Sound System Analysis for Health Smart Home

Eric CASTELLI*, Dan ISTRATE**, Cong-Phuong NGUYEN*
*International Research Center MICA - 1 Dai Co Viet - Hai Ba Trung - Hanoi - Vietnam
Tel: +84-4-868-3087 Fax: +84-4-868-3551 E-mail: Eric.Castelli@mica.edu.vn
**CLIPS-IMAG, BP 53, 38041 Grenoble Cedex 9, France
Tel: +33-4-76-51-4510 Fax: +33-4-76-63 5552 E-mail: Dan.Istrate@imag.fr

Abstract:
A multichannel smart sound sensor capable to detect and identify sound events in noisy conditions is presented in this paper. Sound information extraction is a complex task and the main difficulty consists is the extraction of high-level information from an one-dimensional signal. The input of smart sound sensor is composed of data collected by 5 microphones and its output data is sent through a network. For a real time working purpose, the sound analysis is divided in three steps: sound event detection for each sound channel, fusion between simultaneously events and sound identification. The event detection module find impulsive signals in the noise and extracts them from the signal flow. Our smart sensor must be capable to identify impulsive signals but also speech presence too, in a noisy environment. The classification module is launched in a parallel task on the channel chosen by data fusion process. It looks to identify the event sound between seven predefined sound classes and uses a **G**aussian **M**ixture **M**odel (GMM) method. Mel Frequency Cepstral Coefficients are used in combination with new ones like zero crossing rate, centroid and roll-off point. This smart sound sensor is a part of a medical telemonitoring project with the aim of detecting serious accidents.

Key words: Acoustical Signal Processing, Noise, Multichannel Processing, Real Time, Sound Detection & Classification, Wavelet Transform

1. INTRODUCTION

Our days, sound became a preferred interface and information source for smart home and perceptive spaces. Sound contains low level information like class (silence or music or diverse sounds like door clapping, ringing phone, fall sound), sound type (impulsive or harmonic), speaker identity, and high level information such the lexical parts (words or sentences). The required sensors capabilities are increasing: these sensors become more and more complex, the digital signal processing being a crucial component of them. Extraction of high level information is possible today in real time, thanks to many studies. The most difficult task in digital signal processing for sound sensor is the extraction of high-level information from an one-dimensional signal. The actual challenges are the multichannel processing, attenuation of environmental noise, multispeaker speech recognition.

In this paper we describe a multichannel smart sound sensor, which detect and identify a sound between several predefined sound classes. The input of smart sound sensor is composed of data collected by 5 microphones and its output data is sent through a network (CAN bus or Ethernet). For a real time working purpose, the sound analysis is divided in three steps: sound event detection for each sound channel, fusion between simultaneously events and sound identification. The extracted information is sent through the network and if it is needed, the recorded sound can be transferred for latter analysis by Ethernet network, adapted for large data flow.

This smart sound sensor is a part of a medical telemonitoring project with the aim of detecting serious accidents. In these conditions we consider a *sound event*, an impulsive sound (door clapping, step sounds, dishes sounds, etc.) and a *noise*, a stationary signal (environmental noise, white noise, water flow noise, etc.). The proposed smart sensor is implemented in real time with LabWindows/CVI software on a PC [1]. Evaluation of the sensor has been carried out with real environmental noise on a generated test set. An evaluation methodology is proposed and discussed.

2. SOUND DATABASE

In order to test and validate the smart sound sensor we have generated a sound corpus. It contains recordings made in the Clips laboratory (15 % of the CD), the files of "Sound Scene Database in Real Acoustical Environments" [2] (70 % of the CD) and files from a commercial CD (film effects, 15 % of the CD). Entire corpus is composed of 3354 files; every sound is sampled at 16 KHz and 44 KHz.

We have carried out 7 sound classes from the sound corpus, for training the classification algorithm. The 7 sound classes are presented in the table 1.

The test set used for smart sound sensor validation is a mix between the 7 sound classes and noise at different SNR. The noise is recorded in an experimental apartment named HIS. There are 7 files corresponding to the 7 sound classes. Each file is constituted by all sounds of the corresponding class inserted with silence periods of random time length (1577 events to detect and identify). For each sound, SNR can have a random value either between 10dB and 20dB or between 0dB and 40dB. For the first case, the SNR repartition is uniform and for the second one, the SNR repartition corresponds to real measures in the HIS apartment.

Silence between consecutive sounds varies randomly between 5 and 60 seconds. Total number of useful sounds to be detected is 1577.

Table 1 - Sound class description

Sound class	N° files	N° frames	Total length	Alarm
Door clapping	523	47398	379 s	NO
Glass breaking	88	9338	75 s	**YES**
Ringing phone	517	59188	474 s	NO
Step sounds	13	36480	292 s	NO
Screams	73	17509	140 s	**YES**
Dishes sounds	163	7943	64 s	**YES**
Door lock	200	6050	49	NO

3. SOUND SENSOR HARDWARE

As described in figures 1 & 2, our smart sound sensor is composed of 5 microphones (omni-directional condenser type), an acquisition card (National Instruments PCI-6034E) plugged in the PC which is in charge with the sound analysis software and a CAN Bus adapter card. Condenser microphones are used because of their small dimensions and of their omnidirectional characteristics. Each microphone is equipped with a conditioning card (instrumentation amplifier and anti-aliasing filter). The sampling frequency was fixed at 16 kHz because this value is usual in speech recognition. In our medical telemonitoring application the CAN bus is a dedicated one which provides a big security. It is used to collect information of other types of sensors useful in perceptive space: medical sensors, localization sensors. The CAN bus was chosen as the output interface because its low cost, its good resistance to harsh environments and its deterministic response in collision case [3]. But CAN bus speed is too low for sound wave transmission; in this case it must be replaced by standard Ethernet network.

For sound sample acquisition, low-level functions are used in order to drive the acquisition card in real time. The detected events are saved on PC (on hard-disk) and sent through the Ethernet network. Simultaneously, an history of detected events (detection time, detection type) is recorded in a text file. The abnormal detected signal is recorded in a standard Wave format (without compression) and could be sent in the same format through Ethernet network, if requested.

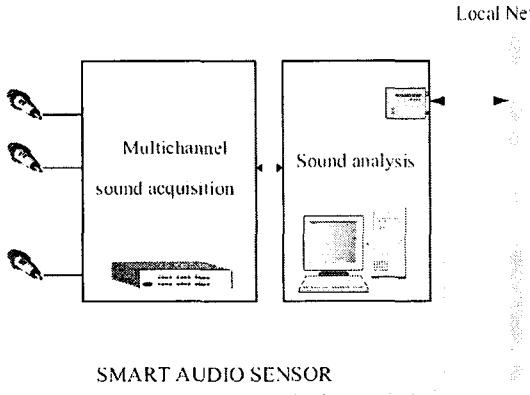

Figure 1 - Smart Sound Sensor diagram

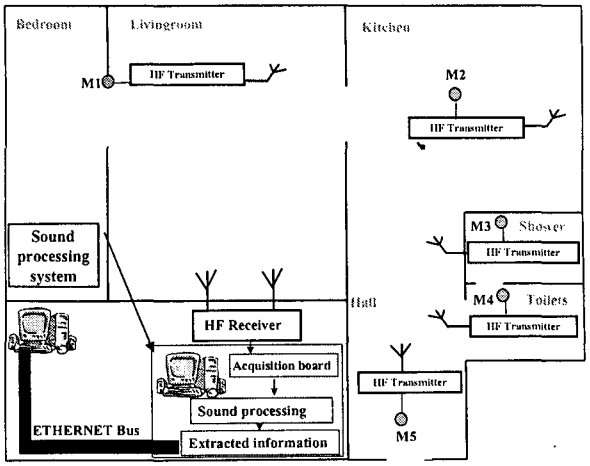

Figure 2 - Smart Sound Sensor hardware

4. SOUND ANALYSIS SYSTEM

The sound analysis system has been divided in three modules as shown in figure 3. The first and second modules are making up the "First parallel task". The third module is the "Second parallel task".

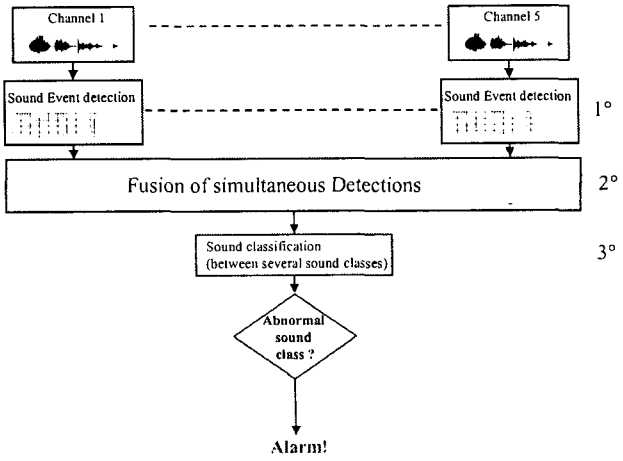

Figure 3 - Sound processing schema

The first module is processed on each channel in order to detect a sound event and to extract it from signal flow. The fusion module selects the best channel in the case of a simultaneously event detection on several channels. The channel with the highest SNR is chosen from estimation of SNR made for each channel. These two modules make up the "First Parallel Task".

The last step is the sound classification or "Second Parallel Task". This module is receiving the sound event extracted (output of "First Task") and it estimates the most probable sound class. The proposed sensor classifies the sound in one of the seven sound classes.

4.1 Sound detection and capture

The event detection aim is to find impulsive signals in the noise and to extract them from the signal flow. Our smart sensor must be capable to identify impulsive signals like door clapping, dishes sounds, fall sounds but also speech presence too, in a noisy environment.

The performances of the first module are very important for the entire system because if an event is lost, it is lost forever.

There are many techniques for sound detection: energy threshold, statistical model [4], energy processing [5] or wavelet processing [6]. We have validated three detection algorithms proposed by Dufaux [7] on our test set but the obtained performances for environmental noise are not suitable. We have proposed two other algorithms [8] with better performances either for environmental noise or for water flow noise but not for the two cases.

A wavelet based event detection algorithm is proposed in the following. Unlike Fast Fourier Transform, Wavelet Transform is well adapted to signals that have very localized features in the time-frequency space. This transform is frequently used for signal detection [8] and audio processing. We have chosen Daubechies wavelets with 6 vanishing moments to compute Disnete Wavelet Transform (DWT). A complete orthonormal wavelet basis consists of scaling (s factor) and translations (u delay) of the mother wavelet function $\psi(t)$, a function with finite energy and fast decay. Continuous wavelet transform is defined by :

$$Wf(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\Psi^*\left(\frac{t-u}{s}\right)dt \qquad (1)$$

Wavelet Transform on a 512 sample frame corresponding to a 32ms window allows good signal enhancement in noisy conditions. Analyzed sounds are impulsive and so, better enhanced by Wavelet Transform. Discrete Wavelet Transform is applied on the sampled data and its output forms a vector of the same length with that of the signal (512). This vector has a pyramidal structure and is composed of 10 wavelet coefficients (figure 3).

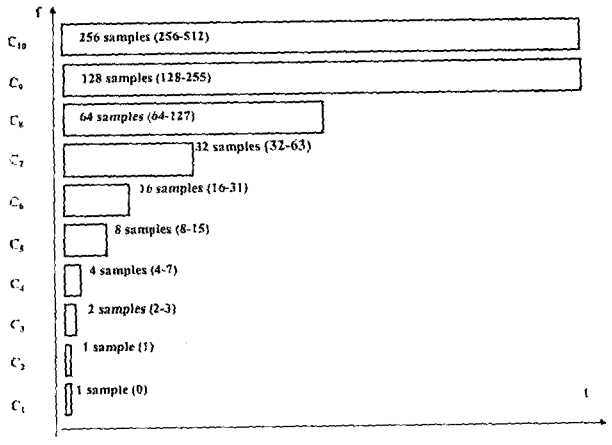

Figure 3 - Wavelet coefficients repartition

The algorithm (flowchart in figure 4) calculates the energy of the 8, 9 and 10 wavelet coefficients (the three higher order coefficients), as the significant wavelet coefficients of the sounds to be detected are rather high order.

The detection is achieved by applying a threshold on the sum of energies of the three highest order wavelet coefficients. The threshold is self-adjustable and depends on the average of the 10 last energy values: $Th = k+\alpha.E_{Average}$ .The used value of α coefficient is 1.2 in order to compensate the small variation of signal around average. Overlap between two consecutive analysis windows is 50%.

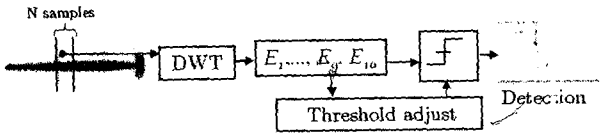

Figure 4 - Flowchart of the wavelet-based algorithm

An example of sound detection achieved by adopting the presented algorithm is shown in the figure 5. The amplitude of the sound signal that contains a mixture between a ringing phone at 3.2 second and a water flow noise at 0 dB of SNR can be seen in the first window, in the figure 5, while the second window shows the wavelet energy outlined in black and the self-adjustable threshold in grey. The detection signal presented in the third window shows clearly that the algorithm detects the signal from noise despite their close amplitudes.

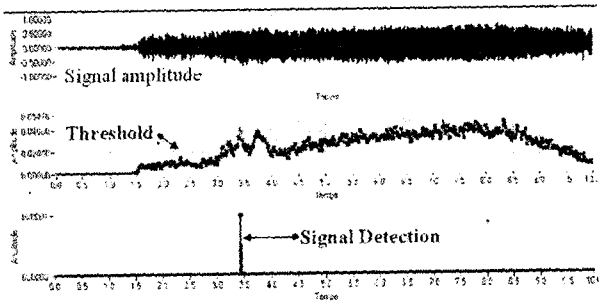

Figure 5 - Detection of a ringing mixed with water flow noise at 0 dB of SNR with proposed algorithm

The presented algorithm detects only the signal beginning and not the end. A first approach is to consider a fix duration sound as detection output. However, the sound classification system is very sensitive to long silence parts. Therefore, we detect the end of the signal by applying the same algorithm on the time-inverted signal.

The procedure used to realize everyday life sound event detection and capture involves the following steps: the output signal starts simultaneously with the detection and lasts 7 seconds. Then, the signal is time inverted and the detection algorithm is applied once again. In the next step the detection of the end of the signal is used to extract the sound (the output signal has a variable length). This procedure is allowing classification algorithm to analyze only the typical part of the detected signal.

4.2 Channels fusion

Event detection is continuously operating on each of five audio inputs. According to figure 3, a fusion is necessary between these 5 inputs in the aim of choosing the best channel and implicitly, to localize

the sound source, when multiple detections occur simultaneously. We consider simultaneous detections two or more detections occurring at less than 0.5 seconds after the first detection (the propagation time and the algorithm structure determine this time).

In case of simultaneous detections, the best channel is considered the one with the highest SNR. The SNR is estimated for each channel like the ratio between average sound power (1 second of signal after event detection) and average noise power (1 second of signal before event detection). The average noise power is buffered continuously in memory. The output of fusion module gives the event signal.

4.3 Sound classification

The classification module looks to identify the event sound between predefined sound classes. This module uses a Gaussian Mixture Model (GMM) method [9]. There are other possibilities for the classification: HMM, Bayesian method and others but GMM classification is easy to implement, procures comparable performances and require low processing time.

This method evolves in two steps: a training step and an identification one. Identification module is the only module involved in real time constraints. Classification does not use directly signal samples, but a vector of acoustical parameters calculated on analysis windows. The acoustical parameters are determined for each analysis window of 16ms with an overlap of 8ms.

The training is initiated for each class ω_k of signals from sound corpus and gives a model containing the characteristics of each Gaussian ($1 \le m \le 4$) of the class: the likelihood $\pi_{k,m}$, the mean vector $\mu_{k,m}$, the covariance matrix and the inverse matrix $\Sigma^{-1}_{k,m}$. These values are achieved after 20 iterations of an "EM" algorithm (Expectation Maximization) following a K-means algorithm. The used matrices are diagonal. Each extracted signal X is a series of n acoustical vectors x_i of p components. The parameters π, μ and Σ have been estimated during the training step. The membership likelihood of a class ω_k for each acoustical vector is calculated for all classes according to:

$$\begin{cases} p(x_i|\omega_k) = \sum_{m=1}^{4} \pi_{k,m} \cdot \dfrac{1}{\sqrt{(2\pi)^d \left|\sum_{k,m}\right|}} \cdot e^{A_{i,k,m}} \\ A_{i,k,m} = \left(-\dfrac{1}{2}(x_i - \mu_{k,m})^t \cdot \dfrac{1}{\sum_{k,m}} \cdot (x_i - \mu_{k,m}) \right) \end{cases} \quad (2)$$

The likelihood of the entire signal is obtained by geometrical average:

$$p(X|\omega_k) = \prod_{i=1}^{n} p(x_i|\omega_k) \quad (3)$$

The signal X belongs to the class ω_l for which $p(X|\omega_l)$ is maximum.

4.4 Selection of Gaussian model number

The Bayesian Information Criterion (BIC) was used in order to determine the optimal number of Gaussian [10]. BIC criterion selects the model through the maximization of integrated likelihood: $BIC_k = -2L_k + \nu_k \ln(n)$, where L_k is the maximum of the logarithmic likelihood, equal to $\log f\left(x \middle| K, \hat{\theta}\right)$ (f is integrated likelihood), K the component number of model, ν_K the number of free parameters, n is the number of frames and θ is parameter space. The minimum value of BIC indicates the best model order.

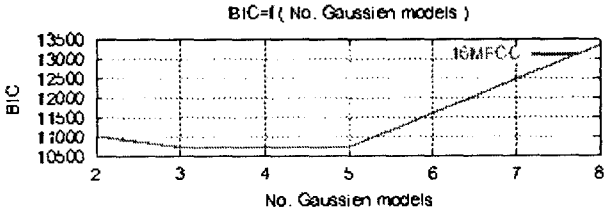

Figure 6 - BIC for 2, 3, 4, 5 and 8 Gaussian

The BIC criterion has been calculated for the sound class with the smallest number of files, for 2, 3, 4, 5 and 8 Gaussian. The results showed in figure 6 are obtained for 16 parameters MFCC. Analyzing these results, a number of Gaussian between 3 and 5 seems to correspond to the best sound modelling. A model with 4 Gaussian distributions has been chosen for following tests.

4.5 Acoustical parameters

There are many types of acoustical parameters like MFCC (Mel Frequencies Cepstral Coefficients), LFCC (Linear Frequencies Cepstral Coefficients), LPC (Linear Predictive Coefficients), LPCC (Linear Predictive Cepstral Coefficients), ZCR (zero crossing rate), RF (Roll-off point), Centroid, etc, but only few of them are appropriate to the sound classification.

After a statistical study based on Fisher Discriminant Ratio (FDR) and a validation on a test set, a combination of 16 MFCC with ZCR, RF and Centroid has been chosen (error classification rate was 10 % on a 1577 test set). MFCC are cepstral coefficients based on triangular filtered energy in different frequency bands (Fourier Transform with a Mel frequency scale followed by logarithm and Inverse Fourier Transform). The ZCR is the number of crossings on time-domain through zero-voltage. The RF measures the frequency which delimits 95 % of the power spectrum, while the Centroid is the frequency which divides the power spectrum in two equal parts.

5. SMART SOUND SENSOR VALIDATION

5.1 Implementation

The whole sound analysis flow-chart (figure 3) is implemented by a software written with

LabWindows/CVI on our smart audio PC. This software drives simultaneously the real time sound acquisition on 5 channels at 16 kHz sample rate. The acquisition is made by a double buffering of 2048 samples by channel. Between 2 acquisitions of 2048 samples, the event detection is made on each of 5 channels. In the same time the signal power is calculated and stored into memory (this value is necessary for SNR estimation).

In case of event detection the fusion module is launched in order to select the best channel. The signal of the selected channel is recorded on the hard disk (7 seconds after event detection). The number of selected channel is shown on the software front panel, like sound localization information.

Before launching a second event detection procedure, the detected signal is time inverted in order to estimate the signal end. After signal extraction at founded signal end, the classification algorithm is started up. The most probable sound class is shown on the front panel. All these modules are launched in a parallel task in relation to real time sound acquisition task.

When a sound event is detected, the smart sound sensor is emitting an information frame through the CAN bus. The frame contains: date and time detection (day, month, year, hour, minute, second, milliseconds) and a character field. This character field is consist of: the three most probable sound classes with their corresponding likelihoods and the localization of the sound event (the channel).

On the software front panel (figure 7) are shown the signal of last detected event, the localization of ast event (the room in our experimental apartment), a chronological account of detected events, and, on apartment plan, and the event localization once again.

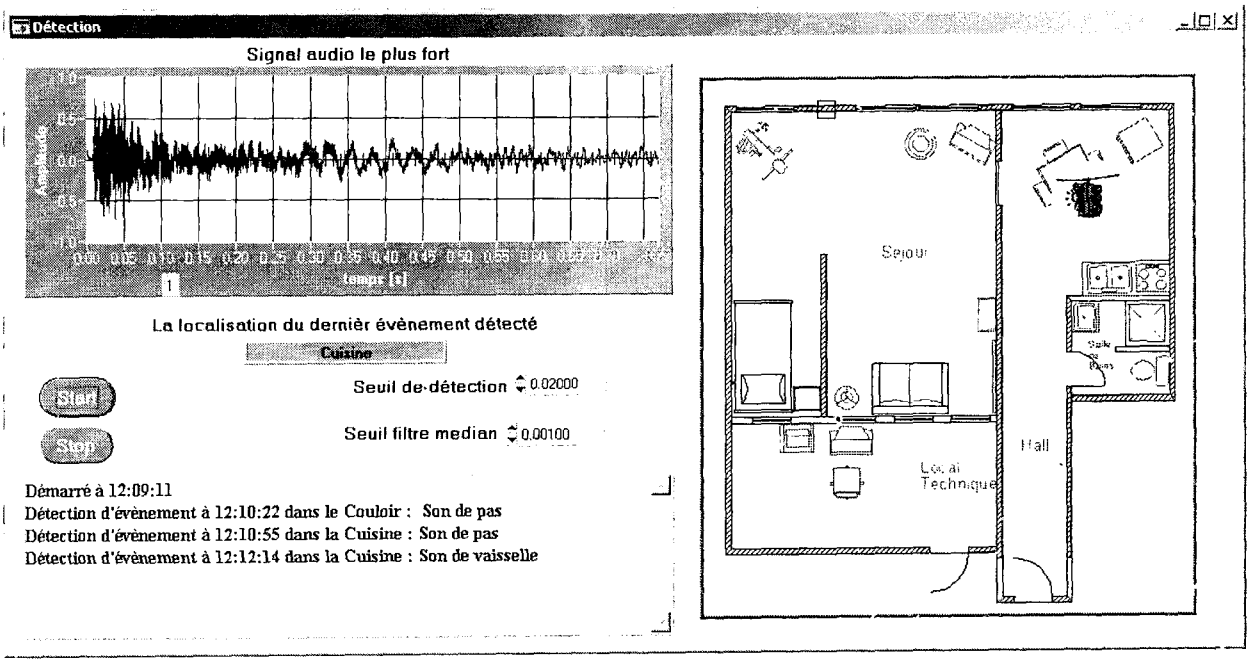

Figure 7 - Front panel of smart sensor. Detection of step and dish sounds

5.2 Coupling between detection and classification

The detection of the end of the signal can have a great influence on classification performances. In order to evaluate the influence of the coupling between the two parallel tasks (detection and classification) a study has been conducted on the same test set. 16 MFCC coupled with ZCR, RF and Centroid are the acoustical parameters. The Error classification rate (ECR) is calculated for the sound obtained with an ideal detection (according to corpus data) and for the sounds coming from automatic detection algorithm. The GMM training is made on the pure sounds with a *leave-one out protocol*. The results are presented in table 2.

The obtained ECR for ideal detection confirm our results of classification in noise conditions. The results obtained with a fixed length of extracted signals are not acceptable. The error introduced by non-adapted coupling is approximately 46%. The signal end detection improves significantly the classification performances. False detections in only noise signal parts may explain that ECR is greater for real detection with length estimation than ideal detection.

Table 2 - Coupling results

	ECR for SNR∈[10 – 20] dB	**ECR** for SNR∈[0 – 40] dB
Ideal Detection & Real length of signals	21.5 %	22.7 %
Automatic Detection & length estimation	**27.7 %**	**25.5 %**
Automatic Detection & fix length of signals	67.8 %	69 %

5.3 Proposed methodology for sensor evaluation

The evaluation of the smart sound sensor for a telemedicine application must take into account different characteristics of both detection and classification modules. For detection module we define a Good detection (G) as an event detection occurring between 0.5 seconds before signal start and signal end. A Missed Detection (MD) is a lack of detection in the previously defined time interval and a False Alarm (FA) is a detection occurring outside of this time interval.

For classification module, the 7 sound classes are divided in two categories: class with alarm (A) and class without alarm ($\overline{A}$). Possible errors are:

- **Error without consequence (W)** = a sound of a class with alarm is classified in another class with alarm *or* a sound of a class without alarm is classified in another class without alarm
- **Missed detection (MD)** = a sound of a class with alarm is classified in a class without alarm
- **False alarm (FA)** = a sound of a class without alarm is classified in a class with alarm

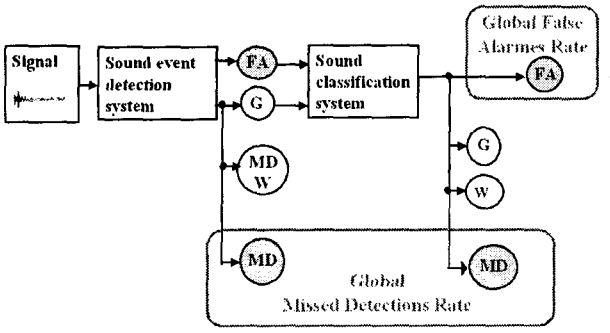

Figure 8 - Global False Alarm Rate and Global Missed Detection Rate representation

All errors for the two tasks are illustrated in the figure 8. The proposed Global Missed Detection Rate (GMDR) and the Global False Alarm Rate (GFAR) are defined in equations (GMDR) and (GFAR) in accordance with figure 8 as follows.

$$GMDR = \frac{MD_{Detect} + MD_{Class}}{Total\ Number\ of\ event\ detections} \quad (4)$$

$$GFAR = \frac{FA_{Class}}{Total\ number\ of\ event\ detections + FA_{Class}} \quad (5)$$

6. RESULTS

The results of smart sound sensor evaluation on the test set are presented in the table 3. Test set contains a mixture between real noise (recorded in the test apartment) and everyday life sounds. These results were obtained by using wavelet-based algorithm for detection of the start and of the end of signal. The classification module was achieved with 4 Gaussian models and 16 MFCC coupled with energy, zero crossing rate, roll-off point and centroid. For classification module a *leave one out* protocol has been used: 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 detected files.

There are not many missed detections, GMDR = 3 % for the two test sets. This value can be considered as acceptable for our application because the sound extraction system will be coupled with other sensors. False alarms remain below 15 %, GFAR ≈ 12 % in the same conditions.

Table 3 - Sound system performances on 1577 tests

	SNR∈[10,20] dB	SNR∈[0,40] dB
GMDR	3 %	3 %
GFAR	12.3 %	12.7 %

5. CONCLUSIONS AND PERSPECTIVES

In this paper we have presented a multichannel smart sound sensor which detects sound events and identifies sounds among 7 predefined sound classes. The smart sound sensor was designed to work in the framework of a medical telemonitoring application. We have proposed an evaluation methodology in relation with the application. The obtained results in real noisy environments may be acceptable.

Actually, the sensor is composed of the data acquisition card plugged in a PC. To make it physically independent, we can implement the sensor system inside the digital signal processor card. We have chosen to test the sensor algorithms using a PC because of the facilities in terms of implementation and verification. This sensor application was the telemedicine field, but it can be generalized to the perceptive spaces.

ACKNOWLEDGEMENT

This work is a part of the DESDHIS-ACI "Technologies for Health" project of the French Research Ministry. This project is a collaboration between the Center MICA (Hanoi - Vietnam), the CLIPS laboratory and the TIMC laboratory (Grenoble - France).

REFERENCES

[1] N.I., *"LabWindows/CVI User Manual"*, National Instruments Corporation, December 1999.

[2] Real World Computing Partnership, "CD - Sound scene database in real acoustical environments," http://tosa.mri.co.jp/sounddb/indexe.htm, 1998-2001.

[3] Can Bus site, *"http://www.can.bosh.com"*, 2003

[4] Takeshi Yamada & Narimasa Watanabe, *"Voice activity detection using non-speech models and HMM composition,"* in Workshop on Hands-free Speech Communication, Tokyo, Japan, 2001.

[5] A. Dufaux, *Detection and Recognition of Impulsive Sounds Signals*, Ph.D. thesis, Faculté des sciences de l'Université de Neuchâtel, 2001.

[6] L. Daudet, *Représentations structurelles de signaux audiophoniques - Méthodes hybrides pour des applications à la compression*, Ph.D. thesis, Marseille, 2000.

[7] M. Vacher, D. Istrate, L. Besacier, E. Castelli & J.F. Serignat, *"Smart audio sensor for telemedicine,"* in Smart Objects Conference 2003, Grenoble, France, 15-17 May 2003.

[8] F.K. Lam & C.K. Leung, *"Ultrasonic detection using wideband discret wavelet transform"*, in IEEE TENCON, August 2001, vol2, pp. 890-893.

[9] D. Reynolds, *"Speaker identification and verification using Gaussian mixture speaker models,"* in Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 1994, pp. 27–30.

[10] G. Schwarz, *"Estimating the dimension of a model"*, Annals of Statistics, vol.6, pp. 461-464, 1978.