SOUND DETECTION AND CLASSIFICATION THROUGH TRANSIENT MODELS USING WAVELET COEFFICIENT TREES

Michel Vacher, Dan Istrate and Jean-François Serignat

CLIPS - IMAG (UMR CNRS-INPG-UJF 5524, Team GEOD
385, rue de la Bibliothèque - BP 53, 38041 Grenoble cedex 9, France (Europe)
phone: +33 4 7663 5795, fax: +33 4 7663 5552, email: Michel.Vacher@imag.fr
web: www-clips.imag.fr

ABSTRACT

Medical Telesurvey needs human operator assistance by smart information systems. Usual sound classification may be applied to medical monitoring by use of microphones in patient's habitation. Detection is the first step of our sound analysis system and is necessary to extract the significant sounds before initiating the classification step. This paper proposes a detection method using transient models, based upon dyadic trees of wavelet coefficients to insure short detection delay. The classification stage uses a Gaussian Mixture Model classifier with classical acoustical parameters like MFCC. Detection and classification stages are evaluated in experimental recorded noise condition which is non-stationary and more aggressive than simulated white noise and fits with our application. Wavelet filtering methods are proposed to enhance performances in low signal to noise ratios.

1. INTRODUCTION

In this paper a sound detection/classification method is presented. This method has been developed as part of a medical telesurvey system intended for home hospitalization. The aim of this system is to detect a distress situation of the patient using sound analysis. In case of distress a medical center is automatically called with the aim in view to give assistance to the patient. The decision of call is taken by a data fusion system from smart sensors and particularly a sound system as explained in [1].

Each sound produced in the apartment is characteristic of:

- a patient's activity: the patient is locking the door, or he is walking in the bedroom,
- the patient's physiology: he his having a cough,
- a possible distress situation for the patient: a scream or a glass breaking are suddenly appearing.

If the system has a good ability of classification for such sounds, it will be feasible to know if the patient is needing help. Several usual sound classes needed for this application have been defined and a corpus has been recorded in our laboratory.

Before sound classification, it is necessary in a first step to establish the start and the stop time of the sound to classify in the environmental noise. The precision of this 2 times must be sufficient to allow the classification step good performances. In the context of audio signal encoding the input signal can be decomposed into "tonal", "transient" and "stochastic" components as described by Daudet in [2][5]; our problem is restricted to transient detection for which large wavelet coefficients are more easily interpreted as transients.

Proposed methods are based on trees of wavelet coefficients, during transition time upper wavelet coefficients being affected: a significant coefficient is likely coming with additional significant coefficients at the same time location and lower scale level [3]. In this paper, two methods based on wavelet tree detection are presented, the obtained results are compared. We also present the results of sound classification method in noisy conditions.

2. SOUND EXTRACTION IN NOISY ENVIRONMENT

2.1 Noise and sounds

As no everyday life sound database was available in the scientific area, we have recorded a sound corpus. This corpus contains recordings made in the CLIPS laboratory, files of "Sound Scene Database in Real Acoustical Environment" (RCWP Japan) and files from a commercial CD: door slap, chair, step, electric shaver, hairdryer, door lock, dishes, glass breaking, object fall, screams, water, ringing, etc. The corpus contains 20 types of sounds with 10 to 300 repetitions per type. The test signal database has a duration of 3 hours and consists of 2376 files.

The sound classes of our corpus are described in the following table; the number of frames for each class is given too. Each frame has a duration of 16ms (256 samples at 16 kHz). Signal duration varies in a 500:1 ratio. Fast variations of the signal are related to short duration parts of the signal (some milliseconds).

Sound Class	Number of Frames (Entire corpus)	Duration (Each Sound)
Door Slap	47 398	375 ms
Breaking Glasses	9 338	15 ms-7.5 s
Ringing Phone	59 188	35 ms-10 s
Step Sound	3 648	1.4-5 s
Scream	17 509	0.37-5.8 s
Dishes Sounds	7943	125 ms-1.35 s
Door Lock	605	24 ms-117 ms

Table 1: Sound classes

This work is a part of the DESDHIS-ACI "Technologies for Health" project of the French Research Ministry. This project is a collaboration between the Clips ("*Communication Langagière et Interaction Personne-Système*") laboratory, in charge of the sound analysis, and the TIMC ("*Techniques de l'Imagerie, de la Modélisation et de la Cognition*") laboratory, charged with the medical sensors analysis and data fusion.

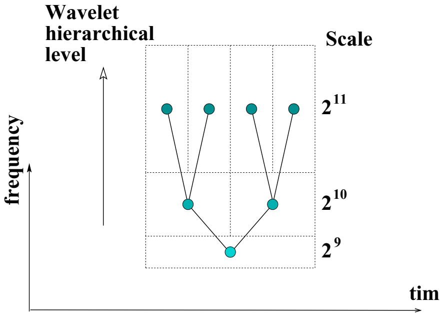

Figure 1: Tree of wavelet coefficients for N=2048 sample window (tree depth of 3 levels)

Two types of noise have been considered, the noise registered inside an experimental apartment[1], which is named HIS noise, and stationary white noise. HIS noise is a result of all noises in the building, he is a transient noise similar to usual sounds to detect, but transients are partially reduced by propagation inside the structure of the building. This kind of noise is not a stationary noise. First investigations showed that, unlike Dufaux studies [4], white noise performances are not sufficient to insure satisfactory performances in our actual case.

For this reason white noise study will only be used for literature result comparison, like in [4]. Evaluation of the algorithms has been made at 4 signal to noise ratios: 0, +10, +20 and +40dB.

2.2 Transients modeling

Methods based on wavelet transforms are often used for singularity characterization and transient detection, because of the compact support of wavelets in conjunction of the dyadic properties of these transforms. These two properties are allowing the analysis of reduced parts of the processing window. The figure 1 shows a wavelet tree with 3 level depth beginning at the highest hierarchical level. Each node is corresponding to a wavelet whose support is drawn in frequency and time domain. For wavelets of highest level the support in time is twice the sampling period.

For our purpose it is not necessary to determine the full tree corresponding to the transient, we limit our study to these 3 levels and we characterize each tree by his energy e, the sum of the energy of all nodes. We have chosen Daubechies wavelets ψ with 6 vanishing moments to compute DWT on 2048 sample windows (128 ms), the wavelet base is generated by translation and dilatation of the mother wavelet ψ [8]:

$$\left\{\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}}\,\psi\left(\frac{t-2^j n}{2^j}\right)\right\}_{(j,n)\in\mathbb{R}} \qquad (1)$$

As we consider the energy e of the tree, the non significant nodes are implicitly not taken into account because they are negligible in the summation. With this approach the tree is not pruned and we don't eliminate nodes at scale 2^{11} if their mother node at scale 2^{10} is not significant, but this might not be very harmful because of the low depth of the tree.

A signal of chair falling with HIS noise is drawn on the bottom sub-figure of figure 2, the sound appears at time $t = 10s$. The top sub-figure displays tree energy evolution across the time. Energy corresponding to useful signal is surrounded by isolated noise pulses which are sometimes greater but useful signal is associated with numerous adjacent trees and in this way could be detected.

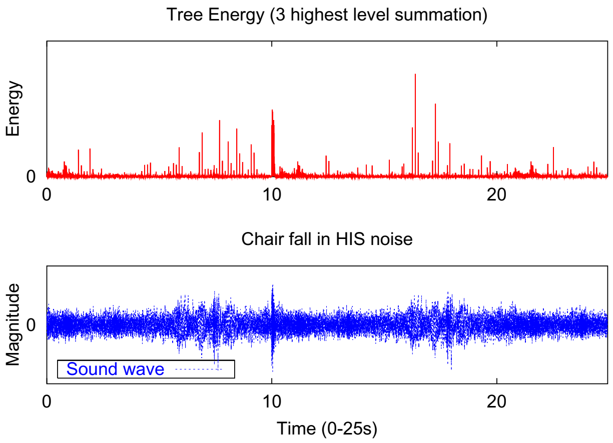

Figure 2: Sound signal and Tree Energy

2.3 Proposed detection algorithms

2.3.1 *Several tree mean*

DWT is calculated on $N = 2048$ sample windows (128ms) as shown in figure 3. From this DWT the energy e of each tree is obtained by time translation ($500\mu s$) across the transform. The means e_{means} of the 64 last values is calculated at each translation step in order to suppress noise influence. Since 16 kHz sampling rate, corresponding frame width is 32 ms. A transient is characterized by a large increase of e_{means}.

The detection th threshold is adaptive: $th = \kappa + 1.2 \cdot \mu_{e_{means}}$, with $\mu_{e_{means}}$ referring to the mean of the last values of e_{means} and κ to an adjusting parameter. The coefficient 1.2 was introduced because of remaining oscillations on e_{means}.

2.3.2 *Threshold on the standard deviation*

This algorithm (see flowchart in figure 3) is computing the DWT of consecutive $N = 2048$ sample windows. From this DWT the energy e of each tree is obtained as above by time translation across the transform. A median filter is applied to eliminate isolated trees which are only relevant of noise, the

Detection Method	SNR	HIS noise	White noise
Several tree mean	0dB	6.7%	5.9%
	⩾+10dB	0%	0%
Standard deviation	0dB	3%	22.7%
	⩾+10dB	0%	0%
Filtered energy	*0dB*	*71.3%*	*19.2%*
(conditioning	*+10dB*	*45.2%*	*6.1%*
median filter)	*+20dB*	*7.5%*	*6.1%*
	+40dB	*6.1%*	*6.1%*

Table 2: Detection EER, 198 tests at each SNR level (99 noised sounds, 99 pure noise)

[1] The HIS apartment is located in the TIMC laboratory building

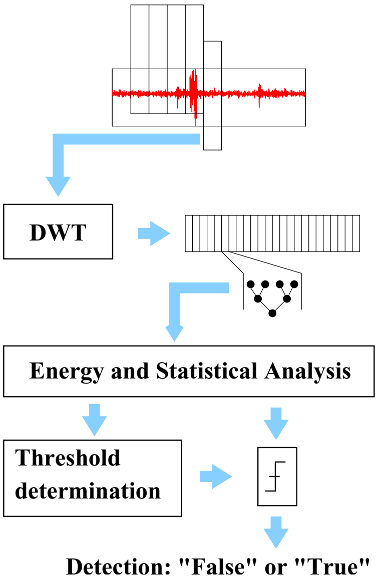

Figure 3: Detection algorithms using energy tree evaluation

Method	0dB	+10dB	+20dB	+40dB
Sev. tree mean	23.6ms	13.9ms	9ms	5.5ms
Standard dev.	30.6ms	13.4ms	11.5ms	8.4ms

Table 3: Mean of detection delay for sound duration shorter than 2s for HIS noise (78 tests at each SNR level)

width of the filter is 3. The standard deviation σ of the last 640 filtered energy values is calculated at each translation step: a high increase of the standard deviation is significant of a transient.

The detection is achieved by increase beyond an adaptive threshold $th = \kappa + \mu_{sigma}$, with μ_{sigma} referring to the mean of the last values of σ and κ to an adjusting parameter.

2.4 Detection results

Evaluation of each algorithm was done from COR curves giving *missed detection rate* (MDR) as function of *false detection rate* (FDR), the Equal Error Rate (EER) being achieved when MDR=FDR. Results for the two algorithms and for the conditioning median filtered energy described in [4] are given in table 2. Best results for HIS noise at 0dB SNR are obtained for *"Standard deviation"* (3%) and *"Several tree mean"* (6.7%), in the case of white noise *"Several tree mean"* (5.9%) is the best.

In order to insure best classification results, a short detection delay is very important. Delay means for the 2 proposed methods are given in table 3 at each SNR in the previous conditions (threshold choice in order to obtain Equal Error conditions) for sounds of short duration for which it is important to extract the most useful part of the signal. Best values at 0dB SNR are obtained for *"Several tree mean"*: 23.6ms; if SNR$\geqslant$+10dB they are below 14ms for the 2 methods. An additional part of signal may be added without critical incidence by deciding that signal is beginning 20 ms before detection time: it is needed neither to cut signal nor to transmit additional noise frames to the classification stage.

3. SOUND CLASSIFICATION

We have used a **G**aussian **M**ixture **M**odel (GMM) method in order to classify the sounds [9]. There are other possibility for the classification: HMM, Bayesian method, etc. GMM has been chosen because with other methods similar results has been obtained, although they are more complex.

3.1 Acoustical parameters

The first step of sound classification is acoustical parameters extraction. Acoustical parameters are a synthetic representation of time signal. Acoustical parameters classically used in speech/speaker recognition are: MFCC(Mel Frequencies Cepstral Coefficients), LFCC (Linear Frequencies Cepstral Coefficients), LPC(Linear Predictive Coefficients). Acoustical parameters used in speech/music/noise segmentation are : ZCR (zero crossing rate), RF (roll-off point), centroïd. **Zero Crossing Rate (ZCR)** is the number of crossings on time-domain through zero-voltage within an analysis frame. **Roll-off Point (RF)** is the frequency which is above 95% of the power spectrum. **Centroïd** represents the balancing point of the spectral power distribution within a frame.

3.2 GMM

The classification with a GMM method suppose that the acoustical parameters repartition for a sound class may be modeled with a sum of Gaussians. This method evolves in two steps: a training step and a classification step. In the training step for each sound class the Gaussian model is estimated. The training step start with a K-Means algorithms followed by EM algorithm(Expectation-Maximization) in 20 steps. In the classification step for each acoustical vector is calculated a likelihood for each sound class. The global likelihood for each class is the geometrical average of all acoustical vector likelihood. The signal belongs to the sound class for which likelihood is maximum.

3.2.1 Model Selection

The BIC (Bayesian Information Criterion) criterion is used in this paper in order to determinate the optimal number of Gaussians [10]. BIC criterion select the model trough the maximization of integrated likelihood: $BIC_{m,K} = -2.L_{m,K} + \nu_{m,K} \ln(n)$. Where $L_{m,K}$ is logarithmic maximum of likelihood, equal to $\log f(x|m,K,\widehat{\theta})$ (f is integrated likelihood), m is the model and K the component number of model, $\nu_{m,K}$ is the number of free parameters of model m and n is the number of frames. The minimum value of BIC indicate the best model.

The BIC criterion has been calculated for the sound class with the smallest number of files, for 2, 4, 5 and 8 Gaussians. The results of the table 4 are obtained for 16 MFCC parameters. Looking at the results, a number of Gaussians between 3 and 5 seem to correspond to the best sound modeling. We have decided to use 4 Gaussians.

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

Table 4: BIC for 2, 3, 4, 5 et 8 Gaussians (1577 tests)

	SNR [dB]				
Filtering	0	10	20	40	$\geqslant 55$
Without	48.3	27.2	13.1	11.1	10.1
With F1	40	20.5	14.6	10.4	10
With F2	40.4	20.9	15.1	10.7	10

Table 5: ECR for 16MFCC+ZCR+RF+Centroïd in the HIS noise presence (1577 tests for each SNR)

3.3 Noise attenuation

In order to increase the classification efficiency, wavelet filtering is applied before sound classification. The Wavelet Transform is more adapted to analyze and process impulsive signals than Fourier Transform which is adapted to periodical signals.

Two methods are tested on our test set. The general steps of the method are : DWT calculation on 256 samples window (9 wavelet coefficients), the application of thresholds on the DWT Coefficients, DWT inverse calculation.

Thresholds are applied to the absolute value of each Wavelet Transform coefficients. For the first method (F1) values under the threshold are cleared and other values are unmodified. For the second method (F2) values under the threshold are cleared; for other values a subtraction of estimated noise value is made ($B^i_{max}/10$). Threshold values for each DWT Coefficient are:

$$\begin{cases} T_i = 1.2 * B^i_{max} & \text{for} \quad i = 1 \dots 4 \\ T_i = 0.9 * B^i_{max} & \text{for} \quad i = 5 \\ T_i = 0 & \text{for} \quad i = 6 \dots 9 \end{cases}$$

where T_i is the threshold applied to coefficient i of DWT and B^i_{max} the maximal value of coefficient i of DWT for the noise. The value B^i_{max} is estimated on the first 100ms of signal which are considered to contain only environmental noise.

This filtering threshold choice results from a study of the HIS noise and sounds. The sounds contain less useful information in the first five DWT coefficients, whereas in the case of HIS noise almost all information is located in low hierarchical level coefficients of DWT.

3.4 Classification results in noisy conditions

The sound classification is validated on the test set with 7 classes (the pure sounds and the sounds mixed with HIS noise at 0 dB, 10 dB, 20 dB and 40 dB of SNR). The sound classification performances are evaluated through the error classification rate (ECR) which represent the ratio between the bad classified sounds and the total number of sounds to be classified.

In the table 5 the classification results for 16 MFCC acoustical parameters coupled with zero crossing rate, Roll-off point and centroïd are presented. We can observe that for ”pure” sounds we have 10% of classification error. In the noise conditions, the wavelet filtering give a gain, in absolute, of 8% for the ECR. The two methods of wavelet filtering has approximately same results.

4. CONCLUSION

We have presented detection and classification methods allowing us to detect and classify a sound event recorded in nursed home. Proposed detection method are resulting in low delay after signal beginning -typically 14 ms- so that link to classification step is not disturbed.

Detection is error-less for 10dB SNR and upper and error classification rate of 20% or better are reached in the same noise conditions; according to these two results we can conclude that this detection/classification system may be used under realistic conditions with moderate noise.

We are working to apply proposed detection techniques to speech recognition in order to allow call for help by the patient in our medical application.

These identification methods may have possible applications in multimedia classification or security sound surveillance.

REFERENCES

[1] G. Virone, D. Istrate, M. Vacher and all, “First Steps in Data Fusion between a Multichannel Audio Acquisition and an Information System for Home Healthcare,” in *Proc. IEEE Engineering in Medicine and Biology Society*, Cancun, Mexico, Sept. 2003, pp. 1364–1367.

[2] L. Daudet, *Représentations structurelles de signaux audiophoniques - Méthodes hybrides pour des applications à la compression*. PhD Thesis, Marseille, 2000.

[3] L. Daudet, S. Mollat, and D. B. Torrésani, “Transient detection and encoding using wavelet coefficient trees,” in *Proc. GRETSI 2001*, Toulouse, France, F. Flandrin Ed., Sept. 2001.

[4] A. Dufaux, L. Besacier, M. Ansorge and F. Pellantini, “Automatic Sound Detection and Recognition for Noisy Environment,” in *EUSIPCO 2000*, Tampere, Finland, Sept. 2000.

[5] L. Daudet and B. Torrésani, “Hybrid representations for audiophonic signal encoding,” *Journal of Signal Processing, Special issue on Image and Video Coding Beyond Standards*, vol. 82(11), pp. 1595-1617, Nov. 2002.

[6] 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*, Jan. 2002.

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

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

[9] D. Reynolds, *Speaker Identification and Verification using Gaussian Mixture Speaker Models*, Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland,pp 27-30, 1994.

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