

Information Extraction from Sound for Medical Telemonitoring

Dan Istrate, Eric Castelli, Michel Vacher, Laurent Besacier and Jean-François Serignat

Abstract— Today, the growth of ageing population in Europe needs an increasing number of health-care professionals and facilities for aged persons. Medical telemonitoring at home (and more generally telemedicine) improves the patient's comfort and reduces hospitalization costs. Using sound surveillance as an alternative solution to video telemonitoring, this paper deals with the detection and classification of alarming sounds in a noisy environment. The proposed sound analysis system can detect distress or everyday sounds everywhere in the monitored apartment and is connected to classical medical telemonitoring sensors through a data fusion process. The sound analysis system is divided in two stages: sound detection and classification. The first analysis stage (sound detection) has to extract significant sounds from a continuous signal flow. A new detection algorithm based on Discrete Wavelet Transform (DWT) is proposed in this paper, algorithm which leads to accurate results when applied to non-stationary signals (such as impulsive sounds). The algorithm presented in this paper was evaluated in a noisy environment and is favourably compared to the state of the art algorithms in the field. The second stage of the system is sound classification which is based on a statistical approach to identify unknown sounds. A statistical study was done to find out the most discriminant acoustical parameters in the input of the classification module. New wavelet based parameters, better adapted to noise, are proposed in this paper. The telemonitoring system validation is presented through various real and simulated test sets. The global sound based system leads to a 3% Missed Alarm Rate and could be fused with other medical sensors to improve performance.

Index Terms— Sound Detection, Wavelet Transform, Sound Classification, GMM, Medical Telemonitoring.

I. INTRODUCTION

THE actual growth of ageing population in Europe needs an increasing number of health-care professionals and facilities for aged persons. A possible solution to this problem is *telemedicine*, the practice of distance medicine by means of telematic tools which includes a wide variety of tasks like telediagnosis, distance teaching and learning, telesurveying [1] and distributed database applications. All these tasks involve the sharing of knowledge, data, expertise and services among health-care professionals. Medical telemonitoring at home, a telemedicine application, is an interesting solution compared to health facility institutions for the elderly since it offers a medical surveillance in a familiar atmosphere for the patient.

Information technology is taking an important role in the progress of health-care service. Many applications have already shown that a rational use of telemedicine becomes a cost-effective solution in the treatment of elderly patients [2], [3].

This research has been supported by the French Ministry of Research and is the result of a collaboration between the CLIPS laboratory and the MICA Center.

Despite the large number of advantages, there are some problems with telemedicine like: software complexity, data compression, data transmission and software reliability. A solution to simplify the software complexity and to increase its reliability is the use of specific informatics agents [4]. Data compression for medical information is submitted to severe constraints in order to conserve all the important medical details [5]. The choice of the transmission channel is also difficult; the Ethernet network, WAP (data transmission through a GSM terminal) [6] and television cable [7] are some of the solutions investigated.

Most of the systems used in this field only take into account medical sensors (blood pressure, pulse, oxymeter) and localization sensors (infrared or contacts) to survey patient [8], [9], [10]. Current systems use sound and video but these supports are dedicated only for communication between the patient and the medical staff [11] and not for patient telesurveillance.

In this paper, we present a system for the detection and the classification of everyday life sounds. The aim of our research is to develop a medical supervising system using sound sensors. The telemonitoring system must cover all the areas of the apartment, including the toilets, the bathroom and the bedroom. If a video camera is installed in every room the patient could have the uncomfortable feeling of being spied on. On the other hand, a sound sensor is more discreet and the patient's privacy is less disturbed as there is no continuous recording of the sound in the room but only a real-time analysis is applied to the the last 10s audio capture.

The originality of this research is to use sound as an informative source simultaneously with other sensors. We propose to extract and classify everyday life sounds such as: door banging, glass breaking, sounds of doing the dishes or falling objects or persons sounds, etc. in the aim of detecting serious accidents such as falls or somebody fainting everywhere in the apartment. Thus, our approach consists of replacing the video camera by a system of multichannel sound acquisition that analyses the sound environment of the apartment in real time and detects distressful situations.

In order to respect privacy, no continuous sound recording is made. Only the latest detected sound event is kept and sent to the alarm monitor if it is considered to be a possible alarm. This signal can also be used by a human operator to take the decision of a medical intervention.

In order to reduce the computation time necessary for a multi-channel real time system, the sound extraction process has been divided in two stages: detection and classification. The sound event detection is a complex task because the audio signals occur in a noisy environment.

Firstly, the medical context and the global telemonitoring

system are introduced in section II. The two steps of the system process are described in section III, the detection stage, respectively in section IV, the classification stage. The way of coupling these two steps has an important influence on the sound classification. Two possible cases of realizing this coupling are discussed in section V of the paper. In order to evaluate the proposed system, we have collected a sound corpus, which is described in section VI. The performance of every step of the system has been evaluated individually in a noisy environment as well as the performance of the global system. These results are illustrated in section VII. The system characteristics, its strengths and applications are presented in section VIII.

II. THE TELEMONITORING SYSTEM

A. About Telemetry

The living area used in our experiments is a 30 m² apartment situated in the TIMC¹ laboratory buildings. The rooms are equipped with medical sensors: blood pressure sensor, oxymeter and a weighting scale, infrared position sensors and sound sensors. The sensor location is the following: the microphones and the infrared sensors are distributed in every room (kitchen, hall, living-room, shower-room and toilet) while the most used medical sensors are wireless. The telemetry system architecture is made up of two computers which exchange information through an Ethernet network as presented in *Figure 1*. The sound extraction and analysis system has a dedicated PC (Sound Analysis PC) which acquires the signals from all five microphones.

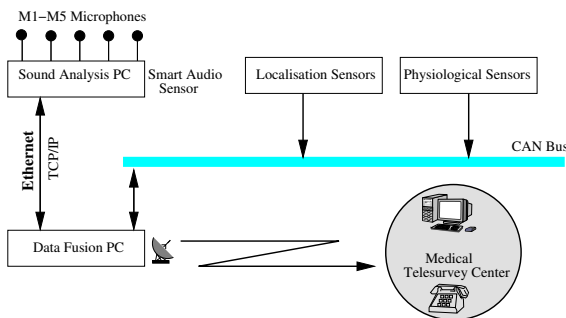


Fig. 1. Acquisition and analysis system

The other PC (Data Fusion PC in *Figure 1*) collects data from fixed and moving sensors as well as the information from the smart audio sensor (the Sound Analysis PC). Depending on the information provided by the sound analysis PC and the rest of sensors, the Data Fusion PC will send an alarm if necessary.

This paper will focus only on the smart audio sensor which will be described in the following.

B. The Sound Analysis

From a daily patient surveillance, a sound signal flow is continuously analysed. Among different everyday life sounds, only some of them are considered alarming sounds: glass breaking, screams, etc. In order to have a reliable sound telemetry

¹Techniques de l'Imagerie, de la Modélisation et de la Cognition (Image, Modelling and Cognition Techniques)

system, every sound event is detected (a sudden change in the environmental noise), extracted and used as input for the classification stage. This second step of the system aims to identify the sound type among several predefined classes which are detailed in section VI.

The sound analysis system has been divided in three modules as shown in *Figure 2*. The advantage of this division is to make real time implementation possible. Other methods (sound segmentation using a Hidden Markov Model or Bayesian Information Criterion joint with a Gaussian Mixture Model) which involves more complex models, would not allow real time processing.

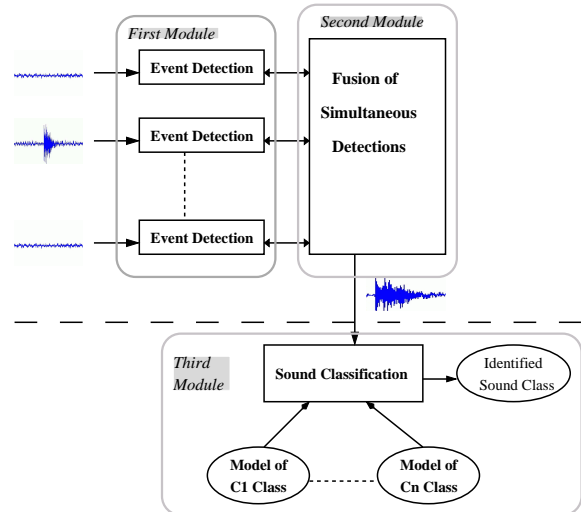


Fig. 2. Sound analysis

The first module is applied to each channel in order to detect sound events and to extract them from the signal flow. The source of sound or speech can be localized by comparing the estimated Signal to Noise Ratio (SNR) for each channel. The fusion module selects the best channel if several events are detected simultaneously. The third module receives the sound event extracted by the previous module and it estimates the most probable sound class.

The system has been designed to respond to several constraints: the real time five channels analysis, the wide dynamic amplitude of the useful signal, the use of this system 24h/24h, the eventual presence of a lot of non-stationary noise, the wide variety of sounds needed to be classified.

III. DETECTION

A. Method

In medical telemetry systems, the signal detection is very important because if a sound event is lost during the first stage of the system, it is lost forever. On the other hand, if there are too many false alarms (signal detected when nothing has occurred) the recognition system is saturated. Therefore, the performance of the detection algorithm is very important for the entire system.

Detection consists of identifying the desired signals in a

noisy environment. The two hypothesis of binary detection are:

$$\begin{cases} H_0 : & o(t) = b(t) \\ H_1 : & o(t) = s(t) + b(t) \end{cases} \quad (1)$$

where $o(t)$ is the analyzed signal, $b(t)$ is the noise and $s(t)$ the signal to be detected. The basic function of a detection algorithm is to extract some measured features or quantities from the input signal and to compare these values with a threshold.

Signal detection is a wide domain that includes: detection of numerical signals in noise [12], radar signals detection, voice activity detection. There are various possibilities in which to define the measured features, for example: energy, likelihood of a statistical model, high-order statistics [13]. Most of the existing systems try to detect the human voice (voice activity detection) and not the impulsive sounds [14]. Voice detection is based on speech properties such as pitch, spectral characteristics [15], Linear Predictive Coefficients (LPC) [16], [17]. There has not been a lot of work done in dealing with impulsive sound detection. Dufaux (2001) proposed three algorithms of impulsive sound detection with good results only in white noise: one based on the variance of the signal energy and the two other algorithms based on the conditioning median filtering of the energy [18]. The algorithm based on conditioning median filtering is used in our study as a state of the art algorithm; its measured feature is the difference between signal energy and the conditioning median filtered energy.

In our first experiments, we noticed that the environmental noise of experimental apartment had quite different properties than of white noise, which lead to a decrease of system performance. This constraint has directed our research on the improvement of the algorithms, notably in the environmental noise which has mainly low frequency components and includes impulsive sounds coming from the neighborhood of the apartment.

B. Proposed detection algorithm based on Wavelet filtering

Compared to the Fourier Transform, the Wavelet Transform is better adapted to signals which have very localized features in the time-frequency space. Therefore, this transform is often used in signal detection and audio processing [19], [20] because of its non-uniform time and frequency resolution.

All signals $x(t)$ can be decomposed in a sum of functions $\psi_{u,s}(t)$ localized and weighted by $\kappa_{u,s}$:

$$x(t) = \sum_{u,s} \kappa_{u,s} \psi_{u,s}(t) \quad (2)$$

where u is the time shift (a constant for Fourier Transform) and s is the scale factor. The type of $\psi_{u,s}(t)$ makes the difference between the Short Time Fourier Transform ("frequency" analysis) and the Wavelet Transform ("time-scale" analysis).

The Discrete Wavelet Transform (DWT) has non-uniform frequency and respectively time resolution. The time resolution, on the contrary to the frequency one, is greater in high-frequencies and poor in low frequencies which involve that DWT is preferred for impulsive signal detection. The wavelet base is generated by translation and dilatation of the

mother wavelet ψ . In signal processing applications (noise filtering, signal compression), the Daubechies wavelets are used as mother wavelet due to their properties: good regularity for high number of moments. In the proposed algorithm we use Daubechies wavelets with 6 vanishing moments in computing the DWT [21], [22].

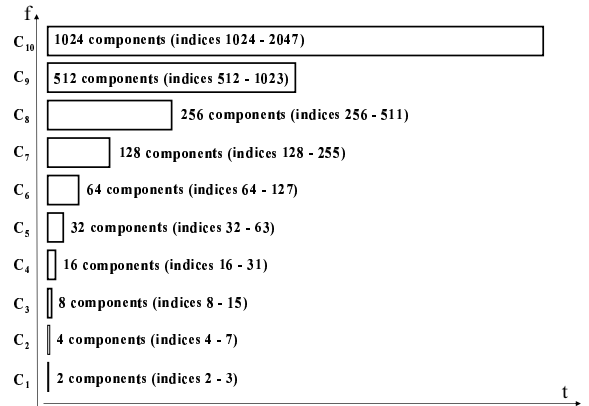


Fig. 3. Layout of the resulting wavelet-transform matrix (frame of 2048 samples)

Discrete Wavelet Transformation is applied to the sampled data and its output forms a vector with the same length as the signal. This vector has a pyramidal structure and is composed of 12 *wavelet transform coefficients* for a frame of 2048 samples. The layout of the coefficients in the vector is given in *Figure 3*.

The proposed algorithm (flowchart in *Figure 4*) calculates the energy of three upper wavelet transform coefficients (higher order coefficients which contain respectively 1024, 512 and 256 samples), because the significant wavelet coefficients of the sounds to be detected are of a rather higher order (corresponding to high frequency in the sounds). The analysis frame length is 128 ms (2048 samples) because of the real time acquisition constraint, but, for a better time resolution analysis, the DWT output vector is analysed by a window of 32 ms (4 windows of 32 ms inside the DWT frame). Thus, the detection threshold is applied to a three depth wavelet tree.

Finally, to complete detection, the system applies a threshold to the sum of energies in the three depth wavelet tree. The threshold is self-adjustable and depends on the average of the N last energy values (in this study we have used 40 values for a statistical representation).

An example of signal detection is shown in *Figure 5* where a phone ring, starting at $t = 3$ s, is mixed with flowing water noise at 0 dB of SNR (upper window). In the lower window of the same figure the energy of the three wavelets coefficients is presented in black and the adaptive threshold in the dotted line. We can see that the phone ringing signal is detected by the proposed algorithm.

IV. SOUND CLASSIFICATION

A. Method

Pattern recognition domain uses many techniques, such as: **G**aussian **M**ixture **M**odel (GMM) [23], **H**idden **M**arkov

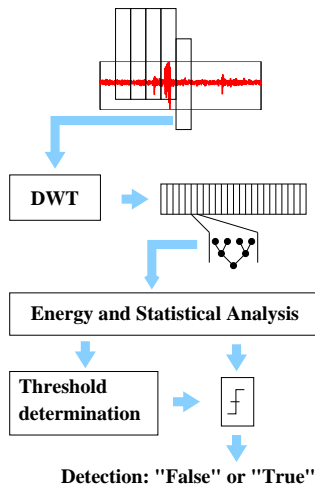


Fig. 4. Flowchart of the wavelet based algorithm

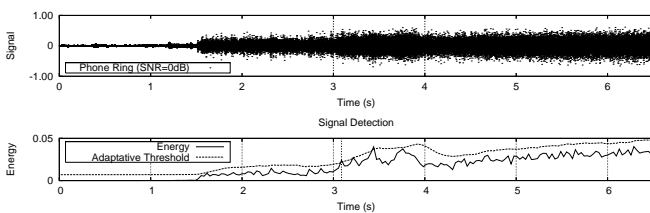


Fig. 5. Detection example of a phone ringing mixed with flowing water noise at SNR=0 dB

Models (HMM), Dynamic Time Warping (DTW), neural networks and others. Since sound classification is a sub-domain of the pattern recognition, all these different techniques can be used. Existing studies on environmental sound classification are quite limited and still at a preliminary stage. Woodard [24] uses the HMM method to classify sounds but his corpus has a few number of sound classes (only 3) and the noise presence is not taken into account. Another classification method is presented by Papadopoulos et al. in [25]. This method is based on a comparison between normalized spectrum and learned sound classes spectrum and the results are given only for 3 sound classes. A comparison between neural network, DTW and vector quantization is presented by Cowling [26]. The results are done for 8 sound classes but the sound database duration is only 35s.

The GMM method is flexible with regard to signal type and it performs well in speaker/sound recognition, as demonstrated by Reynolds [27], which are the main reasons for choosing this method in our study. The HMM are more complex with longer computation time and are not very well adapted to short signal classification. The results obtained by Dufaux [18] with a 3 stage HMM are similar to GMM classification. The GMM method works in two stages: training and classification.

a) Training.: For each class of signals (ω_k) from the corpus a training stage is initiated in order to obtain a model containing the characteristics of each Gaussian distribution (m) of the class: the weight of the Gaussian ($\pi_{k,m}$), the average vector ($\mu_{k,m}$) and the covariance matrix ($\Sigma_{k,m}$). These values are calculated after M iterations ($M=20$) of the "EM" algorithm (Expectation Maximization) [28], which follows a K-Means al-

gorithm. The covariance matrices are diagonal.

b) Recognition.: 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 stage. The size of acoustical vectors (d) is the number of acoustical parameters extracted from the signal. The likelihood of each acoustical vector given for a class ω_k is calculated within the following formula (N is the number of Gaussian distributions):

$$\begin{cases} p(x_i | \omega_k) = \sum_{m=1}^N \pi_{k,m} \cdot \frac{1}{\sqrt{(2\pi)^d |\Sigma_{k,m}|}} \cdot e^{A_{i,k,m}} \\ A_{i,k,m} = \left(-\frac{1}{2}(x_i - \mu_{k,m})^T \cdot \frac{1}{\Sigma_{k,m}} \cdot (x_i - \mu_{k,m}) \right) \end{cases} \quad (3)$$

The likelihood of the entire signal (n frames) is thus given by the following equation:

$$p(X | \omega_k) = \prod_{i=1}^n p(x_i | \omega_k) \quad (4)$$

n represents the number of signal frames. The algorithm determines that the signal X will belong to the class ω_l in which $p(X | \omega_l)$ is maximum.

B. Acoustical parameters

Sound classification does not use direct sound signals but a parametric signal representation in order to eliminate redundancies. In speech classification, the classical acoustical parameters are: MFCC (Mel-Frequency Cepstral Coefficients) [29], LFCC (Linear Frequency Cepstral Coefficients), LPC (Linear Prediction Coefficients), etc. This paper proposes a set of acoustical parameters based on wavelet transform and three acoustical parameters traditionally used in speech/music/noise segmentation: Zero Crossing Rate, Centroid and Roll-off Point. First and second derivative of the acoustical parameters (called Δ and respectively $\Delta\Delta$) are also used in order to introduce the temporal variation of the signal in the GMM modelling [30].

The MFCC parameters are calculated as follows: direct Fast Fourier Transform (FFT), the computation of the energy of 24 non-uniform triangular filters (Mel-Scale), logarithm application on energy values, inverse Discrete Cosine Transform (DCT). The LFCC parameters are calculated in the same way, but the triangular filters are uniform. LPCC parameters are the cepstral of LPC coefficients which represent the vocal tract filtering model.

a) Zero Crossing Rate (ZCR).: The value of the zero-crossing rate is given by the number of zero-voltage crossings in the analysis frame. In order to eliminate noise influence, we have introduced a symmetric clipping threshold. The value of the clipping threshold represents 0.03% of the signal amplitude. In fact, the zero-crossing rate indicates the dominant frequency in frame.

b) Roll-off Point (RF).: This feature is used to measure the frequency which takes 95% of the power spectrum. The roll-off point can be viewed as a measure of the "skewness" of

the spectral shape. The value is higher for right-skewed distributions. The value of the roll-off point is the solution of *equation (5)* with $\Theta = 0.95$.

$$\sum_{k < \text{RF}} X[k] = \Theta \sum_k X[k] \quad (5)$$

c) Centroid.: The centroid represents the balancing point of the power spectrum distribution within a frame [31]. The centroid for a frame at a specific time is computed as the roll-off point, *equation (5)*, where $\Theta = 0.5$.

d) The acoustical parameters proposed: Wavelet Based Coefficients.: The wavelet transform applied in speech recognition has not been studied a lot [32] despite its better time resolution in high frequencies. The acoustical parameters proposed are based on the Discrete Wavelet Transform similar to classical cepstral coefficients. This acoustical parameter type will be referred as DWTC. Firstly the Discrete Wavelet Transform (DWT) is computed in a 256 samples window. Secondly, the energies of the last six wavelet transform coefficients are calculated and followed by a logarithmic amplitude transformation (by analogy with MFCC). The final acoustical vector contains the DWT^{-1} logarithmic energy coefficients. The total number of parameters is six.

1) Selection of the Acoustical Parameters: In order to find the relevant acoustical parameters for classification, a statistical study has been conducted using the Fisher Discriminant Ratio (FDR).

The FDR (*equation (6)*) gives an indication of the separation capacity of every acoustical features. In *equation (6)*, the average of parameter x for the class i is $\overline{x[i]}$, the standard deviation of parameter x for the class i is $\text{Var}(x)[i]$ and the number of sound classes is k .

$$FDR = \frac{\sum_{i=1}^k \sum_{j=1}^k (\overline{x[i]} - \overline{x[j]})^2}{\sum_{i=1}^k \text{Var}(x)[i]} \quad (6)$$

The results of this study are presented in paragraph VII-C.

V. COUPLING BETWEEN DETECTION AND CLASSIFICATION

The final module of the system implements the coupling of the detection system with the classification one. The critical point of the coupling for the classification stage is the precision of the sound delimitation before sending the detected signal to the classification system. The possible errors in the sound delimitation are:

- early detection of the sound (a part of the extracted signal contains silence only);
- signal detection with a delay (a part of the sound is eliminated). This error has a reduced influence on the GMM classification system, because the signal time progression is not taken into account by the GMM method.

One possible approach is to consider detection output as a fix duration sound. However, the sound classification system is

very sensitive to the parts which are silent for a long time. In order to solve this problem we propose to use the detection of the end of the signal. The end is detected by applying the same detection algorithm on the time inverted signal.

The chosen configuration for the coupling of detection and classification involves the following steps:

- 1) the output signal is extracted at the detection time; its duration is seven seconds (the maximum sound duration of our test set).
- 2) the signal is time inverted and the detection algorithm is applied once again.
- 3) the detection of the signal end is used to cut the sound; the resulting signal, sent to the classification system has a variable length of time.

Through this procedure, the classification system analyzes only the typical part of the detected signal, which has a variable length of time.

VI. THE SOUND DATABASE

In order to test and validate the event detection system and the sound recognition system we have collected a sound corpus [33]. It contains recordings made in the Clips laboratory (15% of the CD), the files of "Sound Scene Database in Real Acoustical Environments" (70% of the CD) [34] and files from a commercial CD (film effects, 15 % of the CD) [35]. There are 3354 files and every file is sampled both at 16 kHz and 44.1 kHz.

At Clips laboratory the sounds were recorded with a Beyer Dynamics microphone and a digital tape (sampling rate 44.1 kHz), being transferred to the PC through its sound card. The sound corpus contains: door banging sounds (different types of doors), chair sounds, walking sounds, electric shaver sounds, hairdryer sounds, door locking sounds, dishes sounds, glass breaking, falling objects sounds, screams, flowing water sounds, telephone or door bell ringing, etc. To summarize, the sound corpus contains 20 types of sounds with a minimum of 10 repetitions per type (the maximum is 300 repetitions) $\approx 3\text{h}$ total signal time.

A. Detection Test Set

In order to validate the detection algorithms we have generated a test set which is a mixture of environmental noises and useful sounds. For every sound, there are two signals in the test set: one contains the mixture between the sound and the noise (file with event) and the other one with only the noise (file without event). Every sound and noise has been recorded three times. Each file is 25s long (because of the length of the sound and of the time necessary to initiate the algorithms, which is approximately 5s). The sound starts at 10th second of the signal. In the test signal base, we consider three types of noise (white noise, flowing water noise and environmental noise recorded in the apartment) and 11 types of sounds (screams, falling chairs, falling book, glass breaking, door banging, walking sounds, coughs, sneezes, door locking, phone ringing and speech). For every mixture "sound-noise", there are 4 files with 4 signal to noise ratios (SNR): 0 dB, 10 dB, 20 dB and 40 dB. The SNR is calculated on the total time length of the sound.

In order to validate the results obtained from the simulation test set we have recorded 60 files inside our testing apartment (real life conditions) at different SNR ($2\text{dB} \leq \text{SNR} \leq 30\text{dB}$ with an average of 15 dB). We have used the same sounds (played with a loud speaker) as in the simulation test set.

B. Classification Test Set

The test set used for the sound classification is composed of 7 sound classes: door banging(523 sounds), phone ringing (517 sounds), walking sounds (13 sounds), sounds of doing the dishes (163 sounds), door locking (200 sounds), glass breaking (88 sounds), screams (73 sounds). There are five sets in the seven classes: one contains pure sounds and in the other four, mixtures of sound and environmental noise (named HIS noise) at 0, 10, 20 and 40 dB signal to noise ratio.

C. Coupling Test Set

This test set is used for the performance evaluation of coupling between detection and sound classification and for the validation of the entire audio information system. This final test set contains all the sounds of the recognition test set superposed to HIS environmental noise. There are seven files corresponding to the seven sound classes. Each file is made up of a succession of signals corresponding to all the sound classes and periods of silence at random duration. The SNR for each sound has a random value between 10dB and 20dB with an uniform repartition. Silence between consecutive sounds varies randomly between 5 to 60 seconds. The total number of useful sounds to be detected is 1577.

VII. EXPERIMENTAL RESULTS

A. Evaluation of the Detection Algorithm Performance

Missed Detection Rate (R_{MD}) and False Detection Rate (R_{FD}) are used to characterize detection performance in our test set. They are calculated according to the *formulas* (7) and (8).

$$R_{MD} = \frac{\text{No. missed detections}}{\text{No. events to detect}} \quad (7)$$

$$R_{FD} = \frac{\text{No. false detections}}{\text{No. false detections} + \text{No. events to detect}} \quad (8)$$

A detection is considered to be *false* if an event is detected while there is no real event. A detection is considered to be *missed* when the system detects nothing in the interval: 0.5s before the event and the end of the signal event. A detection occurring during this interval is considered to be a *good* detection of the event.

To compare the algorithms we have determined the equal error rate (EER) from ROC (Receiver Operating Characteristics) curves, defined as the value of R_{MD} for which $R_{MD} = R_{FD}$ (the intersection between ROC curve and the first bisector). The ROC curve plots missed detection rate versus false detection rate.

B. Detection Results with the Test Set

The evaluation of the state of the art algorithm and the wavelet based algorithm on the detection test set is illustrated in *Table I*. The first column represents the algorithm, the second the SNR and two other columns the EER for white noise and HIS noise. For each noise, the performance are presented with a SNR of 0, 10, 20 and 40 dB. Note that HIS noise is the environmental noise recorded in our experimental apartment. For wavelet based algorithm, a constant of the detection threshold (an offset of self-adjustable threshold) is varied in order to obtain a variation of R_{MD} and R_{FD} from 0 to 1.

TABLE I

THE PERFORMANCE OF WAVELET BASED ALGORITHM ARE IN BOLD IN THE TABLE AND COMPARED TO THE STATE OF THE ART ALGORITHM

Detection algorithm	SNR [dB]	EER for different type of noise	
		HIS noise [%]	White noise [%]
Wavelet based	0	7.3	6.1
	+10	0	4.0
	+20	0	0
	+40	0	0
State of the art : Median conditioning filtering	0	64.9	30
	+10	35.8	0
	+20	10.4	0
	+40	0	0

To analyse the results, we must principally compare the corresponding performance of HIS environmental noise and low SNR (real life environmental conditions).

The state of the art algorithm (*median conditioning filtering*) is not suitable because $\text{EER} > 10\%$ for a value of $\text{SNR} \leq 20\text{dB}$. The new algorithm based on *wavelet filtering* gives the best results for HIS noise: $\text{EER}=0\%$ for $\text{SNR} \geq +10\text{dB}$ and $\text{EER}=7.3\%$ for $\text{SNR}=0\text{dB}$. The results are not very good for white noise ($\text{EER}=4\%$ for $\text{SNR}=10\text{dB}$), but are still better in comparison with the state of the art algorithm.

The results shown in *Table I* obtained from the simulation test set are confirmed by the *real detection test set* (paragraph VI-A). The wavelet based algorithm gives an EER of **0%** for this real test set.

C. Sound Classification

1) *Model Selection*: The Bayesian Information Criterion (BIC) is used in this paper in order to determine the optimum number of Gaussians [36]. BIC criterion expressed by *equation* (9) selects the model through the maximization of integrated likelihood.

$$BIC_{m,K} = -2.L_{m,K} + \nu_{m,K} \ln(n) \quad (9)$$

where $L_{m,K}$ is the logarithm of likelihood maximum, equal to $\log f(x | m, K, \hat{\theta})$ (f is integrated likelihood), m is the model, K the component number of the model, $\nu_{m,K}$ is the number of

free parameters of the model m and n is the frame number. The minimum value of BIC indicates the best model.

The BIC has been calculated from the sound class with the smallest number of files, for 2, 4, 5 and 8 Gaussian. The results presented in *Table II* are obtained from 16 MFCC parameters. Looking at these results, a number of Gaussian between 3 and 5 seems to correspond to the best choice within our available training data.

TABLE II

BIC FOR 2, 3, 4, 5 AND 8 GAUSSIAN (OPTIMAL VALUES FOR BIC ARE IN BOLD IN THE TABLE)

No. Gaussian	2	3	4	5	8
BIC	11043	10752	10743	10757	13373

2) Statistical study for the choice of efficient parameters:

This statistical study shows the relevant acoustical parameters using the Fisher Discriminant Ratio (FDR) expressed by the *equation (6)* and allows us to reduce the number of testings. *Table III* shows FDR values in several acoustical parameters. The number that follows the name of acoustical parameters represents the parameter position in acoustical vector (MFCC1 is the first MFCC parameter).

TABLE III

ACOUSTICAL PARAMETERS WITH HIGH FDR ($FDR \geq 2$). THE CHOSEN PARAMETERS FOR TESTS ARE IN BOLD IN TABLE.

Parameter	FDR	Parameter	FDR	Parameter	FDR
MFCC1	2.72	MFCC10	3.34	Centroid	23.75
MFCC2	16.07	MFCC11	2.88	Energy	2.54
MFCC3	10.33	MFCC12	3.20	DWTC3	2.89
MFCC4	10.02	MFCC14	3.61	DWTC4	4.54
MFCC5	2.01	MFCC15	3.26	DWTC5	6.02
MFCC6	2.91	MFCC16	4.41	DWTC6	8.69
MFCC7	3.36	ZCR	18.00		
MFCC8	3.60	RF	16.70		

Given the results shown in *Table III* we can say that the second, third and fourth MFCC coefficients are the most relevant MFCC parameters in separating our sound classes. ZCR, RF, Centroid and the three wavelet based coefficients (DWTC) are relevant parameters; on the contrary, energy seems irrelevant for instance.

3) *Sound Classification results:* The analysis window (for the calculation of the acoustical parameters) was set at 16 ms with an overlap of 8 ms, values usually used in speech/speaker recognition. The GMM model is made of 4 Gaussian distributions. In these experiments, each of the 1577 sounds in the database is used as a test with the so called "leave one out" protocol: when a sound is used as a test, it is *not* used in training step, so the training set consists of the whole database except the test sound.

TABLE IV

SOUND CLASSIFICATION RESULTS FOR PURE SOUNDS. IN BOLD IN THE TABLE THE BEST COMPROMISE PERFORMANCE/COMPLEXITY AND THE FDR SUGGESTED ACOUSTICAL PARAMETERS

Parameters	PN	CER [%]
$\Delta, \Delta \Delta (16\text{MFCC} + \text{Energy} + \text{ZCR} + \text{RF} + \text{Centroid})$	60	8.7
16 MFCC + Energy + ZCR + RF + Centroid	20	11.4
16LFCC+Energy	17	12.2
16LFCC+ZCR+RF+Centroid	19	12.7
16LPCC+Energy	17	14.7
16MFCC+Energy	17	15.2
3MFCC+ZCR+RF+Centroid	6	16.1
DWTC	6	18.7

Experimental results are presented in *Table IV*, showing the average classification error rate (CER: number of recognition errors divided by the number of tests) and the corresponding number of parameters (PN). We can observe that good results are obtained with MFCC parameters (speech specific parameters) but new parameters like zero crossing rate, roll-off point and centroid seem interesting when combined with conventional parameters used in speech. The best results are obtained with 60 acoustical parameters, the first and the second derivatives of 16MFCC, Energy, ZCR, RF and Centroid. These parameters are denoted in the *Table IV* as $\Delta, \Delta \Delta (16\text{MFCC} + \text{Energy} + \text{ZCR} + \text{RF} + \text{Centroid})$.

The three MFCC coefficients have been tested in conjunction with zero crossing rate, roll-off point and centroid as suggested by the statistical study (*Table III*). We have noticed that the parameters considered to be irrelevant after the statistical study can be eliminated with practically no negative influence on the performance of the system; showing a drastical reduction of the number of parameters (6 instead of 20) produces only a 4.5% increase in the classification error rate (in bold in *Table IV*).

4) *Performance in noisy environment:* Our classification system has been tested in HIS noise situation with two types of training: training only on pure sounds or on pure sounds and noisy sounds.

Training on pure sounds gives constant results for $\text{SNR} \geq 20$ dB; the CER decreases beyond this point: for 16 MFCC + ZCR + RF and 16 LFCC parameters, classification error is 26.82% for $\text{SNR} = +10$ dB (*Figure 6*). These results are not acceptable since the SNR in the testing apartment varies between 10 and 20 dB.

DWTC parameters show greater performance than classical MFCC parameters for $\text{SNR} \leq 10$ dB and only 6 parameters are needed for classification, while in the other case a minimum of 17 is required.

The training step on the noisy sounds takes into account several cases: training only on the noisy sounds for a particular SNR, or on a combination between pure sounds and noisy sounds. Better results are obtained from the combination between pure sounds and noisy sounds at 10 dB SNR (*Figure 7*). Our tests suggest using a GMM class model for each SNR

which would involve SNR estimation before the classification stage.

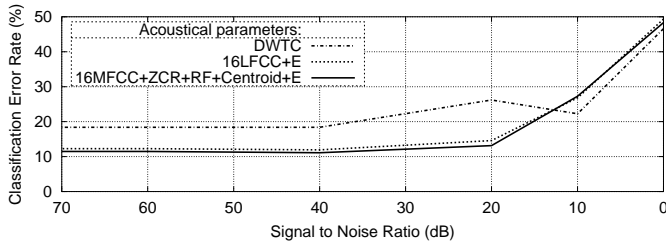


Fig. 6. Classification error in HIS noise (training only on the pure sounds)

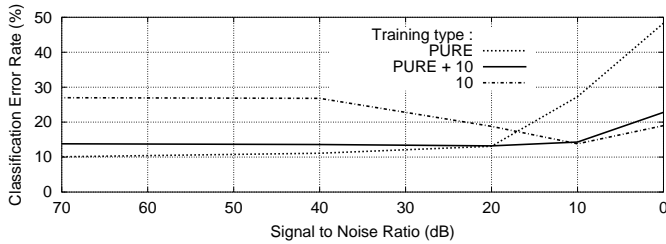


Fig. 7. Classification error in HIS noise (training on the pure sounds and noisy sounds)

D. Evaluation of coupling between detection and classification

In order to evaluate the coupling between detection and classification, we use the wavelet detection algorithm, the same GMM classification system and the detection of the end of the signal. The evaluation is made on the coupling test sets. The threshold of the detection algorithm was set to an optimal value on the detection test set. The used acoustical parameters are 16 MFCC together with ZCR, Roll-off point and Centroid. The GMM training is made on pure sounds with a leave one out protocol.

The efficiency of the coupling between detection and classification stages is studied, and the results are illustrated in Table V. The tests of efficiency have been done on 3 cases:

- the reference case: manual detection and signal delimitation (no algorithm errors);
- the coupling when the automatic detection algorithm and a fixed time length of signal are used;
- the coupling when the automatic detection algorithm and signal time length estimation are used.

The CER obtained for manual detection confirms the classification performance in noisy conditions. The results obtained with a fixed time length extracted signals are not acceptable. The error introduced by the lack of adapted coupling is approximately 46%. The detection of the end of the signal, significantly improves the classification performance. The difference between this coupling and the manual detection reflects the influence of false alarms and missed detections in the overall classification system.

E. Global alarm detection system evaluation

To evaluate the overall alarm detection system, all sound classes are divided into two parts: alarm sounds and non alarm

TABLE V
COUPLING EVALUATION (IN BOLD THE GLOBAL AUTOMATIC SYSTEM)

	CER for SNR 10-20 dB
Reference :	
Manual detection and real length of signals $R_M=0, R_F=0$	21.5 %
Automatic detection and fix length of signals $R_M=1\%, R_F=1\%$ $10 \text{ dB} \leq \text{SNR} \leq 20 \text{ dB}$ $R_M=5\%, R_F=2.6\%$ $0 \text{ dB} \leq \text{SNR} \leq 40 \text{ dB}$	67.8 %
Automatic detection and length estimation $R_M=1\%, R_F=1\%$ $10 \text{ dB} \leq \text{SNR} \leq 20 \text{ dB}$ $R_M=5\%, R_F=2.6\%$ $0 \text{ dB} \leq \text{SNR} \leq 40 \text{ dB}$	27.7 %

sounds.

The possible cases after detection are: good detection event (GD), false detection event (FD), missed detection event (MD). The detected events, GD and FD, are sent to the classification system. A part of the missed detection events does not have serious consequences (we call this MD W) because they belong to the non alarm sound classes.

The classification stage, similarly as the previous step, may produce: good alarms (GA), false alarms (FA) and missed alarms (MA). A part of the missed alarms has no bad consequences on the final results (we call this MA W). The Figure 8 illustrates an analysis of the possible errors of each stage of the system and their propagation.

The *Global Missed Alarm Rate* (GMAR) is defined in this paper as the sum of missed detection (MD) and missed alarm (MA). The *Global False Alarm Rate* (GFAR) is defined as the False Alarm Rate at the output of the sound classification stage and not the sum of false alarms generated by each system stage. Despite the fact that the False Detection errors generated by the detection stage are injected in the classification stage, these errors can influence or not the GFAR after the classification stage. For example, if a false detection is classified in a non alarm sound class, it is not used in the GFAR computation.

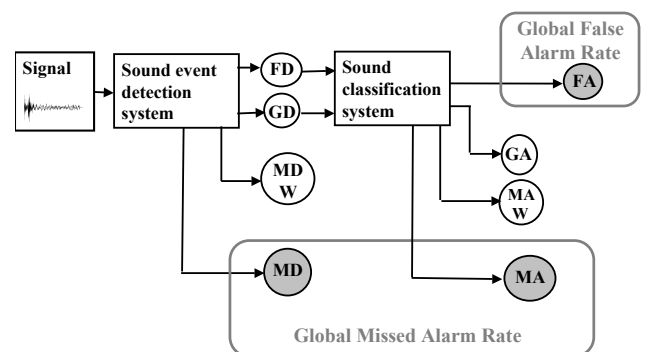


Fig. 8. Global False Alarms Rate and Global Missed Detection Rate

The *Global Missed Alarm Rate* finally obtained is 3% from

the global system. This value may be acceptable if the sound extraction system is joined to results from other sensors. The *Global False Alarm Rate* is 12% from the global system, which in the investigated working conditions represents approximately 12 false alarms per day. Some of these false alarms could be eliminated by the fusion of the sound analysis system with the classical medical telemonitoring system. For example, when the localization of the sound alarm does not correspond with that of infrared sensor output, the alarm is eliminated. The fusion between the sound analysis system and a classical medical telemonitoring system is the aim of a future research.

VIII. CONCLUSIONS

a) Focus on the obtained results.: A summary of the performance of our detection and classification system is illustrated in *Table VI*. Although some of the tasks were evaluated on a simulated data and still need to be validated in real conditions, we can say that we have proposed and tested efficient algorithms both for sound detection and classification, the goal of this paper.

TABLE VI

SUMMARY OF THE PERFORMANCE OF OUR DETECTION AND CLASSIFICATION SYSTEM. THE GLOBAL SYSTEM PERFORMANCE IS OUTLINED.

Task	Data	% Error
Detection	Real	0
Classification	Simulated	8.7
Global System	Simulated	3 (Missed Alarm Rate)

More precisely, the main results of this study can be summarized as follows:

- a sound detection algorithm has been proposed and validated; this new algorithm is based on the wavelet transform with good performance in a noisy environment (SNR between 0 and 10 dB);
- the new acoustical parameters resulting from wavelet transform are the best adapted to noise among the tested parameters;
- a technique to detect the end of the signal has also been proposed;
- an original methodology for the evaluation of the sound medical telesurvey system has been presented.

The algorithm for signal detection has been compared with the state of the art algorithm. The system allows us to detect a sound event in the apartment with a 7% error rate for a SNR of 0 dB. A GMM system has been implemented for sound classification. At the beginning, classical parameters usually used in speech recognition are tested. Later on, we proposed and tested non-conventional and new parameters resulting from the DWT. Non-conventional parameters like ZCR, RF and Centroid appear to be very discriminant for the sound classification while those resulting from the DWT seem better adapted

to environmental noise. The global system (detection coupled with classification) has an acceptable rate of 3% missed detection. The sound extraction process described here could also be applied to the classification of multimedia documents and to security surveillance.

ACKNOWLEDGMENTS

This system is part of the DESDHIS² project, a collaboration between CLIPS³ laboratory (UMR CNRS-INPG-UJF 5524), responsible for the sound analysis, and TIMC laboratory, charged with the medical sensors analysis and data fusion. This project is financed by the French Ministry of Research (ACI - "Technologies pour la Santé").

REFERENCES

- [1] I. Korhonen, J. Parkka, and M. V. Gils, "Health monitoring in the home of the future," *IEEE Eng. Med. Biol. Mag.*, pp. 66–73, May 2003.
- [2] R. L. Bashshur, "State-of-the-art telemedicine/telehealth: Ch.1 - telemedicine and health care," *Telemedicine Journal and e-Health*, vol. 8, no. 1, pp. 5–12, 2002.
- [3] P. A. Jennett, L. A. Hall, D. Hailey, A. Ohinmaa, C. Anderson, R. Thomas, B. Young, D. Lorenzetti, and R. E. Scott, "The socioeconomic impact of telehealth: A systematic review," *Journal of Telemedicine and Telecare*, vol. 9, no. 6, pp. 311–320, 2003.
- [4] V. D. Mea, "Agents acting and moving in healthcare scenario - a paradigm for telemedical collaboration," *IEEE Trans. Inform. Technol. Biomed.*, vol. 5, no. 1, pp. 10–13, Mar. 2001.
- [5] Z. Lu, D. Y. Kim, and W. A. Pearlman, "Wavelet compression of ECG signals by set partitioning in hierarchical trees algorithm," *IEEE Trans. Biomed. Eng.*, vol. 47, p. 849856, 2000.
- [6] K. Hung and Y. T. Zhang, "Implementation of a WAP-Based telemedicine system for patient monitoring," *IEEE Trans. Inform. Technol. Biomed.*, vol. 7, no. 2, pp. 101–107, June 2003.
- [7] R. G. Lee, H. S. Chen, C. C. Lin, K. C. Chang, and J. H. Chen, "Home telecare system using cable television plants - an experimental field trial," *IEEE Trans. Inform. Technol. Biomed.*, vol. 4, no. 1, pp. 37–44, Mar. 2000.
- [8] M. Takizawa, S. Sone, K. Hanamura, and K. Asakura, "Telemedicine system using computed tomography van of high-speed telecommunication vehicle," *IEEE Trans. Inform. Technol. Biomed.*, vol. 5, no. 1, pp. 2–9, Mar. 2001.
- [9] J. Reina-Tosina, L. Roa, and M. Rovayo, "NEWBET: telemedicine platform for burn patients," *IEEE Trans. Inform. Technol. Biomed.*, vol. 4, no. 2, pp. 173–177, June 2000.
- [10] E. Jovanov, A. D. Lords, D. Raskovic, P. G. Cox, R. Adhami, and F. Andrasik, "Stress monitoring using a distributed wireless intelligent sensor system," *IEEE Eng. Med. Biol. Mag.*, pp. 49–55, May 2003.
- [11] P. Varady, Z. Benyo, and B. Benyo, "An open architecture patient monitoring system using standard technologies," *IEEE Trans. Inform. Technol. Biomed.*, vol. 6, no. 1, pp. 95–98, Mar. 2002.
- [12] Y. Wu and S. Y. Kung, "Signal detection for MIMO-ISI channels: An iterative greedy improvement approach," *IEEE Trans. Signal Processing*, vol. 52, no. 3, pp. 703–720, Mar. 2004.
- [13] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 217–231, Mar. 2001.
- [14] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 109–118, Feb. 2002.
- [15] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505, Sept. 2003.
- [16] S. G. Tanyer and H. Ozer, "Voice activity in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 4, pp. 478–482, July 2000.
- [17] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, pp. 441–457, May 2001.

²Detection de Situations de Détresse en Habitat Intelligence Santé (Distress situations detection in a medical intelligent habitat)

³Communication Langagière et Interaction Personne-Système (Linguistic Communication, Human System Interaction)

- [18] A. Dufaux, "Detection and recognition of impulsive sounds signals," Ph.D. dissertation, Faculté des sciences de l'Université de Neuchâtel, 2001.
- [19] F. K. Lam and C. K. Leung, "Ultrasonic detection using wideband discrete wavelet transform," in *IEEE TENCON*, vol. 2, Tokyo, Japon, Aug. 2001, pp. 890–893.
- [20] M. Unser and T. Blu, "Wavelet theory demystified," *IEEE Trans. Signal Processing*, vol. 51, no. 2, pp. 470–483, Feb. 2003.
- [21] S. Mallat, *Une exploration des signaux en ondelette*, ser. ISBN 2-7302-0733-3. Palaiseau, France: Les Editions de l'École Polytechnique, 2000.
- [22] P. L. Dragotti and M. Vetterli, "Wavelet footprints: Theory, algorithms, and applications," *IEEE Trans. Signal Processing*, vol. 51, no. 5, pp. 1306–1323, May 2003.
- [23] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. SAP-3, no. 1, pp. 72–83, Jan. 1995.
- [24] J. P. Woodard, "Modeling and classification of natural sounds by product code hidden markov models," *IEEE Trans. Signal Processing*, vol. 40, no. 7, pp. 1833–1835, July 1992.
- [25] G. Papadopoulos, K. Efsthathiou, Y. Li, and A. Delis, "Implementation of an intelligent instrument for passive recognition and two-dimensional location estimation of acoustic targets," *IEEE Trans. Instrum. Meas.*, vol. 41, no. 6, pp. 885–890, June 1992.
- [26] 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, Australia, Jan. 2002.
- [27] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91–108, Jan. 1995.
- [28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of Acoustical Society of America (JASA)*, vol. 39, pp. 1–38, 1977.
- [29] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.
- [30] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, Feb. 1981.
- [31] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Lett.*, vol. 11, no. 2, pp. 258–261, Feb. 2004.
- [32] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE Signal Processing Lett.*, vol. 8, no. 7, pp. 196–198, July 2001.
- [33] CLIPS-IMAG Equipe GEOD - Dan Istrate, "Base de données. Sons de la vie courante," www-clips.imag.fr, Grenoble, France, November 2001.
- [34] R. W. C. Partnership, "CD - Sound scene database in real acoustical environments," <http://tosa.mri.co.jp/soundb/indexe.htm>, Tokyo, Japan, 1998–2001.
- [35] S. Sciascia, "CD - bruitages - vol.3," Paris, France, 1992.
- [36] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.



Dan Istrate Born in 1976, D. Istrate is an Associate Professor at ESIGETEL Fontainebleau - France being involved in embedded systems for sound processing. He defended his Ph.D. thesis in 2003 with the subject "Sounds Detection and Classification for Medical Telemonitoring" in CLIPS laboratory. He participated to RESIDE-HIS project and DESDHS, financed by IMAG federation and respectively by French Research Minister. He joined LIA laboratory in 2004 during his post-doctoral researches on the speaker diarization task with evolutive HMM (Multimodal Biometry) project financed by French Research Minister, to ESTER and NIST RT05 campaigns for speaker diarization systems evaluation.

Scientific topics: sound detection and classification, wavelet, real time processing, speaker diarization, embedded systems.



Eric Castelli Born in 1962, Eric Castelli is an Associate Professor at the "Institute National Polytechnique de Grenoble (INPG)". He defended his PhD thesis in electronic system field in 1989 and his HDR thesis in 1999. He worked at "Institut de la Communication Parlée (ICP)" laboratory from 1984 to 1997, and then he joined the CLIPS laboratory in January 1998. Now he works in Hanoi - Vietnam in the framework of an international co-operation project in order to setup a new Franco Vietnamese research laboratory: International Research Center MICA (Multimedia Information, Communication and Applications). He is now vice-director of the MICA Center in charge of the scientific program management. He published about 50 papers on various aspects of speech analysis, speech production, speech recognition & general instrumentation.

Scientific topic: Speech and sound processing, instrumentation, multimedia.



Michel Vacher Born in 1954, M. Vacher is Research Scientist at the French "Centre National de la Recherche Scientifique" since 1986. He received the Ph.D. degree in acoustical science from the INSA of Lyon, France, 1982. The subject of his work was the "Nonlinear Behaviour of Micro-bubbles in a Liquid". M. Vacher have then worked on high resolution electron microscopy (HREM) image simulation and analysis in the LTPCM laboratory (CNRS - UMR 5614). He joined the CLIPS laboratory (CNRS - UMR 5524) at the end of 2000 to work on "Habitat Intelligent Santé" project, a sound analysis based medical telesurveillance project. He works on Wavelet Transform applications and carries out research on sound classification and keyword/speech recognition.

Scientific topics: Smart Rooms, Signal Processing, Automatic Sound Tracking & Classification, Real-Time Processing.



Laurent Besacier Laurent Besacier defended his PhD thesis in Computer Science in April 1998 on A parallel model for automatic speaker recognition at the University of Avignon (France). Then he spent one and a half year at IMT (Switzerland) as an associate researcher working on M2VTS European project (Multimodal Person Authentication). Since September 1999 he is an associate professor at the University Joseph Fourier (Grenoble). He carries out research on automatic speaker recognition and segmentation, on multilingual speech recognition and multimedia information retrieval, within the GEOD team at CLIPS Lab. He published about 50 papers on various aspects of these domains. A complete list of publications, as well as an extended resume can be found on : <http://www-clips.imag.fr/geod/User/laurent.besacier/>



Jean-François Serignat In 1974, J.F. Serignat obtained his Doctorate in engineering from the I.N.P.G. in the speech analysis-synthesis field with Linear Predictive Coding (LPC) methods. Then he continued speech analysis research by means of "Autoregressive moving average" (ARMA) models. In 1985, he became the manager of the "Databases and Knowledge-bases" team at the Institut de la Communication Parlée (I.C.P.) at Grenoble. He has been in charge of the management of the French GRECO speech sound database (BDSONS). Then, in 1994-95

he contributed to the recording management of the French part of the EUROM1 Speech Database for an European ESPRIT-SAM project. Since 1997, he is at CLIPS laboratory in GEOD team. In 1999-2000, he managed the recording of BRAF100, a French Database for Automatic Speech Recognition with 100 speakers, in cooperation with "Interactive System Labs" (Karlsruhe University) and "Carnegie Mellon University" (USA) Since January 2000, he is in charge of GEOD team, a group whose research topics are with automatic speech recognition, man-machine dialogue and smart rooms.

Scientific Topics: Speech and sound analysis, speech and knowledge databases