circadian activity in telemedicine," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 12, pp. 1463–1469, December 2002.

[2] G. Virone, D. Istrate, M. Vacher, et al., "First steps in data fusion betwenn a multichannel audio acquisition and information system for home healthcare," in *IEEE Engineering in Medecine and Biology Society, Cancun*, September 2003.

[3] F.K. Lam and C.K. Leung, "Ultrasonic detection using wideband discret wavelet transform," in *IEEE TENCON*, August 2001, vol. 2, pp. 890–893.

[4] Stéphane Mallat, *Une exploration des signaux en ondelette*, ISBN 2-7302-0733-3. Les Editions de l'Ecole Polytechnique, 2000.

[5] L. Lu, H.J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transaction on Speech and Audio Processing*, vol. 10(7), pp. 504–516, October 2002.

[6] M.Cowling and R. Sitte, "Analysis of speech recognition techniques for use in a non-speech sound recognition system," in *Digital Signal Processing for Communication Systems, Sydney-Manly*, January 2002.

[7] I.Magrin-Chagnolleau, G.Gravier, and R.Blouet, "Overview of the ELISA consortium research activities," *2001 : a Speaker Odyssey*, pp. 67–72, June 2001.