HAL
open science

Information Extraction From Sound for Medical Telemonitoring

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

To cite this version:

Dan Istrate, Eric Castelli, Michel Vacher, Laurent Besacier, Jean-François Serignat. Information Extraction From Sound for Medical Telemonitoring. IEEE Transactions on Information Technology in Biomedicine, 2006, 10(2), pp.264-274. hal-00419915

HAL Id: hal-00419915

https://hal.science/hal-00419915

Submitted on 28 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Information Extraction From Sound for Medical Telemonitoring

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

***Abstract*—Today, the growth of the aging 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) must extract significant sounds from a continuous signal flow. A new detection algorithm based on discrete wavelet transform is proposed in this paper, which leads to accurate results when applied to nonstationary signals (such as impulsive sounds). The algorithm presented in this paper was evaluated in a noisy environment and is favorably compared to the state of the art algorithms in the field. The second stage of the system is sound classification, which uses 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*—Gaussian mixture model (GMM), medical telemonitoring, sound classification, sound detection, wavelet transform.**

I. Introduction

THE ACTUAL growth of the aging population in Europe results in the needs for 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 medical surveillance in a familiar atmosphere for the patient.

Information technology is taking on an important role in 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].

Despite the large number of advantages, there are some problems with telemedicine, such as software complexity, data compression, data transmission, and software reliability. One 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]–[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 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 applied to the last 10 s 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, falling objects, person, sound, 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 multichannel sound acquisition system that analyzes 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

Manuscript received September 10, 2004; revised February 11, 2005, May 23, 2005, July 28, 2005. This work was supported by the French Ministry of Research (ACI—"Technologies pour la Santé").

D. Istrate is with the Ecole Supérieure d'Informatique et Genie des Telécommunication (ESIGETEL), Avon-Fontainebleau, France.

E. Castelli is with the Multimedia Information, Communication and Applications Laboratory, Hanoi, Vietnam.

M. Vacher, L. Besacier and J.-F. Serignat are with the Communication Langagière et Interaction Personne-Système Laboratory, Grenoble, France.

Digital Object Identifier 10.1109/TITB.2005.859889

1089-7771/$20.00 © 2006 IEEE

to the alarm monitor, if it is considered to be a possible alarm. This signal can also be used by a human operator to make the decision of a medical intervention.

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

First, the medical context and the global telemonitoring system are introduced in Section II. The detection stage of the system is described in Section III, and the classification stage is described in Section IV. The method 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, in addition to the performance of the global system. These results are illustrated in Section VII. The system characteristics, strengths, and applications are presented in Section VIII.

II. The Telemonitoring System

A. About Telemonitoring

The living area used in our experiments is a 30 m^2 apartment situated in the TIMC[1] laboratory buildings. The rooms are equipped with medical sensors: blood pressure, oxymeter and a weighing scale, infrared position and sound. The sensor locations are 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 telemonitoring system architecture is made up of two computers which exchange information through an Ethernet network as presented in Fig. 1. The sound extraction and analysis system has a dedicated PC (sound analysis PC) which acquires the signals from all five microphones.

The other PC (data fusion PC in Fig. 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 daily patient surveillance, a sound signal flow is continuously analyzed. Among different everyday life sounds, only some of them are considered alarming sounds: glass breaking, screams, etc. In order to have a reliable sound telemonitoring system, every sound event is detected (a sudden change in the environnemental 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.

[1]Techniques de l'Imagerie, de la Modélisation et de la Cognition (image, modeling, and cognition techniques).

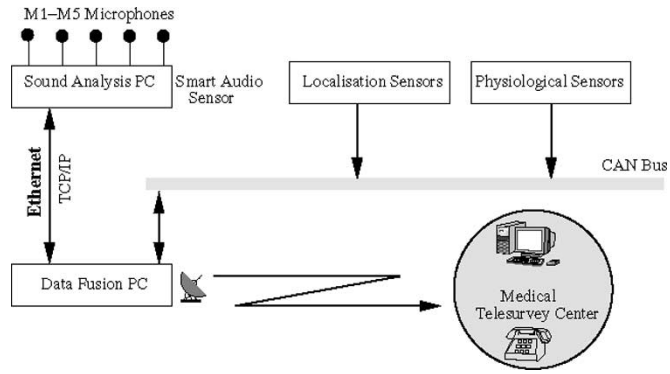

Fig. 1. Acquisition and analysis system.

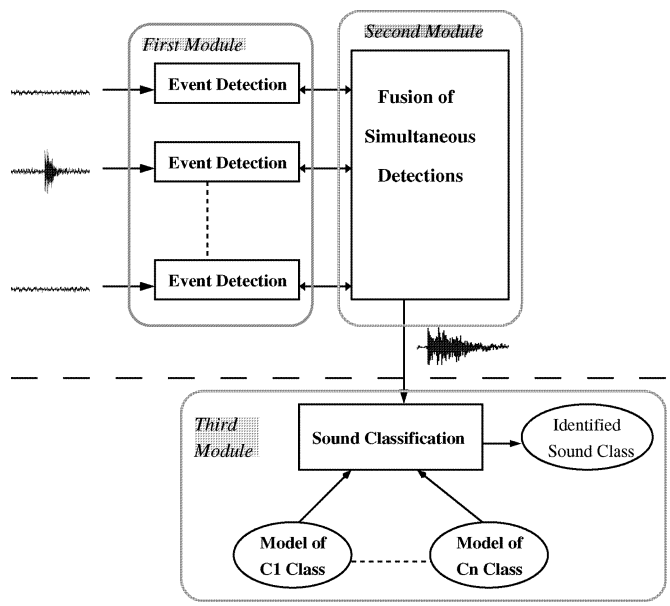

Fig. 2. Sound analysis.

The sound analysis system has been divided in three modules as shown in Fig. 2. The advantage of this division is to make real time implementation possible. Other methods [sound segmentation using a hidden Markov model (HMM), or (BIC) joint with a Gaussian mixture model (GMM)] which involve more complex models would not allow real time processing.

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 24 h/24 h,

the eventual presence of significant nonstationary noise, and the wide variety of sounds that need to be classified.

III. Detection

A. Method

In medical telemonitoring systems, 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 becomes 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 hypotheses of binary detection are

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

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 signal detection, and voice activity detection. There are various possibilities with which to define the measured features; for example, energy, likelihood of a statistical model, and 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], and linear predictive coefficients (LPC) [16], [17]. There has not been much 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 other two 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 that 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 nonuniform time and frequency resolution.

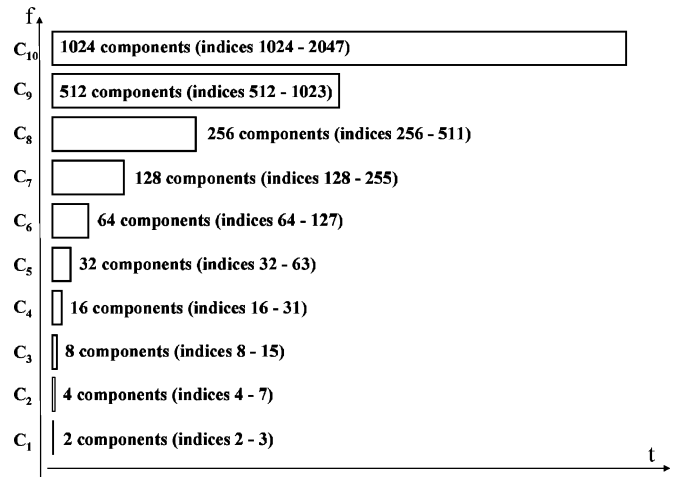

Fig. 3. Layout of the resulting wavelet transform matrix (frame of 2048 samples).

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

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

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 nonuniform frequency and time resolution. The time resolution, compared to frequency one, is greater in high-frequencies and poor in low frequencies which explain the use of DWT for impulsive signal detection. The wavelet base is generated by translation and dilatation of the mother wavelet ψ. In signal processing applications (e.g., noise filtering, and signal compression), the Daubechies wavelets are used as the mother wavelet due to their properties, including good regularity for high number of moments. In the proposed algorithm, we use Daubechies wavelets with six vanishing moments in computing the DWT [21], [22].

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 Fig. 3.

The proposed algorithm (flowchart in Fig. 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 analyzed by a window of 32 ms (four 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

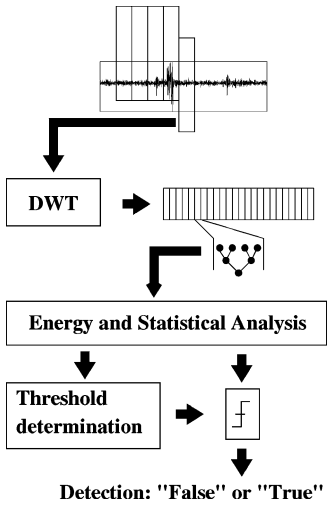

Fig. 4. Flowchart of the wavelet based algorithm.

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 Fig. 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 adaptative threshold is shown with 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 the Gaussian mixture model (GMM) [23], HMMs, dynamic time warping (DTW), neural networks, and others. Since sound classification is a subdomain of 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 only three sound classes, 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 three sound classes. A comparison between neural network, DTW, and vector quantization is presented by Cowling [26]. The results are done for eight sound classes, but the sound database duration is only 35 s.

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 HHMs 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 three stage HMM are similar to GMM classification. The GMM method works in two stages: training and classification.

1) 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 expectation maximization (EM) algorithm [28], which follows a K-means algorithm. The covariance matrices are diagonal.

2) 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)

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

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

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

where 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 mel-frequency cepstral coefficients (MFCC) [29], linear frequency cepstral coefficients (LFCC), LPC, 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 $\Delta\Delta$, respectively) are also used in order to introduce the temporal variation of the signal in GMM modeling [30].

The MFCC parameters are calculated as follows: direct fast Fourier transform (FFT), the computation of the energy of 24 nonuniform triangular filters(Mel-scale), logarithm application on energy values, and inverse discrete cosine transform (DCT). The LFCC parameters are calculated in the same way, but the triangular filters are uniform. The LPCC parameters are the cepstral of LPC coefficients which represent the vocal tract filtering model.

1) 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.

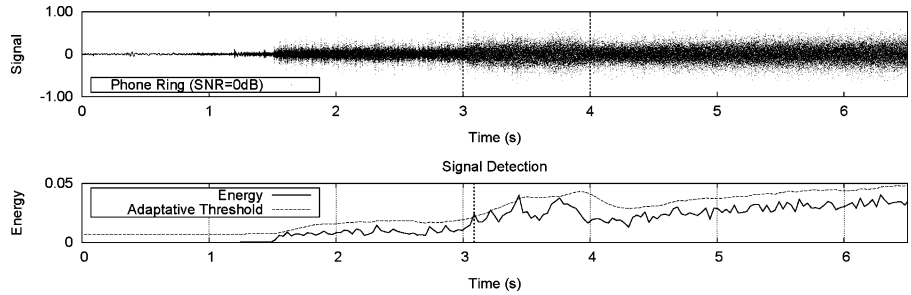

Fig. 5. Detection example of a phone ringing mixed with flowing water noise at $\mathrm{SNR} = 0$ dB.

2) Rolloff Point (RF): This feature is used to measure the frequency which takes 95% of the power spectrum. The rolloff 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 rolloff point is the solution of (5) with $\Theta = 0.95$

$$\sum_{k<\mathrm{RF}} X[k] = \Theta \sum_{k} X[k]. \tag{5}$$

3) 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 rolloff point, (5), where $\Theta = 0.5$.

4) The Acoustical Parameters Proposed: Wavelet-Based Coefficients: The wavelet transform applied in speech recognition has been little studied [32], despite its better time resolution at 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 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.

5) 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 [in (6)] gives an indication of the separation capacity of every acoustical features. In (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 $Var(x)[i]$, and the number of sound classes is k

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

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

V. Coupling Between Detection and Classification

The final module of the system implements the coupling of the detection with the classification. 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: 1) early detection of the sound (a part of the extracted signal contains silence only); 2) 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 7s (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 at both 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), and 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 ten

repetitions per type (the maximum is 300 repetitions) for approximately 3 hours of 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 contains only the noise (file without event). Every sound and noise has been recorded three times. Each file is 25 s long (because of the length of the sound and of the time necessary to initiate the algorithms, which is approximately 5 s). The sound starts at the tenth second of the signal. In the test signal base, we considered three types of noise (white noise, flowing water noise, and environmental noise recorded in the apartment) and eleven 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 four files with the following SNRs: 0, 10, 20, and 40 dB. The SNR is calculated based 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 dB $\leq$SNR $\leq$ 30 dB with an average of 15 dB). We have used the same sounds (played with a loudspeaker) as in the simulation test set.

B. Classification Test Set

The test set used for the sound classification is composed of seven 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), and screams (73 sounds). There are five sets in the seven classes: one contains pure sounds, and the other four contain mixtures of sound and environmental noise (named HIS noise) at 0, 10, 20, and 40 dB SNR.

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 10 and 20 dB with a uniform distribution. Silence between consecutive sounds varies randomly between 5 to 60 s. The total number of useful sounds to be detected is 1577.

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 [%]	EER for different type of 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

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 (7) and (8)

$$R_{\mathrm{MD}} = \frac{\text{No. missed detections}}{\text{No. events to detect}} \tag{7}$$

$$R_{\mathrm{FD}} = \frac{\text{No. false detections}}{\text{No. false detections} + \text{No. events to detect}}. \tag{8}$$

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 0.5 s 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 receiver operating characteristics (ROC) curves, defined as the value of R_{MD} for which $R_{\mathrm{MD}} = R_{\mathrm{FD}}$ (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 SNR, and the other two 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 a 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.

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

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

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

The state of the art algorithm (median conditioning filtering) is not suitable because EER is greater than 10% for a value of SNR ≤ 20 dB. The new algorithm based on wavelet filtering gives the best results for HIS noise: EER $= 0\%$ for SNR $\geq +10$ dB, and EER $= 7.3\%$ for SNR $= 0$ dB. The results are not very good for white noise (EER $= 4\%$ for SNR $= 10$ dB), but are still better in comparison to 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 (Section VII-A. The wavelet based algorithm gives an EER of 0% for this real test set.

C. Sound Classification

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

$$\mathrm{BIC}_{m,K} = -2.L_{m,K} + \nu_{m,K} \ln(n) \tag{9}$$

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.

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 (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 the acoustical vector (MFCC1 is the first MFCC parameter).

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.

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, which are values usually used in speech/speaker recognition. The GMM model is made of four 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 III
ACOUSTICAL PARAMETERS WITH HIGH FDR ($\mathrm{FDR} \geq 2$). THE CHOSEN PARAMETERS FOR TESTS ARE IN BOLD

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**		

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

Parameters	PN	CER [%]
Δ,$\Delta\Delta$(16MFCC+Energy+ZCR+RF+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 is the 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, rolloff point, and centroid seem interesting when combined with conventional parameters used in speech. The best results are obtained with 60 acoustical parameters, and the first and the second derivatives of 16 MFCC, energy, ZCR, RF, and centroid. These parameters are denoted in the Table IV as $\Delta, \Delta\Delta$ (16 MFCC + Energy + ZCR + RF + Centroid).

The three MFCC coefficients have been tested in conjunction with zero crossing rate, rolloff 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 (six instead of 20) produces only a 4.5% increase in the classification error rate (shown in bold in Table IV).

4) Performance in Noisy Environment: Our classification system has been tested in HIS noise situations with two types

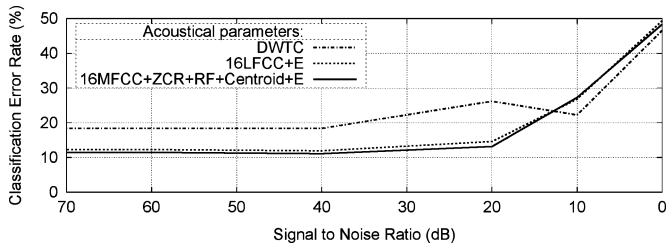

Fig. 6. Classification error in HIS noise (training only on the pure sounds).

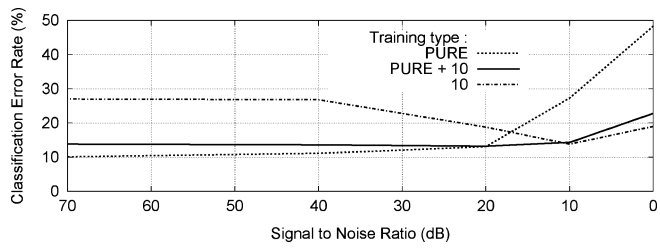

Fig. 7. Classification error in HIS noise (training on the pure sounds and noisy sounds).

of training: training only on pure sounds or on pure sounds and noisy sounds.

Training on pure sounds gives constant results for SNR ≥ 20 dB; the CER decreases beyond this point: for 16 MFCC + ZCR + RF and 16 LFCC parameters, classification error is 26.82% for SNR $= +10$ dB (Fig. 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 SNR ≤ 10 dB and only six 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 of pure sounds and noisy sounds. Better results are obtained from the combination of pure sounds and noisy sounds at 10 dB SNR (Fig. 7). Our tests suggest using a GMM class model for each SNR which would involve SNR estimation before the classification stage.

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 acoustical parameters are 16 MFCC together with ZCR, rolloff 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 three cases:

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

	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 dB$\leq$SNR$\leq$20 dB	67.8 %
Automatic detection and length estimation R_M=1%, R_F=1% 10 dB$\leq$SNR$\leq$20 dB	**27.7 %**

1) the reference case: manual detection and signal delimitation (no algorithm errors);
2) the coupling when the automatic detection algorithm and a fixed time length of signal are used;
3) 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 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 on 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 nonalarm sounds.

The possible cases after detection are good detection event (GD), false detection event (FD), and 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 consequence (we call this MDW) because they belong to the nonalarm sound classes.

The classification stage, similar to 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 MAW). Fig. 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 as the sum of missed detection (MD) and 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 FAs 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 may or may not influence the GFAR after the classification stage. For example, if a false

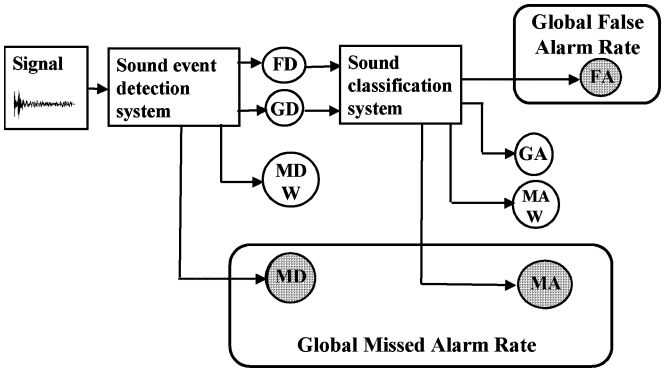

Fig. 8. Global false alarms rate and global missed detection rate.

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)**

detection is classified in a nonalarm sound class, it is not used in the GFAR computation.

The global missed alarm rate is 3% from the global system. This value may be acceptable if the sound extraction system is joined with results from other sensors. The global false alarm rate is 12% from the global system, which in the working conditions investigated represents approximately 12 FAs per day. Some of these FAs could be eliminated with 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. CONCLUSION

A summary of the performance of our detection and classification system is illustrated in Table VI. Although some of the tasks were evaluated on simulated data and still need to be validated in real conditions, we can conclude that we have proposed and tested efficient algorithms both for sound detection and classification.

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

1) 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);
2) the new acoustical parameters resulting from wavelet transform are the best adapted to noise among the tested parameters;
3) a technique to detect the end of the signal has also been proposed;
4) 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. Initially, classical parameters usually used in speech recognition are tested. Subsequently, we proposed and tested nonconventional and new parameters resulting from the DWT. Nonconventional 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 environnemental 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.

ACKNOWLEDGMENT

This research is part of the DESDHIS[2] project, a collaboration between CLIPS[3] laboratory (UMR CNRS-INPG-UJF 5524), responsible for the sound analysis, and TIMC laboratory, charged with the medical sensors analysis and data fusion.

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," *Telemed. J. 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 socio-economic impact of telehealth: A systematic review," *J. Telemed. 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. Inf. 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, pp. 849–856, Jul. 2000.
[6] K. Hung and Y. T. Zhang, "Implementation of a WAP-Based telemedicine system for patient monitoring," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 2, pp. 101–107, Jun. 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. Inf. Technol. Biomed.*, vol. 4, no. 1, pp. 37–4, 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. Inf. 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. Inf. Technol. Biomed.*, vol. 4, no. 2, pp. 173–177, Jun. 2000.
[10] E. Jovanov, A. D. Lords, D. Raskovic, P. G. Cox, R. Adhami, and F. An-drasik, "Stress monitoring using a distributed wireless intelligent sensor system," *IEEE Eng. Med. Biol. Mag.*, pp. 49–55, May 2003.

[2]DEtection de Situations de Détresse en Habitat Intelligence Santé (Distress situations detection in a medical intelligent habitat).

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

[11] P. Varady, Z. Benyo, and B. Benyo, "An open architecture patient monitoring system using standard technologies," *IEEE Trans. Inf. 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 Process.*, 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 Process.*, 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 Process.*, 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 Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.
[16] S. G. Tanyer and H. Ozer, "Voice activity in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–82, Jul. 2000.
[17] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, May 2001.
[18] A. Dufaux, "Detection and recognition of impulsive sounds signals," Ph.D. dissertation, Electronics and Signal Processing Dept., Faculte´des Sciences de I'Univ. de Neuchatel, Neuchatel, Switzerland, 2001.
[19] F. K. Lam and C. K. Leung, "Ultrasonic detection using wideband descrete wavelet transform," in *Proc. IEEE TENCON*, vol. 2, Tokyo, Japan, Aug. 2001, pp. 890–893.
[20] M. Unser and T. Blu, "Wavelet theory demystified," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 470–83, Feb. 2003.
[21] S. Mallat, *Une exploration des signaux en ondelette*, ser. ISBN 2-7302-0733-3. Palaiseau, France: Les Editions de 1'Ecole Polytechnique, 2000.
[22] P. L. Dragotti and M. Vetterli, "Wavelet footprints: theory, algorithms, and applications," *IEEE Trans. Signal Process.*, 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 Process.*, vol. 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 Process.*, vol. 40, no. 7, pp. 1833–1835, Jul. 1992.
[25] G. Papadopoulos, K. Efstathiou, 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, Jun. 1992.
[26] M. Cowling and R. Sitte, "Analysis of speech recognition techniques for use in a non-speech sound recognition system," in *Proc. 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 Commun.*, 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," *J. Acoust. Soc. Amer.*, 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 Process.*, vol. ASSP-28, pp. 357–366, 1980.
[30] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-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 Process. 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 Process. Lett.*, vol. 8, no. 7, pp. 196–198, Jul. 2001.
[33] CLIPS-IMAG Equipe GEOD-Dan Istrate "Base de donnes. Sons de la vie courante," [Online]. Available: www-clips.imag.fr, Grenoble, France, Nov., 2001.
[34] R. W. C. Partnership "CD-Sound scene database in real acoustical environments," [Online]. Available: http://tosa.mri.co.jp/soundb/indexe.htm, Tokyo, Japan.
[35] S. Sciascia, "Sound CD, Bruitages," vol. 3, Paris, France, 1992.
[36] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–64, 1978.

Dan Istrate received the Ph.D. degree in signal processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 2003.

He is an Associate Professor at Ecole Supérieure d'Informatique et Génie des Télécommunications (ESIGETEL), Fontainebleau, France, involved in embedded systems for sound processing. He participated in the Reconnaissance de Situations de Détresses en Habitat Intéligent Santé (RESIDE-HIS) project and Détection de Situations de Détresses en Habitat Intélligent Santé (DESDHIS) within the Communication Langagière et Interaction Personne-Système (CLIPS) laboratory, Grenoble, France. He joined the Laboratoire d'Informatique d'Avignon (LIA) laboratory in 2004 during his postdoctoral research on the speaker diarization task with evolutive HMM for meeting recordings. At LIA, he participated in the BIOMUL (multimodal biometry) project, and the ESTER and NIST RT05 campaign for speaker diarization systems evaluation.

Eric Castelli received the Ph.D. degree in electronic system field and the Habilitation à Diriger des Recherches degree from the Institut National Polytechnique de Grenoble, Grenoble, France, in 1989 and 1999, respectively.

He is an Associate Professor at the Institute National Polytechnique de Grenoble (INPG), Grenoble, France. He was with the Institut de la Communication Parlée (ICP), Grenoble, France, from 1984 to 1997, and then he joined the Communication Langagière et Interaction Personne-Système (CLIPS) laboratory, Grenoble, France, in January, 1998. Now he works in Hanoi, Vietnam, in an international co-operation project to establish 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 scientific program management. He has published about 50 papers on various aspects of speech analysis, speech production, speech, recognition, and general instrumentation.

Michel Vacher received the Ph.D. degree in acoustical science from the INSA of Lyon, Lyon France, 1982.

He has been a Research Scientist at the French "Centre National de la Recherche Scientifique" since 1986. He worked on high-resolution electron microscopy (HREM) image simulation and analysis in the LTPCM laboratory. He joined the CLIPS laboratory at the end of 2000 to work on the "Habitat Intelligent Sante" project, a sound analysis-based medical telesurveillance project. He also works on wavelet transform applications and carries out research on sound classification and keyword/speech recognition.

Laurent Besacier received the Ph.D. degree in computer science from the University of Avignon, Avignon, France.

He spent one and a half year at IMT (Switzerland) as an Associate Researcher working on the M2VTS European project (multimodal person authentication). Since Sptember 1999, he has been an Associate Professor at the University Joseph Fourier, Grenoble, France. He carries out research on automatic speaker recognition and segmentation, multilingual speech recognition, and multimedia information retrieval,

with the Groupe d'Etude sur l'Oral et le Dialogue (GEOD) team at the Communication Langagière et Interaction Personne-Système (CLIPS) laboratory, Grenoble, France. He published some 50 papers on various aspects of these topics.

Jean-François Serignat received the Ph.D. degree in engineering from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1974.

In 1985, he became the Manager of the "Databases and Knowledge-Bases" team at the Institut de la Communication Parlée (I.C.P.), Grenoble, France. He managed the French GRECO speech sound database (BDSONS). In 1994–1995, he contributed to the recording management of the French part of the EU-ROM1 speech database for the European ESPRIT-SAM project. Since 1997, he has been with CLIPS laboratory in the GEOD team. In 1999–2000, he managed the recording of BRAF100, a french database for automatic speech recognition with 100 speakers, in cooperation with the Interactive System Laboratories, Karlsruhe University, and Carnegie Mellon University. Since January 2000, he has been in charge of the GEOD team, a group whose research topics involve automatic speech recognition, man-machine dialogue, and smart rooms.