Life Sounds Extraction and Classification in Noisy Environment

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

Life Sounds Extraction and Classification in Noisy Environment

M. Vacher and D. Istrate and L. Besacier and J.F.Serignat and E. Castelli
CLIPS - IMAG Team GEOD
Grenoble , France
Michel.Vacher@imag.fr, Dan.Istrate@imag.fr

ABSTRACT
This paper deals with the sound event detection in a noisy environment and presents a first classification approach. Detection is the first step of our sound analysis system and is necessary to extract the significant sounds before initiating the classification step. We present three original event detection algorithms. Among these algorithms, one is based on the wavelet and gives the best performances. We evaluate and compare their performance in a noisy environment with the state of the art algorithms in the field. Then, we present a statistical study to obtain the acoustical parameters necessary for the training and, the sound classification results. The detection algorithms and sound classification are applied to medical telemonitoring. We replace video camera by microphones surveying life sounds in order to preserve patient's privacy.

KEY WORDS
Acoustic Signal Processing, Noise, Sound Detection, Sound Classification

1 Introduction

In this paper, we present a system of everyday life sound classification. In order to reduce the calculation time necessary for a multi-channel real time system, our sound extraction process is divided in two steps: detection and classification. The sound event detection is a complex task because the audio signals occur in a noisy environment. In detection step, we compare the performances of the state of the art algorithms and of the three new proposed detection algorithms in the real noisy conditions. The best performances resulted from our proposed algorithm based on the wavelet analysis of sound that allows us to eliminate the noise influence on the detection results. In recognition step using a statistical study applied to acoustical parameters, we can choose the appropriate parameters that give the best classification results with a GMM system.

The possible applications of our sound extraction process are numerous: multimedia documents classification, security sound surveillance, medical telemonitoring etc. The aim of our study is a medical supervising application through the sound. Sound surveillance seems to be better accepted by patients than video camera monitoring.

2 The Detection Algorithms

The detection of a signal (useful sound) is very important because if an event is lost during the first step of the system, it is lost forever. On the other hand, if there are too many false alarms the recognition system is saturated.

Therefore, the performance of the detection algorithm is very important for the entire system. There are many techniques for the sound detection: very simple as functional principle (a threshold on energy), or with a statistical model [1]. We have tested three state of the art detection algorithms [2]: a very simple one, based on the variance of the signal energy and two algorithms based on the median filtering of the energy. We propose three algorithms: one based on the cross-correlation of two successive windows, a second one based on the error of energy prediction and an other one based on wavelet transform. In the next sections we will give a brief presentation of the proposed algorithms and we will pursue with the results of all the tested algorithms.

Cross-Correlation Detection. Knowing that the cross-correlation function is the measure of similarity between two signals, we have used the cross-correlation between two successive signal windows in order to find abrupt changes of the signal (see flowchart in *Figure 1*). The algorithm calculates the cross-correlation between two successive normalized windows of 2048 samples (128ms) and keeps the maximum value of the cross-correlation. Finally, we apply a threshold on this signal (if the signal is under the threshold we generate an event detection). We normalize the signal by the square root of window energy. The overlap between two consecutive series of two analysis windows is 50%.

Energy Prediction based Detection. This algorithm calculates the signal energy on N (2048) samples windows.

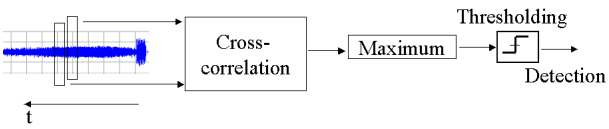

Figure 1. Flowchart of the cross-correlation algorithm

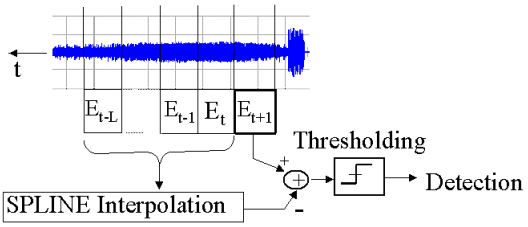

Figure 2. Flowchart of the algorithm for prediction error

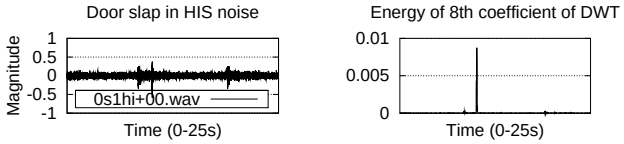

Figure 3. Time evolution of 8^{th} DWT coefficient's energy

First step: the next value of the energy is predicted based on the ten previous values (L=10, prediction length) using the Spline Interpolation method. Next step: a self-adjustable threshold is settled on the prediction error (the absolute difference between the real value and the predicted value). If the energy varies with a small slope, the error is small. However, the error is important for a fast variation of energy (the case of an event to detect) (see flowchart in *Figure 2*). The self-adjustable threshold depends on the standard deviation and the average of the signal.

Wavelet filtering based Detection. Unlike sines and cosines, wavelets are well adapted to signals that have more localized features than time independent wave-like signals: door slap, breaking glasses, step sound, etc... They are more and more used for signal detection [3]. We have chosen Daubechies wavelets with 6 vanishing moments to compute DWT [4]. The wavelet transform on a 512 sample frame allows a good signal enhancement in HIS and white noise. This algorithm computes the energy of the 8, 9 and 10 wavelet coefficients, because most significant wavelet coefficients for sounds to be detected are rather high order, as shown in *Figure 3*: two parasitic noises which are flanking the sound are nearly cleared. Sound appears at 10s.

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

3 Sound Classification

3.1 Selection of the Acoustical Parameters

In order to find the relevant acoustical parameters, we have made a statistical study: the average, the standard deviation, the repartition histograms of parameters by sound classes and the Fisher Discriminant Ratio (FDR).

The FDR (see *Equation 1*) gives an indication about the separation capacity of every acoustical parameters (its value is bigger than 1 for good separation capacity).

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

The results of this study are presented in paragraph 5.4.

After this statistical study, we have tested the classification based on different types of parameters classically used in speech recognition: 16 MFCC ([5]), 16LFCC, 16LPCC, 16 energy coefficients with linear rectangular filters, 16 energy coefficients with linear triangular filters, 16 energy coefficients on a Mel (logarithmic) scale. But we have also tested new parameters generally less used in speech processing like: zero crossing rate (ZCR), roll-off (RF) point (a measure of skewness of the spectral shape) and centroïd (the barycenter of the spectrum). The last three parameters are shortly presented in the next section.

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

Roll-off Point (RF). This feature is used to measure the frequency which delimits 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 2*.

$$\sum_{k<\mathrm{RF}} X[k] = 0.95 \sum_{k} X[k] \quad (2)$$

Centroïd. The centroïd represents the balancing point of the spectral power distribution within a frame. The centroïd for a frame at a specific time is computed as the roll-off point with 0.50 instead of 0.95 in *Equation* 2.

3.2 The Classification Method

We have used a **G**aussian **M**ixture **M**odel (GMM) method in order to classify the sounds [6]. There are other possibilities for the classification: HMM [7], Bayesian method and other [8]. This method evolves in two steps: a training step and a recognition step. We have chosen to use

a model with only 4 Gaussian components, since preliminary experiments have shown no improvement with more components.

The Training Step. The GMM training has been done on the ELISA [9] platform. The training is initiated for each class ω_k of signals of our corpus and gives a model containing the characteristics of each Gaussian distribution ($1 \leq m \leq 4$) of the class: the likelihood $\pi_{k,m}$, the mean vector $\mu_{k,m}$, the covariance matrix and the inverse matrix $\Sigma_{k,m}^{-1}$. These values are achieved after 20 iterations of an "EM" algorithm (Expectation Maximization). The matrices are diagonal.

The Recognition Step. Each extracted signal, X, is a series of n acoustical vectors, x_i, of p components . The parameters π, μ and Σ have been estimated during the training step. The size of acoustical vectors, d, is the number of acoustical parameters used for training. The likelihood of membership of a class ω_k for each acoustical vector is calculated for each class with:

$$\begin{cases} p(x_i \mid \omega_k) = \sum_{m=1}^{4} \pi_{k,m} \cdot \frac{1}{(2\pi)^{d/2} \mid \sum_{k,m} \mid^{\frac{1}{2}}} \cdot \exp(A_{i,k,m}) \\ A_{i,k,m} = \left(-\frac{1}{2}(x_i - \mu_{k,m})^{\mathcal{T}} \cdot \Sigma_{k,m}^{-1} \cdot (x_i - \mu_{k,m}) \right) \end{cases} \quad (3)$$

The likelihood of the entire signal is given by their multiplication :

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

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

4 Sound Database

In order to test and validate the event detection system and the sound recognition system we have recorded a sound corpus. It contains recordings made in the Clips laboratory (15% of the CD), the files of "Sound Scene Database in Real Acoustical Environments" [10] (70% of the CD) and files from a commercial CD (film effects, 15 % of the CD). There are 3354 files and every file is sampled at 16 KHz and 44 KHz.

The sounds picked up in the Clips laboratory were recorded with a Beyer Dynamics microphone and a digital tape (sampling rate 48KHz), then transferred to the PC by sound card. The sound corpus contains: door slap sound (different types of doors), chair sound, step sound, electric shaver sound, hairdryer sound, door lock sound, dishes sound, glass breaking, objects fall sounds, screams, water sound, different ringing, etc. To summarize, the sound corpus contains 20 types of sounds with minimum 10 repetitions per type (the maximum is 300 repetitions).

4.1 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. We consider to be *useful* (impulsive and short) sounds as: door slap, glass breaking, objects fall, etc.; and *environmental* (long and stationary) noises like: water flow, hairdryer, electric shaver, etc. For every sound, there are two signals in the test set : one contains the mixture between the sound and the noise (with event) and the other one only the noise (no 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 initialize the algorithms $\approx$ 5s). The sound starts at second 10 of the signal. In the test signal base, we consider three types of noise (white noise, water flow noise and environmental noise recorded in the habitat) and 11 types of sounds (scream, chair fall, book fall, glass breaking, door slap, step sound, cough, sneeze, door lock, telephone 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 length of the sound.

4.2 Recognition Test Set

The test set used for the sound recognition task is composed of 7 sound classes: door clapping (523 files), ringing phone (517 files), step sound (13 files), dishes sound (163 files), door lock (200 files), breaking glasses (88 files), screams (73 files). There are 5 set of the 7 classes: one with the pure sounds and other four, mixing sound and HIS noise at 0, 10, 20 and 40 dB signal to noise ratio.

5 Experimental Results

5.1 Evaluation of Detection Algorithm Performance

To find the best algorithm for our application, we have calculated the Missed Detection Rate (R_M) and the False Detection Rate (R_F) on the test set, with the *formulas (5)* and *(6)*.

$$R_M = \frac{\text{No. missed detections}}{\text{No. events to detect}} \quad (5)$$

$$R_F = \frac{\text{No. false detections}}{\text{No. false detections} + \text{No. events to detect}} \quad (6)$$

A detection is considered to be false if an event is detected while actually there is no event. We consider a detection to be missed when the system detects nothing in the interval: 0.5s before the event and, the end of the signal event (*Figure 4*). A detection occurring in this interval is considered to be a good detection of the event.

When we have created the test set, for every file we have generated a file type ".sam" (the SAM standard [11]) containing temporary labels: the start and stop labels of the

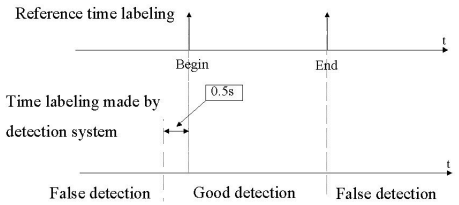

Figure 4. Definition of a missed and false detection

useful noise. Every algorithm has been applied to the all files from the test set and has generated a label file. For every algorithm we have varied the detection threshold in order to obtain a variation of R_M and R_F from 0 to 1 (for the Self-Adjustable Thresholds algorithms we have varied the κ coefficient). To compare the algorithms we have determined the equal error rate (EER), defined as value of R_M for $R_M = R_F$.

5.2 Detection Results with Our Test Set

The EER of the different algorithms applied on our test set are given in *Figure 5* for white noise *(a)* and HIS noise *(b)* and several signal to noise ratio 0, $+10$, $+20$ and $+40dB$ (note that HIS noise is the environmental noise recorded in our experimental habitat in Grenoble). The numerous calculation necessary for the entire sound system (5 channels to analyze simultaneously at 16 KHz sample rate) and the necessity of a real time processing (medical conditions) forced us to make a trade-off between the performance and the complexity of the algorithm. Besides, medical care imposes a very small missed detection rate.

After the results analysis, we can conclude that most algorithms are very efficient in case of white noise (EER=0% for SNR$\geq$+10dB). Only the *energy prediction error* and especially *cross-correlation* (EER$>$70% $\Rightarrow$ the curve is outside of graphic area) give bad results: sound event detection is difficult because white noise is not correlated. But in real conditions, our preliminary tests have shown that white noise is not realistic for our application. Therefore to analyze the results, we must compare their corresponding performances only for HIS environmental noise for a 10dB SNR (our real environmental conditions).

The *median filtering*, *variance* and *self-adjusting threshold* are not suited because EER$>$10% for a 20 dB SNR. We can state that the the *energy prediction* algorithm is fast and gives good results (EER=7% at +20dB) except for HIS noise at low SNR. The *cross-correlation* algorithm is better for the HIS noise but requires long calculation time and can not be used with white noise.

In conclusion, the *wavelet filtering* algorithm gives the best results for HIS noise: EER=0% for SNR$\geq$+10dB and EER=7.6% for SNR=0dB. The results are roughly less for white noise (EER=4% for SNR=10dB), but they are enough to allow good performances for similar noises like water flow.

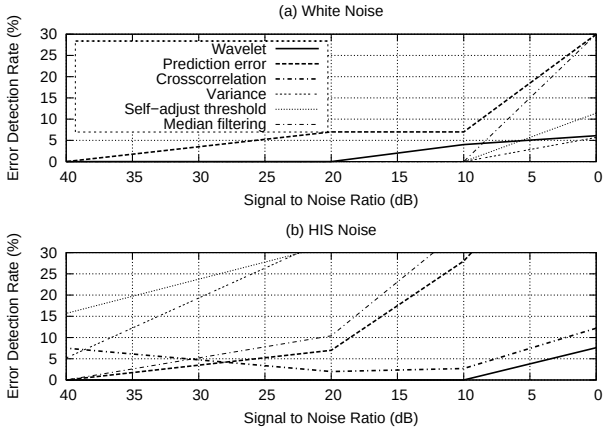

Figure 5. Detection results in different noisy environment

5.3 Detection Results in Real Conditions

We have recorded 60 files inside our test-apartment (real conditions) at different SNR (minimum 2dB, maximum 30 dB, average 15 dB). We have used the same sounds (played with a speaker) as in the test base. The results obtained with the well adapted algorithms (*wavelet filtering*, *cross-correlation* and *prediction error*) are presented in *Table 1* and confirm previous results (*Figure 5*).

Algorithm	Wavelet	Cross-corr.	Predict.Err
EER[%]	**0**	**4.4**	**10**

Table 1. Results of sound event detection in real conditions

5.4 Preliminary Results with Sound Classification

Statistical study for the choice of parameters. The statistical study results gives the relevant acoustical parameters (especially the Fisher Discriminant Ratio - *formula 1*) and reduces the number of tests. *Table 2* shows FDR values for some parameters.

As resulting from *Table 2* : the second, third and fourth MFCC coefficients are the only relevant MFCC parameters for our classes. Otherwise ZCR, RF and Centroïd are relevant but not the energy.

Sound Classification results. The analysis window was set to 20 ms with an overlap of 10 ms. The GMM model is made of 4 Gaussian distributions. The training/test protocol is a "leave one out" protocol: the model of each class is trained on all the signals of the class, excepting one. Next, each model is tested on the remaining sounds of all classes. The whole process is iterated for all files (1577 tests).

The experimental results are in *Table 3*. The average of error classification rate (ECR: number of recognition er-

Parameter	FDR	Parameter	FDR
MFCC1	2.72	MFCC11	2.88
MFCC2	**16.07**	MFCC12	3.20
MFCC3	**10.33**	MFCC13	1.48
MFCC4	**10.02**	MFCC14	3.61
MFCC5	2.01	MFCC15	3.26
MFCC6	2.91	MFCC16	4.41
MFCC7	3.36	**ZCR**	**17.99**
MFCC8	3.60	**RF**	**16.70**
MFCC9	0.53	**Centroïd**	**23.75**
MFCC10	3.34	Energy	2.54

Table 2. FDR for some acoustical parameters

Parameters	PN	ECR [%]
Δ,$\Delta\Delta$(16MFCC+Energy +ZCR+RF+Centroïd)	60	8.71
16 MFCC + Energy+ ZCR+RF+Centroïd	20	**11.47**
16LFCC+Energy	17	12.26
16LPCC+Energy	17	14.74
16MFCC+Energy	17	15.21
3MFCC+ZCR+RF +Centroïd	6	**16.11**
16 Coef.Mel	16	23.50

Table 3. Results of sound classification methods

ror divided by the number of tests) and the correspondent number of parameters (PN) are given. For each parameter, we calculate the average of the error value of all the classes. This first sound classification results are encouraging. We can observe that the best results are obtained with the MFCC parameters (speech specific parameters) but new parameters like zero crossing rate, roll-off point, centroïd seem interesting when combined with conventional parameters used in speech.

We have tested the combination of three MFCC coefficients with the zero crossing rate, roll-off point and centroïd, suggested by the statistical study (see *Table 2*). We have noticed that the parameters considered to be irrelevant after the statistical study can be eliminated with practically no negative influence on the performances of the system; drastically reducing the number of parameters (6 instead of 20 parameters) produces only 4.5% increase of the error classification rate (in bold in *Table 3*).

Performances in noisy environment. We have tested our classification system in HIS noise. The results are roughly constant for SNR$\geq$20dB, but they decay beyond: for 16 MFCC + ZCR + RF and 16 LFCC parameters, error classification is 26.82% for SNR=+10dB (see *Figure 6*). Real conditions are between 10 and 20dB of SNR and these first results are not sufficient. We are actually working to improve performances by signal enhancement.

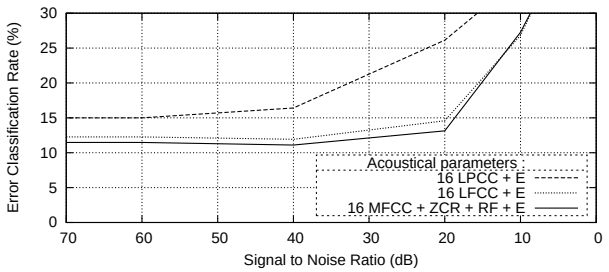

Figure 6. Classification error in HIS noise

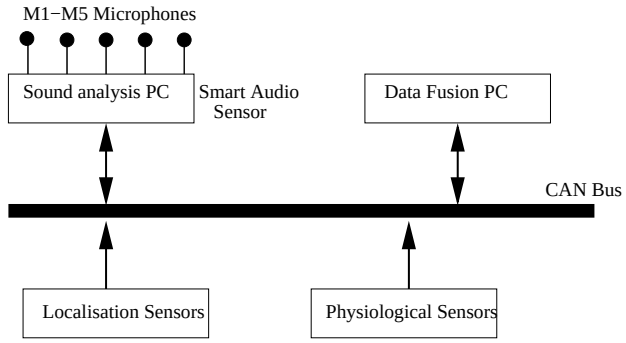

Figure 7. Acquisition and analysis system

6 Application to Medical Supervision

We have developed the detection and classification system in order to obtain a smart audio sensor for surveillance of the elderly, convalescent persons or pregnant women [12]. Its main goal is to detect serious accidents as falls or faintness (which can be characterized by a long idle period of the signals) everywhere in the apartment. Thus, the originality of our approach consists in replacing the video camera by a system of multichannel sound acquisition. The system analyzes in real time the sound environment of the apartment and detects abnormal sounds (objects or patient's falls) and calls for help (or groans), that could indicate a distress situation in the habitat [13].

To respect again privacy, no continuous recording or storage of the sound is made, since only last 20s of the audio signal are kept in a buffer and sent to the alarm monitor if a sound event is detected. That can be used by the human operator to make the decision of a medical intervention.

Telemonitoring. The habitat we used for experiments is a 30 m^2 apartment situated in the TIMC laboratory buildings, filled with various sensors, especially microphones. The entire telemonitoring system is comprised of two computers which exchange information through a CAN bus (see *Figure 7*).

The master computer is in charge of data fusion and analyzes both data coming from fixed and moving sensors and information coming from the slave computer, which is

continuously surveying the microphones. The final sound analysis system (implemented inside the slave computer) should be the following: each time a sound event is detected, a message is sent to the master computer, notifying occurrence time of detection, type of event (speech or other sound), localization of the emitting source ; it also should indicate either the most probable sound classes (with the corresponding confidence index), or the most probable words (calls for help), with their confidence index. From this the master computer could send an alarm if necessary.

Sound Analysis System. A microphone is located in every room (toilet, kitchen, shower-room, hall and living-room). Each of the 5 microphones is connected to the slave computer. The sound or speech source can be localized by comparing the sound levels of the microphones.

The sound analysis system is a two-step analysis system : the first one for the detection of a sound event and the second one for sound classification. In the complete implementation of the system the second step includes sound classification and words recognition. In the first step, signals from the 5 channels are used to detect events. It is a difficult task because of the environmental noise.

If a sound event is detected, extracted signal is transmitted to the second step and sound classification is initiated. At the moment, the recognition system is only in test and the detected events are classified by a human operator.

7 Conclusions

We have proposed three new algorithms for signal detection. They have been tested with three state of the art algorithms, bests results being achieved with the wavelet filtering algorithm. This allows us to detect a sound event in the habitat. We are currently testing a GMM system for the sound class recognition. Firstly, we have used classical parameters of speech recognition; secondly we have tested new parameters. The first results are encouraging and non-conventional parameters like ZCR, RF and Centroïd seems to be very discriminant for the sound classification task. We are working to improve sound classification, because insufficient performances have been encountered in noisy environment. This system is developed for medical supervision application in the framework of RESIDE-HIS project.

8 Acknowledgments

This system is a part of the RESIDE-HIS (REconnaissance de SItuations de DEtresse en Habitat Intelligent Santé) project, a collaboration between the CLIPS and TIMC laboratories. This project is financed by IMAG.

References

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

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

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

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

[5] S.B.Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, pp. 357–366, 1980.

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

[7] R.S. Goldhor, "Recognition of environmental sounds," in *ICASSP '93, USA*, 1993, pp. 149–152.

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

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

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

[11] "http://www.icp.grenet.fr/relator/standsam.html," .

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

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