# SOUND CLASSIFICATION IN A SMART ROOM ENVIRONMENT: AN APPROACH USING GMM AND HMM METHODS

Michel VACHER, Jean-François SERIGNAT, Stéphane CHAILLOL

Laboratoire d'Informatique de Grenoble, Team GETALP, UMR CNRS-INPG-UJF 5524,
385, rue de la Bibliothèque, BP 53, 38041 Grenoble cedex 9, France, Michel.Vacher@imag.fr

Corresponding author: Michel VACHER

Because of cost or convenience reasons, patients or elderly people would be hospitalized at home and smart information systems would be needed in order to assist human operators. In this case, position and physiologic sensors give already numerous informations, but there are few studies for sound use in patient's habitation. However, sound classification and speech recognition may greatly increase the versatility of such a system: this will be provided by detecting short sentences or words that could characterize a distress situation for the patient. Analysis and classification of sounds emitted in patient's habitation may be useful for patient's activity monitoring. GMMs and HMMs are well suited for sound classification. Until now, GMMs are frequently used for sound classification in smart rooms because of their low computational costs, but HMMs should allow a finer analysis: indeed the use of 3 states HMMs should allow better performances by taking into account the variation of the signal according to time. For this framework a new sound corpus was recorded in experimental conditions. This corpus includes 8 sound classes useful for our application. The choice of needed acoustical features and the two approaches are presented. Then an evaluation is made with the initial corpus and with additional experimental noise. The obtained results are compared. At the end of this framework a segmentation module is presented. This module has the ability of extracting isolated sounds in a record by the means of a wavelet filtering method which allows the extraction in noisy conditions.

*Key words:* Gaussian mixture model, hidden Markov model, background noise, sound classification in smart rooms, Wavelet transform.

## 1. INTRODUCTION

It is well known that ageing is emerging as an important concern for European countries. In this context the central challenge of health and long-term care policies is to provide full access to high-quality services for all, while ensuring the financial sustainability of these services. Progress in aids and assistive technologies might be a cost-effective way to support the supply of informal care and care provisions. In this way speech analysis and sound classification can give interesting information by taking into account distress calls from the patient and fall sounds. Therefore sound classification and speech analysis can give information on the patient and may help the decision-making by the medical monitoring system [1].

The medical monitoring system, described in [1], uses two kinds of information. Information issued from the medical sensors, the actimeter and door contacts are analysed in order to detect a difference in the behaviour and state of the patient. Information given by the sound analysis system [2] may be analyzed in the same manner but can also be used to detect critical or distress situation. It will be the case when a sentence like "Help me", "Doctor quickly" is recognized or when a scream, an object fall or a glass breaking is classified.

The implementation of the sound analysis system must meet two different aims: the real-time ability of the system, a good precision for speech recognition and sound classification. The real-time ability is achieved if sounds and speech are detected on the flow and not missed. Concerning speech recognition and sound classification, results may be known some seconds after the sound event but it is very important that neither false nor missed alarm occurs. In this paper we will only discuss sound classification; Gaussian Mixture Model (GMM) [3] and Hidden Markov Model (HMM) [4], [5], [6] based methods are often used in

this area. The GMM method is easy to implement while the HMM method takes into account the shape of the signal. For this framework, the ALIZE library was used as well for the GMM method and for the HMM method.

### 1.1. Short Overview of the Sound Analysis System

The aim of our global project is to obtain useful sound information and to transmit it through network to a medical supervising application in a medical centre. The habitat we used for experiments is a 30m$^2$ apartment situated at the TIMC laboratory inside the Faculty of Medicine of Grenoble. It is equipped with various sensors, especially microphones in every room (hall, toilet, shower-room, living-room) [1]. The entire tele-monitoring system is composed of three computers, which exchange information through a local network (see figure 1).

This system is designed for the surveillance of the elderly, convalescent persons or pregnant women. Its main goal is to detect serious accidents or falls or faintness at any place of the apartment. Each time a sound event is analysed, a message is sent to the Data Fusion PC, notifying occurrence time of detection, most probable sound class or recognized sentence, localization of the emitting source. From this and from other data obtained from localisation and physical sensors, the Data Fusion PC can send an alarm if necessary.

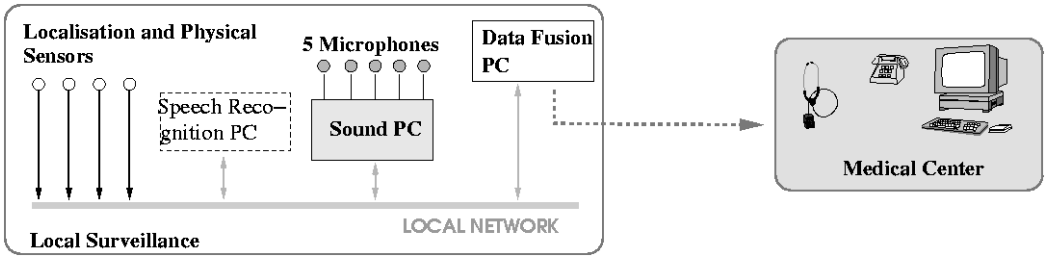


**Figure 1.** Medical tele-monitoring system.

The sound analysis system has been divided in four modules as shown in figure 2. The first module is the detection module in charge of extraction of audio events from the signal flow. Extracted signals are then transmitted to the segmentation stage, which switches them to the classification module in case of life sounds or to the RAPHAEL recognition module [2] in case of speech. At the end, the obtained information will be sent to the data fusion system, which will respond to the question: “Is the patient in a normal or a distress situation?”

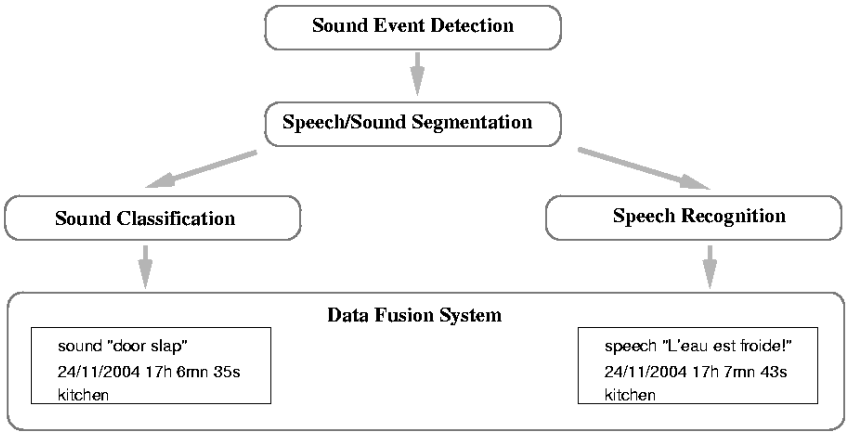


**Figure 2.** Sound analysis system.

## 2. GMM AND HMM METHODS

GMM and HMM methods are well suited for sound classification [8], [9]. HMM-based methods are used in the acoustical stage of speech recognition systems, prosodic analysis [10] and isolated word recognition [11]. The implementation of these two methods in this framework uses the ALIZE library [3]. It was necessary to add some specific extensions for our application.

### 2.1. Gaussian Mixture Models - GMMs

The classification with a GMM-based method supposes that the acoustical parameter repartition for a sound class may be modelled with a sum of Gaussian distributions. This method evolves in two steps: a training step and a classification step, as shown in figure 3. The acoustic pre-processing stage in charge of feature extraction will be described in subsection 4.1.

During the training step and for each sound class the characteristics of each Gaussian model are estimated, the number of these models $N$ will be discussed later in subsection 4.2. Parameters of the Gaussian distribution $m\ (1 \le m < N)$ are for the sound class $k\ (1 \le k \le 8)$: the likelihood $\pi_{k,m}$, the mean vector $\mu_{k,m}$ and the inverse covariance matrix $\Sigma_{k,m}^{-1}$. During the initialisation step, these parameters are initialized by the mean of the arbitrary partition of the training corpus in $N$ equal sized parts. This step is followed by a second step including 12 iterations of the EM algorithm (Expectation Maximisation) on 20% of the corpus (randomly drawn). The last step is made of 12 iterations of the EM algorithm on the full corpus.

In the classification step the likelihood of each frame of the signal is calculated for each sound class (see equation 1). A frame is a vector of $d$ acoustical features. The global likelihood $p(X|\omega_k)$ for one class is the geometrical average of the likelihood of the $n$ frames as expressed in equation 2, and the signal belongs to the class for which the likelihood is maximal.

$$p(x_i|\omega_k) = \sum_{m=1}^{N} \pi_{k,m} . \frac{1}{(2\pi)^{d/2} \det(\sum_{k,m})^{1/2}} . \exp(A_{i,k,m}) \tag{1}$$

with $A_{i,k,m} = \left( -\frac{1}{2}(x_i - \mu_{k,m})^{\mathrm{T}} . \Sigma_{k,m}^{-1} . (x_i - \mu_{k,m}) \right)$

$$p(X|\omega_k) = \prod_{i=1}^{n} p(x_i|\omega_k) \tag{2}$$

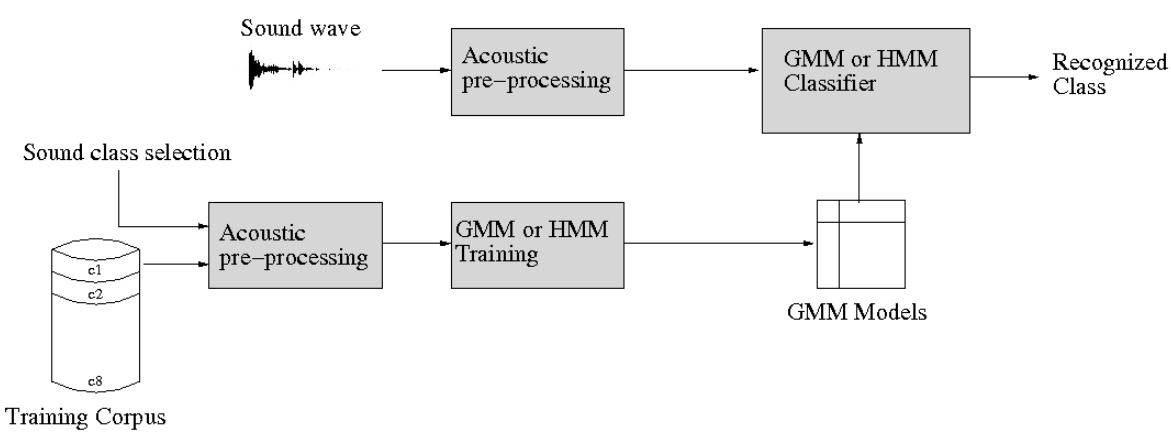


**Figure 3.** Block diagram of the GMM and HMM methods: training and classification.

### 2.2. HMM Classification of Sounds

In the context of audio signal encoding, the input signal can be decomposed into “transient”, “tonal” and “residual” components as described by Daudet in [12]. We choose then to use 5 states HMM as shown in figure 4, the states q0 and q4 corresponding to the silent part at the beginning and at the end of the signal. There is no transition possibility from q0 to q4 because they represent the same state. The states q1, q2 and q3 are related to the three components of the signal. A transition is possible from each state qi to a state qj if j is greater than or equal to i, except for the transition from q0 to q4. For any sound, some of the 3 states may be empty. $P_{ij}$ denotes the transitional probabilities from the state qi to the state qj.

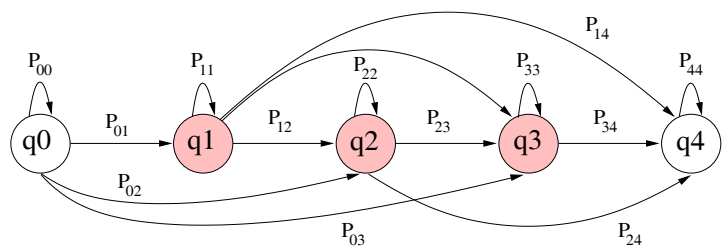


**Figure 4.** HMM state transitions.

An example is given in Figure 5 in the case of a scream; it is the result of the training stage described in section 2.3. The first state q0 has a very short duration and is not visible because of the scale of the figure. The state q1 is short and made of the establishing part of the signal. The state q2 is corresponding to the highest energy part and q3 to the decreasing energy part. During q3 resonant frequencies are decreasing too.

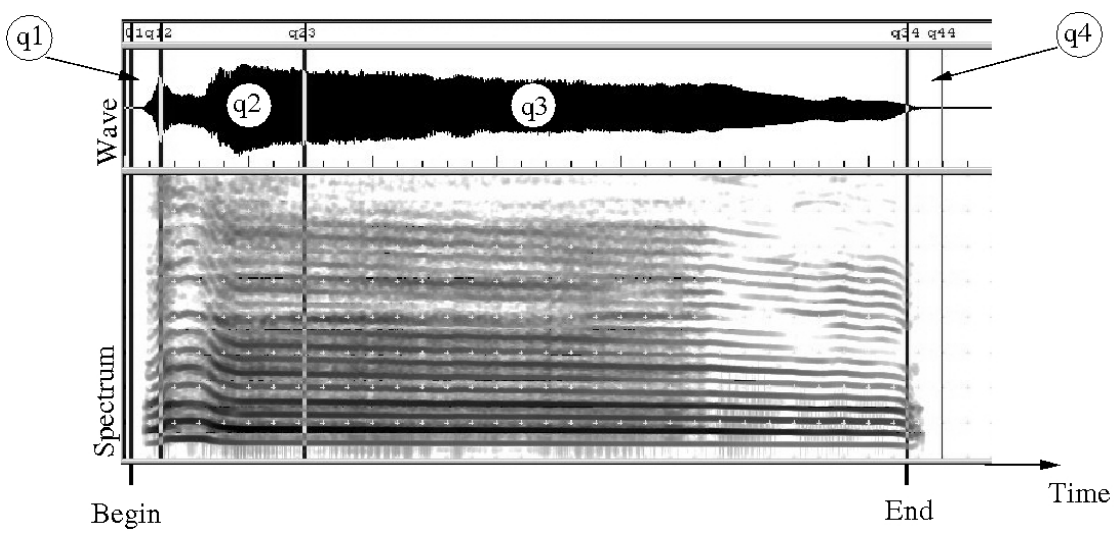


**Figure 5.** HMM states of a scream.

The classification step uses a Viterbi algorithm [13] to estimate the class $k$ of a sound $X$ between $m$ models $M_j$, a sound being represented by a sequence of vectors of n components $X_1^n$. The first member of equation 3 can be simplified because of the equal probability of each sound class.

$$k = \arg\max_j p(X|M_j)\,P(M_j) = \arg\max_j p(X|M_j),\ 0 \le j < m \tag{3}$$

The probabilities are estimated by using a Viterbi method. The Viterbi algorithm is a forward probability method of best path estimation. For simplification all the sums are replaced by a maximum function, and then the estimated probability for the best partial path $q_i^p$ from initial state $q_1$ to the state $q_i$, after emission of the $p$ first vectors $X_1^p$ of $X$ must be expressed by Equation 4,

$$\overline{p}(q_i^p, X_1^p | M_j) \quad = \quad \max_k \overline{p}(q_k^{p-1} | M_j)\ p(q_k^{p-1}, X_1^{p-1}, M_j)\ ,\ 0 \le p < n \tag{4}$$

where $\overline{p}(q_i, X_1^p | M_j)$ denotes the probability of the partial path from the state $q_1$ to the state $q_i$ of the model $M_j$, $X_1^{p-1}$ the first $(p-1)$ vectors of $X$, $x^p$ the $p^{\text{th}}$ vector of $X$, $X_1^p$ the sequence of the $p$ first vectors of $X$, $q^{p-1}$ the state $q$ when the $(p-1)^{th}$ components is reached. Considering the equal probability of each sound class, the equation can be written as in equation 5.

$$\overline{p}(q_i^n, X_1^p | M_j) \quad = \quad \max_k \left[\overline{p}(q_k^{p-1}, X_1^{p-1} | M_j) P(q_i | q_k, M_j)\right] p(x^p | q_i) \tag{5}$$

The probabilities of each vector are evaluated through Gaussian models, each state $q_k^p$ or $q_i$ being modelled by a GMM in conjunction with the probability of transition. The probability $P(X | M_j)$ is then estimated for each model $M_j$ using equation 6.

$$\overline{p}(X | M_j) \quad = \quad \overline{p}(q_F^n\ ,\ X_1^n | M_j) \qquad \text{Where } q_F \text{ denotes the final state.} \tag{6}$$

The signal belongs to the sound class $q$ associated to the model $M_q$ for which the probability has the highest value.

### 2.3. HMM Training and Automatic Labelling

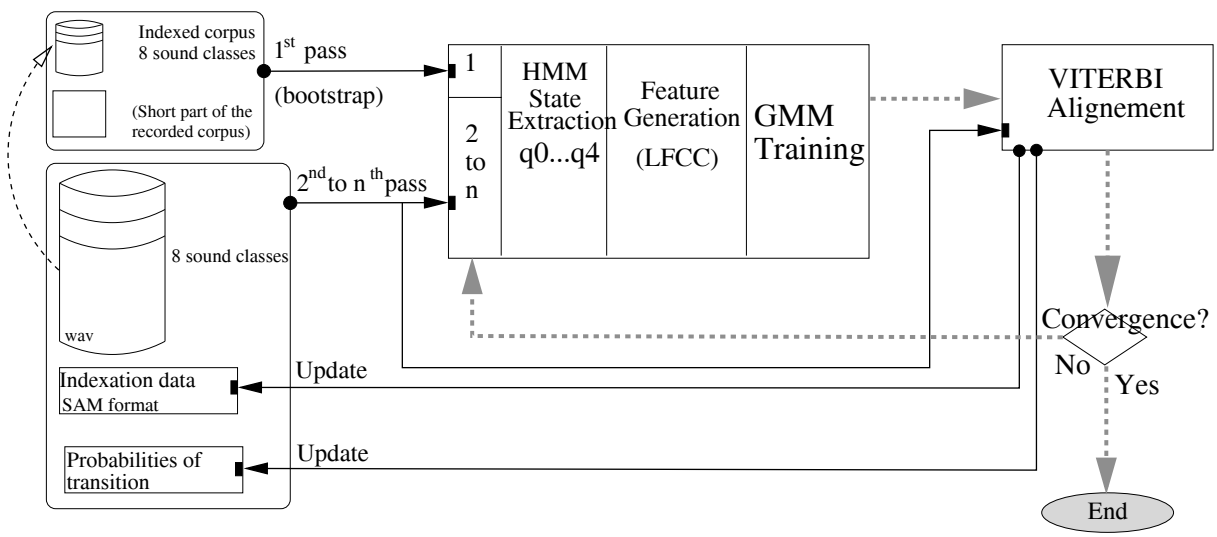


**Figure 6.** HMM Training and Automatic Labelling System.

The global HMM training and automatic labelling system is presented in figure 6. The most used method for HMM training is the Baum-Welch algorithm [13], as it is not available in the ALIZE library we have developed a HMM training system using the Viterbi algorithm. This system, described in figure 6, requires an indexed starting corpus. Thus a short part, around 10%, of the corpus was manually indexed by examining the spectrogram of each sound; indexing marks refer to HMM state transitions. Training is performed for each sound class separately.

At the beginning of the first stage, segmentation data are used in order to extract the 5 states q0…q4 from the initialisation corpus. Then, the acoustic pre-processing step is initiated and produces for all the sound classes the corresponding acoustical features. It is then possible to obtain a GMM model for the 5 states of all the classes. The same method as in section 2.1 is used. The Viterbi algorithm is then applied on the entire corpus using these GMM models. The output of this algorithm is the best path across each sound wave and then, for each frame of the sound wave, the corresponding HMM state. Indexation data for the full

corpus and probabilities of transition are then extracted from these outputs. From the second pass, the same process is started again but GMM models may then be evaluated from the full corpus.

After *n* iterations of the process the convergence is reached if the indexation data remains quite constant. This requires between 20 and 50 steps. At the end of the final training step, the indexation values for the initial indexed corpus have changed because of the optimisation process.

## 3. SOUND DATABASE

Each sound produced in an apartment is characteristic for a normal patient's activity (door slap, dishes, etc.), a possible distress situation (object fall, scream, etc.) or a patient's physiology (cough, etc.). Sounds related to the patient's physiology are not yet taken into account because of the difficulty in recording such sounds.

Therefore a new corpus, adapted to this framework, was recorded; this corpus is made of 8 everyday sound classes related to two categories:

- **Normal** sounds related to a usual activity,
- **Critical** sounds related to the possibility of a distress situation for the patient and thus giving very important information to be sent to the remote monitoring system.

A small share of the corpus consists of sounds extracted from a preceding corpus recorded at the time of former studies [7]. New sounds have been recorded in the CLIPS laboratory using omni-directional wireless microphones (SENNHEISER eW500). Some sounds were obtained from the Web [18]. Some characteristics of this corpus are given in table 1. Each sound is recorded in one file; the sampling rate is 16 kHz. Because of the use of an HMM classifier, each file begins and ends with a silence part of minimal duration 32 ms. The average SNR of the corpus is +27 dB.

With this corpus we have generated a noised corpus with 4 levels of signal to noise ratio (SNR = +8 dB, +17 dB, +22 dB, +26 dB). The noise was recorded in an apartment. The original corpus and the noised corpus have been used for the classification tests.

**Table 1.** Everyday sound corpus

| Class of Sound | Former Corpus | New Records | Internet | Number of Files | Average Duration of one Sound (ms) |
|---|---|---|---|---|---|
| Dishes Sounds | 45% | 50% | 5% | 363 | 606 |
| Door Lock | - | 100% | - | 507 | 390 |
| Door Slap | 40% | 60% | - | 372 | 1022 |
| Glass Breaking | 40% | 50% | 10% | 118 | 269 |
| Object Falls | - | 100% | - | 128 | 1039 |
| Ringing Phone | 15% | 70% | 15% | 319 | 991 |
| Screams | 80% | - | 20% | 102 | 432 |
| Step Sounds | 10% | 60% | 30% | 76 | 86 |
| Entire Corpus | 29% | 61% | 10% | 1985 | 276 |

Some examples of sounds are shown in figure 7. The spectra are very different but in each case high frequency components must be taken into account. In case of the door slap sound, there are two parts: in the first part sound is like a decreasing white noise, in the second part some resonant frequencies are important. The synthetic ringing bell is constituted of discrete and regularly spaced frequencies. The scream is very similar to voice signal with a high number of harmonics. Resonant frequencies are important during all the dishes sound, impact between a cup and a saucer.

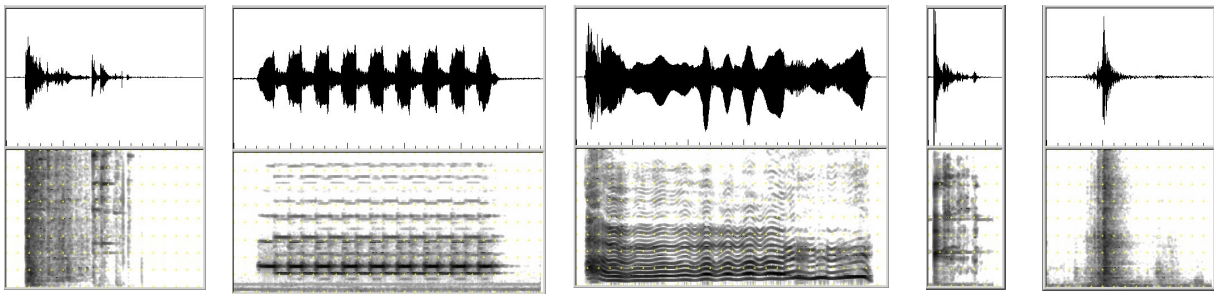


**Figure 7.** Some examples of sounds with the corresponding sonograms

## 4. FEATURES AND MODEL SELECTION

GMM and HMM classification methods are not performed directly on the signal, but use extracted acoustical parameters which are synthetic representations of the time signal. Analysis window width for each frame was set to 16 ms with an overlap of 8 ms. For a sampling rate of 16 kHz the analysis window is composed of $2^8$ samples, an integer power of two being required by Fast Fourier Transform analysis. This width is a compromise between the time precision of state transitions and with frequency analysis constraints. A great number of the signals being as short as 86 ms, it might be impossible to use a wider analysis frame.

### 4.1. Features

Acoustical parameters classically used in speech/speaker recognition are: **MFCC** (Mel Frequency Cepstral Coefficients), **LFCC** (Linear Frequency Cepstral Coefficient) and **LPC** (Linear Predictive Coefficients). MFCC are frequently used in speech recognition because of their characteristics that are very similar to human hearing mainly thanks to the logarithmic Mel frequency scale.

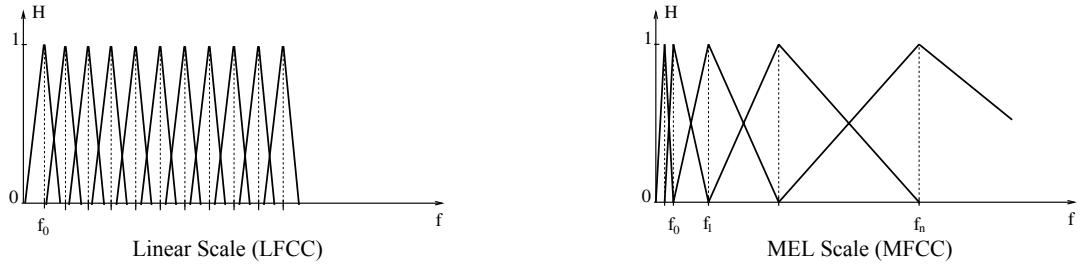


**Figure 8.** LFCC and MEL Triangular Filter Response

As discussed in section 3, the bandwidth of sound signals is very large and includes frequently high frequency components. The computing steps for the LFCC and MFCC parameters are: pre-emphasis and windowing; FFT of the analysis frame signal; triangular filtering; logarithmic calculus of the filtered coefficients and inverse cosine transform. The inverse cosine transform is obtained according to equation 6. As shown in figure 8, the bandwidth is constant over the spectrum for LFCC but larger in high frequencies for MFCC because of MEL logarithmic scale. In our study, it is important to use components allowing an equal sensitivity over the full bandwidth as allowed by LFCC. All the 24 coefficients are considered in order to take into account the full bandwidth.

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n\left(m - \frac{1}{2}\right)}{M}\right), \; 0 \le n < M \tag{6}$$

Normalized energy is not used as additional parameter, this parameter being too dependent of experimental recording conditions. Derivatives of first ("delta") and second order ("delta-delta") of LFCC parameters are preferred. The total amount of used parameters is then 72.

### 4.2. Number of Models in the Case of the GMM-based Method

The Bayesian Information Criterion (BIC) is used in this paper in order to determine the optimal number of Gaussian models [14]. This criterion is well suited for Gaussian mixture as proved by Roeder and Wassermann [15]. BIC criterion selects the model trough the maximization of integrated likelihood (1).

$$BIC_{m,K} = -2\,L_{m,K} + \nu_{m,K} \ln(n) \tag{7}$$

where $L_{m,K}$ is logarithmic maximum of likelihood, equal to $\log f(x|m,K,\tilde{\theta})$ ($f$ is the integrated likelihood), $m$ is the model and $K$ the component number of the model, $\nu_{m,K}$ is the number of free parameters of the $m$ model and $n$ is the number of frames. The minimum value of *BIC* indicates the best model.

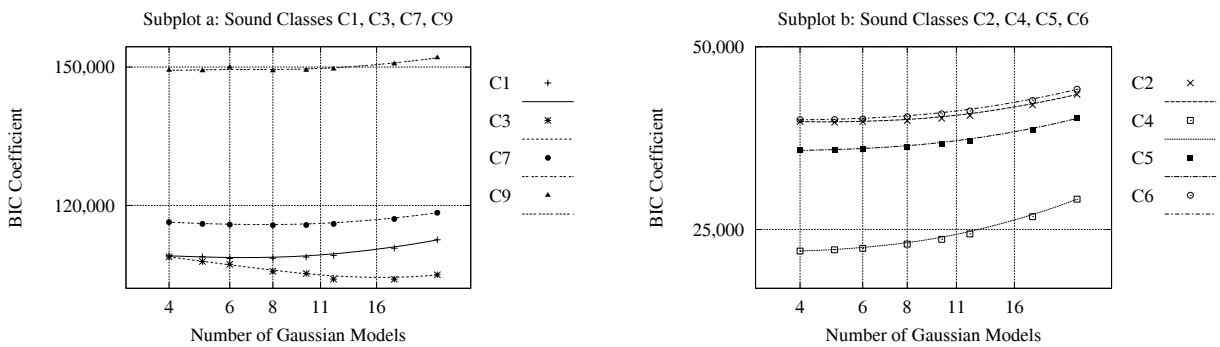


**Figure 9.** BIC Coefficient Evolution for the 8 sound classes, GMM Evaluation with 24 LFCC features in conjunction with derivatives of first and second order.

**Table 2.** Correspondence between number and class of sounds

| *Number* | *Class of Sounds* | *Duration* | *Number* | *Class of Sounds* | *Duration* |
|---|---|---|---|---|---|
| C1 | Door Slap | 6 min 20 s | C5 | Screams | 3 min 30 s |
| C2 | Glass Breaking | 2 min 53 s | C6 | Object Falls | 2 min 13 s |
| C3 | Ringing Phone | 5 min 17 s | C7 | Dishes Sounds | 3 min 40 s |
| C4 | Step Sounds | 44 s | C9 | Door Lock | 11 min 1 s |

The BIC criterion has been used first for the sound class and for the speech class in noiseless conditions, for 4, 5… and 24 Gaussian models in case of 24 LFCC parameters in conjunction with derivatives of first and second order. The results of the figure 9 are given for a number of Gaussian models between 4 and 24 in case of C1, C3, C7, C9 sound classes (subplot a) and C2, C4, C5, C6 sound classes (subplot b). According to the BIC criterion, performances will be optimal when the log-likelihood of the observations, given the GMMs, is minimal. As it appears on these 8 curves, the optimal Gaussian number is different for each sound class. In subplot a, the criterion is minimal between 12 and 20 models for the bell

ringing class (C3), the records of this class are very heterogeneous (old style bells, synthesised bells…). Curves are very flat for C1, C7 and C9 but increase below 16 models. In subplot b, curves are slowly increasing except for step sounds (C4), so a number of 12 Gaussian models seem to be acceptable. The C4 class could be neglected because these sounds are very low level and not often detected in the real environment. We have decided to use 12 Gaussian models, which may be a good compromise between classification performances and calculus consumption (real time constraints).

### 4.3. Number of Models in the Case of the HMM-based Method

Since the likelihood of one frame of signal is evaluated using GMMs, the number of models may be chosen in the same conditions with the BIC criterion. In figure 10 and for the sound classes C5 (screams) and C7 (dishes) the BIC coefficient is represented as function of the Gaussian number for the three states q1, q2 and q3.

The curve shapes are not very different as the preceding related to GMM configuration. We then chose the same number of Gaussian for HMM evaluation.

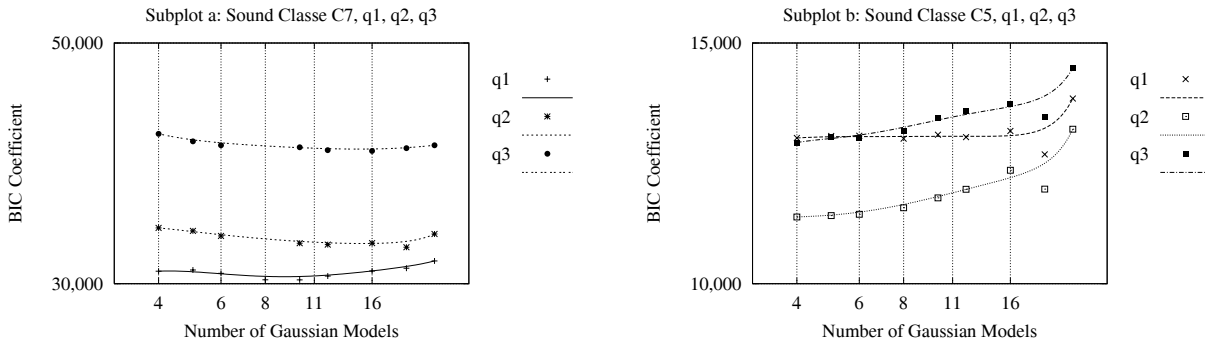


**Figure 10.** BIC Coefficient Evolution for C5 and C7 sound classes, states q1, q2 and q3, 24 LFCC features.

## 5. CLASSIFICATION EVALUATION

Training is made with original sounds but testing is made with original sounds and sounds mixed with experimental noise at 4 different SNR levels. The tests use a “cross validation protocol”, training is achieved with 90% of the original sound corpus and each of the 10% remaining files is evaluated at these 4 SNR levels and for the original sounds.

The sound classification performances are evaluated through the Classification Error Rate (CER), which represents the ratio between badly classified sounds and the total number of sounds to be classified. The number of Gaussian is fixed to 12 for the GMM-based method and for the HMM-based method. The sampling rate is 16 kHz, the bandwidth is then 8 kHz. The analysis window width is 16 ms with an overlap of 8 ms. Results are shown in table 3.

We can observe that in all cases (except at SNR = +17 dB) results are the best for the HMM-based method against the GMM-based method, even if the differential / acceleration coefficients are used for the GMM-based method and not for the HMM-based method. We can conclude that HMMs allow a finer analysis by taking into account the temporal shape of the signal. Then the use of 3 states HMMs allows better performances.

Best results are reached by the use of 24 LFCC parameters with derivatives of first and second order and the HMM-based method. The Classification Error Rate is 1.7% for the original corpus and below 6% at SNR = +22 dB. These values are good according to former results [2] and other results in the literature [16], [17].

**Table 3.** Classification Error Rate (%)

| *SNR* [dB] | *24 LFCC only* | | *24 LFCC and* $\Delta - \Delta\Delta$ | |
|---|---|---|---|---|
| | *GMM* | *HMM* | *GMM* | *HMM* |
| Original | 6.3 | 3.1 | 4.4 | 1.7 |
| 26 | 9.3 | 5.9 | 7.3 | 4.2 |
| 22 | 13 | 6.6 | 10.4 | 5.7 |
| 17 | 21.3 | 16.3 | 15.1 | 9.7 |
| 8 | 36.6 | 29.8 | 43.8 | 28.3 |

## 6. HMM SOUND SEGMENTATION

The proposed global sound recognition system is composed of two modules: the first is the segmentation system and the second is the HMM classification system yet presented. A segmentation system must be able to detect the beginning and the end of each isolated sound in a flow and that may be achieved by the way of a HMM segmentation model with only one class of sounds. This class includes all the sounds of the previous classes. HMM state transitions are shown in figure 11. A possible transition has been added from q4 to q0 which are actually the same state (the "silent" state).

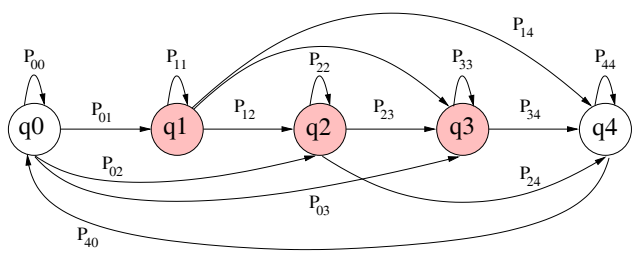


**Figure 11.** HMM State Transitions for Segmentation.

Training is operated in the same conditions as in subsection 2.3 but with a unique model over the complete corpus after a bootstrap step using a short labelled part of the corpus. The number of Gaussian mixtures is 4; a greater value is not needed because we must only make the distinction between silence and one of the 3 sound states. We recorded in real conditions a dedicated corpus in our laboratory using the same omni-directional wireless microphones (SENNHEISER eW500). This test corpus is made of 10 sound wave files; so each wave file contains a sequence of about 10 sounds of all the sound classes except step sounds. The total duration of the corpus is 9 minutes. It contains 129 sounds and the average SNR is +28 dB.

**Table 4.** Classification Results (Number of files)

| | C1 | C2 | C3 | C5 | C6 | C7 | C9 |
|---|---|---|---|---|---|---|---|
| Classified as C1 | 8 | | | | | | |
| Classified as C2 | | 7 | | | | | |
| Classified as C3 | | | 14 | | | | |
| Classified as C5 | | | | 25 | | | |
| Classified as C6 | 1 | | | | 13 | | |
| Classified as C7 | | 4 | | | | 41 | |
| Classified as C9 | | | | | | | 15 |

All the 129 sounds are correctly detected; there is no missed or false sound detection. The results after the classification step are given in table 4.

We can notice that one door slap sound is classified as object fall and that 4 glass breaking sounds are classified as dishes sounds. This is due to the great similarity between these sounds, because they are produced in a similar manner. An impact between a cup and a saucer is not very different of an impact between a cup and the floor even if the cup is broken at the end. So the recognition error rate is 3.9%.

Segmentation results are very poor in noise conditions. So we studied an algorithm of noise reduction in the parts of the signal corresponding to the silent part. In order to avoid the degradation of the sound signal with filtering artefacts, we try not to modify this part; indeed the noise is not stationary and not known at any moment in time. The proposed method supposes that at least one frame of signal containing only noise may be isolated at any moment before the sound signal; these frames of noise will be used as reference.

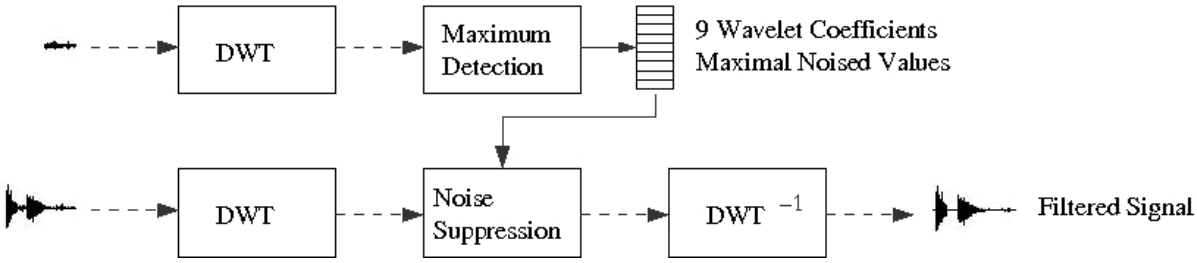


**Figure 12.** Wavelet Noise Reduction.

The Discrete Wavelet Transform (DWT) of each frame of noise is calculated on 256 sample windows; wavelets are distributed over 9 wavelet coefficients. For each coefficient level, the maximal absolute value is memorised. The processing of the signal to be segmented is described in figure 12; it is operated frame after frame. During the first step, the absolute value of each wavelet is compared to the memorised threshold at the corresponding level value; the number of wavelets below the threshold is $n$. An attenuation coefficient is then evaluated as function of $n$; this coefficient is linearly decreasing between 1.0 ($n$ = 129) and 0.2 ($n$ = 256). Inverse DWT is then operated after applying the same attenuation over all of the wavelets. This system is at this time in the course of evaluation and only the first results are available. The final results should be presented at the time of the conference.

## 7. CONCLUSION

In this paper we have presented a comparison between two methods for sound classification in the framework of a medical remote monitoring application. Analysis and classification of sounds emitted in patient's habitation may be useful for patient's activity monitoring. An adapted sound corpus was recorded in experimental conditions and used for evaluation purpose; this corpus includes 8 sound classes that are useful for this application. GMM-based methods are frequently used for sound classification in smart rooms because of their low calculus consumption, but HMM-based methods should allow a finer analysis: indeed the use of 3 states HMMs should allow better performances by taking into account the temporal shape of the signal.

The two approaches are presented, as well as the needed acoustical features. Then, an evaluation is made with the initial corpus and with additional experimental noise in order to compare these two methods. In the same noise conditions, HMM results are always the best. Best results are achieved with the original corpus (SNR = +28 dB), the Classification Error Rate is below 2%. At +17 dB, the CER is below 10% with 24 LFCC parameters and their derivatives of first and second order. However, the time consumption is very important in the case of the HMM algorithm and its implementation in a real-time system will involve to greatly optimize the algorithm and to use fast processors.

At the end, a segmentation module is presented. This module is able to extract isolated sounds in a record by the means of a wavelet filtering method which allows the extraction in noisy conditions. We are

working to add the possibility of speech segmentation in order to extract at once speech and sounds from a wave record.

## REFERENCES


1. RIALLE V., LAMY J.B., NOURY N., BAJOLLE L., *Remote monitoring of patients at home: A software Agent approach*, Computer Methods and Programs in Biomedicine, **Vol. 72**, Issue 3, pp. 257-268, 2003.
2. VACHER M., SERIGNAT J.-F., CHAILLOL S., ISTRATE D., POPESCU V., *Speech and Sound Use in a Remote Monitoring System for Health Care*, Lecture Notes in Computer Science, Artificial Intelligence, Text Speech and Dialogue, Brno, Czech Republic, **Vol. 4188**, Springer, pp. 711-718, 2006.
3. BONASTRE J.-F., ALIZE: A software toolkit for Speaker Recognition, http://www.lia.univ-avignon.fr/heberges/ALIZE/, 2004.
4. AJMERA J., MCCOWAN L., BOURLARD H., *Speech/music segmentation using entropy and dynamism features in a HMM classification framework*, Speech Communication 2003, **Vol. 40**, pp.351-363, 2003.
5. LEFEVRE S., MAILLARD B., VINCENT N., *A two level classifier for audio segmentation*, IEEE International Conference on Pattern Recognition, ICPR'02 Proceedings, **Vol. 3**, Aug. 2002.
6. REYES-GOMEZ M. J., ELLIS D. P., *Selection Parameter Estimation and Discriminative Training of Hidden Markov Models for General Audio Modeling*, IEEE International Conference on Multimedia and Expo, ICME'03 Proceedings, **Vol. 1**, pp. 73-76, July 2003.
7. ISTRATE D., CASTELLI E., VACHER M., BESACIER L., SERIGNAT J.-F., *Information Extraction From Sound for Medical Telemonitoring*, IEEE Transactions on Information Technology in Biomedicine, **Vol. 10**, NO. 2, pp. 264-274, April, 2006.
8. PINQUIER J., SENAC C., ANDRE-OBRECHT R., Speech *and music classification in audio documents*, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASP 2002, **Vol. 4**, pp. 4164, 2002.
9. YAMADA T., WATANABE N., *Detection using non-Speech Models and HMM Composition*, Workshop on Hands-free Speech Communication, Tokyo, Japan, 2001.
10. VICSI K., SZASZAK G., *Prosodic Cues for Automatic Phrase Boundary Detection in ASR*, Lecture Notes in Computer Science, Artificial Intelligence, Text Speech and Dialogue, Brno, Czech Republic, **Vol. 4188**, Springer, pp. 547-554, 2006.
11. CARUNTU A., TODOREAN G., *A Comparative Study of the Methods Used in Isolated Word Recognition*, Speech Technology and Human- Computer Dialogue, pp. 155-159, 2003.
12. DAUDET L., TORRESANI B., *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.
13. BOITE R., BOULARD H., DUTOIT T., HANCQ J., LEICH H., *Traitement de la parole*, Presses polytechniques et universitaires normandes, Lausanne, pp. 175-321, 2000, ISBN 2-88074-388-5.
14. SCHWARZ G., *Estimating the dimension of a model*, Annals of Statistics, **Vol. 6**, pp. 461-464, 1978.
15. ROEDER K., WASSERMANN L., *Practical bayesian density estimation using mixtures of normals*, Journal of the American Statistical Association, **Vol. 92**, pp. 894-902, 1997.
16. COWLING M., SITTE R., *Analysis of speech recognition techniques for use in a non-speech recognition system*, IEEE Transactions on Speech and Audio Processing, **Vol. 10**, pp. 504-516, 2002.
17. DUFAUX A., BESACIER L., ANSORGE M., PELLANDINI F., *Automatic Sound Detection and Recognition for Noisy Environment*, Eusipco 2000, Tampere, Finland, Sep. 2000.
18. "Bruitage", *Bruitages Gratuits : Sound-Fishing.net*, http://www.sound-fishing.net/bruitages.htm, Nov. 2005.