SPEECH AND SOUND ANALYSIS: AN APPLICATION OF PROBABILISTIC MODELS

M. Vacher[1], D. Istrate[2], J.F. Serignat[1]

[1]LIG, UMR 5217, CNRS-INPG, Grenoble, France
[2]RMSE-ESIGETEL, Fontainebleau, France

Abstract: Sound signal carry rich information which can be extracted and used by analysis systems in different modes. The medical remote monitoring systems may help elderly people to live home in security. We have already proposed an environmental sound analysis algorithm to detect distress sound or sentences. The sound and speech recognition modules based on Gaussian Mixture Models and Hidden Markov Models are detailed. A real time implementation on an embedded PC is proposed and evaluated. This implementation is flexible either for the hardware target (desktop, laptop or embedded PC) or on the alarm sending (E-Mail, SMS or remote monitoring center).

Keywords: Signal processing algorithms, Gaussian distributions, Markov models, Pattern recognition, Real time computer systems, Embedded systems, Speech Analysis

1. INTRODUCTION

1.1 Sound Analysis Applications

Sound analysis is involved in several and various fields of investigation. It is often related to an increase interest for automatic monitoring systems. Sounds can be speech, music, songs or more generally a sound of the everyday life.

With regard to speech, Clavel *et al.*(2007) studied the detection and analysis of abnormal situations through fear-type acoustic manifestations. The framework of the application is the detection of abnormal situations in the context of civil safety. A large set of features are used in his work in order to characterize the emotional content: especially spectral features, pitch (F0) of the voice, jitter, shimmer and unvoiced rate. Classification is performed using the Gaussian Mixture Model (GMM) based approach and a Maximum *a Posteriori* decision rule. The parameters of the GMM are estimated during a training step. The results of the evaluation lead to an average accuracy for *Fear* vs. *Neutral* classification of 70%. The emotion recognition task is difficult given the diversity of fear manifestations.

Audio classification and scene classification are frequently studied in the context of hearing aids (Allegro *et al.*, 2001)(Ravindran and Anderson, 2005)(Fillon and Prado, 2003)(Nielsen *et al.*, 2006). These systems provide a variety of options and settings that can be tuned based on the audio environment such as e.g. quiet environment, noisy environment, music, etc. All the type of sounds and not only speech are concerned with this application. Features are estimated from the spectrum of the signal as firstly the so called auditory features (Allegro *et al.*, 2001) -spectral separation, frequency variations, harmonicity, etc.-, and secondly the pitch and his dynamic properties (Nielsen *et al.*, 2006). Classifiers are simple Bayes classifiers, GMMs or Hidden Markov Models (HMMs). Results obtained by Nielsen are 5% classification error rate between music, noise and speech for 1 s analysis windows.

An other kind of application is automatic alarm detection and robotics (Kraft *et al.*, 2005)(Dufaux, 2001)(Cowling, 2004). Only sounds related to the specific application are taken into account. Features are Linear Predictive Coding (LPC) coefficients or cepstral coefficients like Mel Frequency Cepstral Coefficients (MFCC). These features are classically

used in speech recognition. Classification is achieved by employing varied methods, GMMs, HMMs, Learning Vector Quantization (LVQ) or Artificial Neutral Networks (ANN). Results depend on the application and on the corresponding corpus. Dufaux evaluated the robustness of the proposed method (Dufaux, 2001) by studying the influence of the Signal to Noise Ratio (SNR). Classification error rate is about 30% at +10 dB SNR in the case of GMM, 24% in the case of HMM.

Audio documents enclose information that can be used to index them thanks to MPEG-7. Moreover it provides informative methods for automatic description and sound matching (Casey, 2001) that use dimension-reduced, decorrelated spectral features (Audio Spectrum Projection) and Continuous HMMs as classifier. Kim *et al.* (2004) showed that MFCC yield slightly higher performances in most cases.

1.2 Aids for Ageing

All the industrialised countries and particularly Europe are affected by the ageing of the population. In 2030, 37% of the European population will be over the 60 years (European Commission, 2002). The number of elderly people which live alone will be very important, these persons have an increased risk of home accident due to their everyday life activity and moreover they account for 84% of the cases of falls. Medical remote monitoring may bring an appreciate help in the event of accident in the home by a fast transmission of alarm to a close relation or to emergency services. Currently these systems use several fixed sensors (infra-red or medical) and mobile sensors (fall detector, movement) (Bellogo *et al.*, 2006) but we already proposed a system analysing sounds in order to extract additional information (Istrate *et al.*, 2006). This system may be used for distress situation detection but it may be adapted in order to help the patient to interact with household appliances.

2. SOUND RECOGNITION

The general organisation of the sound recognition system is shown in **Error! Reference source not found.**. The microphone is connected to "Mic" input of the PC sound card.

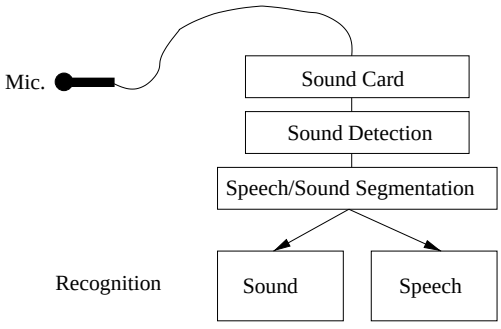

Fig. 1. - Overall Diagram of Sound Recognition

The recorded signal is transmitted to a detection stage in charge of sound extraction. This step is not the topic of this article but is fully detailed in (Istrate *et al.*, 2006). The following step is in charge of the segmentation of the sound signal between "life sound" and "speech" through statistical analysis, the signal is then transmitted to the corresponding recognizer. Actually it is very important to send non linguistic signal to the sound classifier and not to the speech recognizer. Segmentation and sound classification processes will now be explained.

2.1 GMM and HMM Methods

These methods evolve in two steps : a training step using a corpus and a classification step applied to a detected sound. The implementation in this framework uses the ALIZE library (Bonastre, 2004).

GMM METHOD. Each sound class k is modelled by N Gaussian distributions. The parameters of the Gaussian distribution m ($1 \le m \le N$) 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 $\Sigma_{k,m}^{-1}$. 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 Expectation Maximisation (EM) algorithm 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 of the n frames of the signal is evaluated according to equation (1) and then for the entire signal according to (2).

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

With:

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

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

HMM METHOD. HMMs take into account the temporal shape of the signal because the sound signal is decomposed into 3 components as shown in Fig. 2 (states q_1, q_2 and q_3 represent the "transient", "tonal" and "residual" components of the signal). The signal is surrounded by two silent states q_0 and q_4. A transition is always possible from each state q_i to a state q_j with a probability P_{ij} if j is greater than or equal to i, except for the transition from q_0 to q_4 which are the same state of silence. Each state of a class of sound is then modeled by a GMM in conjunction with the probability of transition. The classification step uses a Viterbi algorithm to estimate the class k of a sound X between m models M_j, a sound being represented by a sequence of p vectors of features X_1^p.

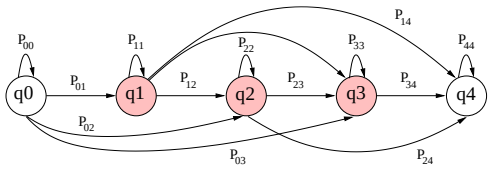

Fig. 2. - HMM State Transitions

k is the solution of equation (3) and $0 \le j < m$. All of the classes have an equal probability.

$$k = \arg\max_{j} p(X|M_j) \qquad (3)$$

The Viterbi algorithm is a forward probability method of best path estimation. The probabilities of each vector are evaluated through GMM models in conjunction with the probability of transition. The HMM training is more complex to describe and therefore it is not in the framework of this article.

2.2 Acoustical Features

Linear Frequency Cepstral Coefficients (LFCC) are used because of their equal sensitivity over the full bandwith contrary to MFCCs which use a MEL logarithmic scale. Energy is not used, this parameter being too dependent of experimental recording conditions. Derivatives of first (Δ) and second order are preferred (ΔΔ).

2.3 Sound and Speech Corpora

A sound corpus and a speech corpus needed for our study were recorded in the GETALP team of the LIG laboratory. Each sound or sentence is characteristic for a normal patient's activity (ringing phone, door lock, etc.), a possible distress situation (object fall, scream, distress key word like "Help me!", etc.), or a patient's physiology (cough, etc.). Sounds related to the patient's physiology are not taken into account because of the difficulty in recording such sounds. Sounds were recorded using omni-directional wireless microphones (SENNHEISER eW500), the sampling rate were 16 kHz. Each sound or sentence is recorded in one file in order to make further evaluation easier.

Sound Corpus. The sound corpus is made of 8 sound classes useful for our application, Glass Breaking, Object Falls and Screams are related to a possible distress situation. A small share of the sound corpus consists of 29% of sounds extracted from a preceding corpus recorded at the time of former studies in the CLIPS laboratory (Istrate *et al.*, 2006). Some sounds were obtained from the Web (Bruitages, 2005) and 61% are new records. The total duration of this corpus is 35 min 38 s.

Speech Corpus. The speech corpus has a total duration of 38 min and is constituted of 2646 audio files. This corpus was recorded by 21 speakers (11 men and 10 women) between 20 and 65 years old.

Table 1 - Everyday Sound Corpus

Sound Class	Number of Files	Average Duration of one Sound (ms)
Dishes Sounds	363	606
Door Lock	507	390
Door Slap	372	1022
Glass Breaking	118	269
Object Falls	120	1039
Ringing Phone	319	991
Screams	102	432
Step Sounds	76	86
Entire Corpus	1985	276

Each sentence is read by the speaker, it is not spontaneous speech. The corpus is composed of 126 sentences in French, 63 without any distress key word ("Il fait chaud", "Bonjour") or with distress keywords ("J'ai mal", "A l'aide"). This short corpus was recorded in order to evaluate the segmentation between sound and speech and the distress keyword extraction. The speech recognition system was trained with bigger corpora.

*NOISY CORPORA. E*valuation in realistic conditions requires noisy corpus. for this reason audio noise is used, it was recorded in our test apartment. it is a not stationary noise. a noisy sound corpus and a noisy speech corpus were generated from audio noise by adding to pure sounds for four Signals to Noise Ratios (SNR): 0, +10, +20 and +40 dB. For each SNR, the noised corpus is made of 2646 speech files and 1985 every day sound files.

2.4 Segmentation and Classification Evaluation

The evaluation is made using a cross validation protocol, training is made with 90% of the corpus and each the 10% remaining files is evaluated at the 4 signal to noise ratios. The Gaussian number depends on the training corpus and is determined through the Bayesian information criterion: 24 for segmentation between speech and sound and 12 for sound classification. The performances are evaluated through the Segmentation Error Rate (*SER*) and the Classification Error Rate (*CER*).

Results for segmentation are presented in Table 2. *SER* is better or equal to 5.1% if SNR is less than +10dB. Tables 3 and 4 present the *CER* for GMM and HMM methods. *CER* is less than 15.1% (GMM) or 9.7% (HMM) in realistic noise conditions ($SNR \ge +10\,dB$).

Table 2 - GMM Segmentation between Speech and Sound, 16LFCC, GMM, 24 Gaussian

Signal to Noise Ratio (dB)	0	+10	+20	+40
16 LFCC only	17.3%	5.1%	3.8%	3.6%

Table 3 - GMM Sound Classification, ΔΔ24 LFCC, 12 Gaussian

Signal to Noise Ratio (dB)	0	+10	+20	+40
24 LFCC	36.6%	21.3%	13%	9.3%
ΔΔ 24 LFCC	43.8%	15.1%	10.4%	7.3%

Table 4 - HMM Sound Classification, ΔΔ 24 LFCC, 12 Gaussian

Signal to Noise Ratio (dB)	0	+10	+20	+40
24 LFCC	29.8%	16.3%	6.6%	5.9%
ΔΔ 24 LFCC	28.3%	9.7%	5.7%	4.2%

HMM results are better at any SNR but this method is difficult to implement in real-time conditions because of the very high analysis time. Therefore the GMM method is implemented in our real-time application.

3. DISTRESS KEYWORD EXTRACTION

Isolated word recognition systems use either DTW algorithm (Gelin and Wellekens, 1997), neural networks or HMM (Ming, 1992). The LIG laboratory has an important experience in continuous speech recognition acquired in European Project C-STAR (Akbar, 1998). The Autonomous Speech Recognition (ASR) system RAPHAEL is used in order to detect distress keywords (Vacher *et al.*, 2006). This ASR has been adapted to distress sentences recognition. HMMs are used for phoneme recognition by the acoustical module of the system. Acoustical models have been trained with corpora recorded by near 300 speakers in order to assure speaker independence.

It is important that the keywords, and only these, related to a distress situation are well recognized. The speech recognition system has been evaluated with the sentences from 5 speakers of our corpus. For 16% of the distress sentences the distress keyword is missed and for 6% of the normal sentences an unexpected distress keyword is introduced by the system.

4. REAL TIME SYSTEM ON EMBEDDED PC

The sound analysis system has been implemented on an Embedded PC running WindowsXP Embedded. The advantages of this implementation are reduced dimensions, silence (fan less) and low cost. In the same time, this software implementation is flexible and can be installed also on desktop or laptop PC equipped with an internal/external sound card. The software can run on any Windows OS (Windows 2000/XP).

The system is divided in four parallel threads and implemented under LabWindows/CVI (National Instruments) as shown in the Fig. 3. The sound signal acquisition is made through sound card using the low Win32 functions which allow the use of a double circular buffer processed via software interruptions. The sample frequency is fixed to 16 KHz and the buffer dimension to *2 x 2048* samples corresponding to algorithm constraints. Each time that the half of the sound buffer is full, an interruption calls the detection algorithm. In the case of sound event detection the signal is recorded temporarily in a wave file. Since the file is recorded, the detection thread sends also a message (the wav file name) through a safe communication queue to the recognition thread.

The detection algorithm uses the Discrete Wavelet Transform (DWT) to analyse in time and frequency the continuous sound signal. A useful signal is detected if the energy of high frequency Wavelet Transform Coefficients exceeds a self-adjustable threshold (Istrate, 2006).

The recognition thread is started in parallel with the detection one and waits a message from detection algorithm. As soon as a message is received, the Sound/Speech Classification algorithm is executed. Then, if the signal is labelled as life sound, the Sound Recognition algorithm is started; otherwise, if the signal is labelled as speech, the corresponding wav file is sent to the speech recognition engine (RAPHAEL). In the two cases, the Event Analysis sub-module decides the action to be started according to the recognized event: if an alarm sound or a distress sentence has been detected, an alarm with the recorded sound is sent using the activated modality (email, SMS or TCP/IP to the remote monitoring center); if the processed event does not indicate an alarm situation the recorded file is deleted but however the type of event and the corresponding time are written in the history file. The possible choices of the action to carry out in the case of distress event detection allow an autonomous utilisation of the remote monitoring system.

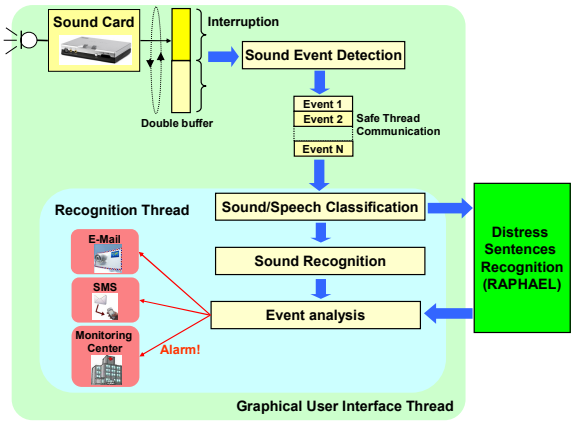

Fig. 3. - Parallel Threads Configuration for the Real Time Implementation

In this case the system can be used without interaction with a remote monitoring centre, only using an alarm message sending to a patient close person.

The application front-panel, presented in Fig. 4, displays in real time the sound signal, the list of previous detected events and a summary of main alarm action parameters. A special menu allows the user to specify the sound card to use (if more than one), to activate the action(s) to carry out in the case of alarm and to configure the parameters of these actions (email of the close person, SMTP email server, IP address of the remote monitoring center).

5. REAL-TIME EVALUATION

The presented software implementation was tested at LIG laboratory but also at RMSE laboratory. In the two cases five scenarios has been played by a person and the obtain results were compared to the system output. At LIG laboratory, the sound analysis software was tested on a desktop PC using internal sound card, a cardioids microphone Beyer Dynamics with his pre-amplifier. The scenarios were played in an office. At RMSE laboratory, the sound analysis software was tested on an embedded PC AEON - AEC 6810 using an USB sound card and an omnidirectional microphone Sennheiser (ME102) with pre-amplifier.

An example of played scenario is: “The phone ringing followed by a door lock sound and a door slap. Next a sound of chair fall precedes a sentence. Finally, a sound of box falling precedes a new sentence”. The detection module has no error on tested scenarios. The speech/sound segmentation module has an error rate of 9 % which confirm the obtained results on evaluation tests. The errors of segmentation module are rather to classify speech like sound (80% of errors) than oppositely. The sound classifier has an error rate about 45% but tested scenarios contain also 17.64 % unknown sounds like key sound.

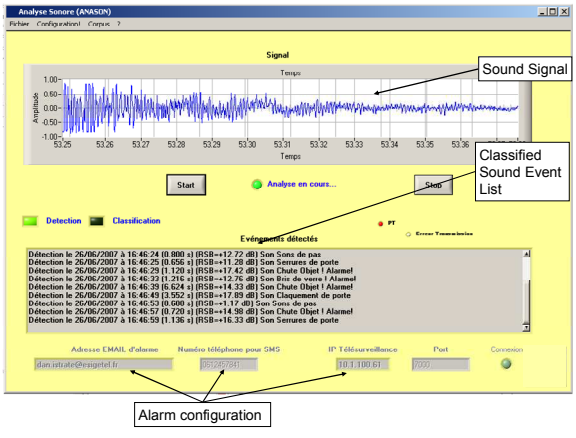

Fig. 4. - Software Front Panel

These results can be explained on the one hand by the SNR of signals (between 5 and 20 dB) and on the other hand by the fact that the current system cannot reject the unknown sounds.

Globally, the system has no false alarm and 20 % of missed detections. We consider *a missed detection* if an abnormal sound (glass breaking, screams, object falls) or a distress sentence was not detected, and in the same manner, *a false alarm* if a normal sound or sentence has been classified like an alarm. These tests showed that the real time implementation can be used in normal conditions in order to detect abnormal sound or distress sentences.

6. CONCLUSIONS AND PERSPECTIVES

We have presented an application of probabilistic models in a speech and sound analysis system. Objectives are to help distress situation identification in the domain of medical remote monitoring. This system has been implemented in real time on an embedded PC. The software implementation is flexible and can run on Windows OS PC equipped with a sound card and a microphone.

We are working now to implement the possibility of sound rejection in the case of an unknown sound by the sound recognition module.

REFERENCES

Allegro, S., M. Büchler and S. Launer, Automatic Sound Classification Inspired by Auditory Scene Analysis, Consistant and Reliable Acoustic Cues for Sound Analysis, CRAC, Eurospeech, Aalborg, DenMark, Sept. 2001.

Akbar, M. and J. Caelen, *Parole et traduction automatique : le module de reconnaissance RAPHAEL*, COLING-ACL’98, Montréal, Quebec, vol.2, p. 36-40.

Bellogo, G.L., N. Noury, G. Virone, M. Mousseau and J. Demongeot, *Measurement and model of the activity of a patient in his hospital suite*, IEEE Transactions on Information Technology in Biomedicine, vol. 10, pp. 92-99, January 2006.

Bonastre, J.F., ALIZE: A software toolkit for Speaker Recognition , http://www.lia.univavignon.fr/heberges/ALIZE/, 2004.

“Bruitage”, Bruitages gratuits : Sound-Fishing.net, http://www.soundfishing.net/bruitages.htm, Nov. 2005.

Casey, M., *MPEG-7 Sound Recognition Tools*, IEEE Transactions on Circuits and Systems for Video, vol. 11, no 6, June 2001.

Clavel, C., L. Devillers, G. Richard, I. Vasilescu and T. Ehrette, *Detection and Analysis of Abnormal Situations through Fear-Type Acoustic Manifestations*, IEEE Transactions on Speech and Audio Processing, vol 4, pp. 21-24, 2007.

Cowling, M., Non-Speech Environmental Sound Classification System for Autonomous Surveillance, Ph.D. dissertation, Faculty of Engineering and Information Technology, Griffith University, USA, 2004.

Dufaux, A., "Detection and Recognition of Impulsive Sounds Signals", Ph;D. dissertation, Electronics and Signal Processing Dept., Faculté des Sciences de l'Univ. de Neuchatel, Switzerland, 2001.

European Commission, "Europe's response to World Ageing. Promoting economic and social progress in ageing world", Second World Assembly on Ageing, 18 March 2002.

Fillon, T. and J. Prado, *Evaluation of an ERB frequency scale noise reduction for Hearing Aids: a comparative study*, Speech Communication, vol. 39, pp. 23-32, Jan. 2003.

Gelin, P. and C.J. Wellekens,, *Keyword spotting for multimedia document indexing*, Proc. SPIE Vol. 3229, p. 366-377, Multimedia Storage and Archiving Systems II, 1997

Istrate, D., E. Castelli, M. Vacher, L. Besacier and J.-F. Serignat, *Information extraction from sound for medical telemonitoring,* IEEE Transactions on Information Technology in Biomedicine, vol. 10, no. 2, pp. 264-274, April 2006.

Kim, H., J. Burred and T. Sikora, "How efficient is MPEG-7 for general sound recognition", in Proceedings of AES 25th International conference, United Kingdom, June 2004.

Kraft, F., R. Malkin, T. Schaaf and A. Waibel, "Temporal ICA for Classification of Acoustic Events in a Kitchen Environment", in Proceedings of Interspeech, Lisbon, Portugal, pp. 2689-2692, 2005.

Ming, Z.; *Speaker independent recognition of isolated words using vector quantization and neural networks*, Ph.D. Thesis Technische Univ., Berlin, 1992

Nielsen, A., L. Hansen and U. Kjems, Pitch Based Sound Classification, in Proceedings of Acoustics, Speech and Signal Processing, ICASP 2006, vol. 3, Toulouse, France, 2006.

Ravindran, S. and D.V. Anderson, "Audio Classification and Scene Recognition for Hearing Aids", in Proceedings of Circuits and Systems, ISCAS, vol. 2, pp. 860-863, May 2005.

Vacher, M., J.-F. Serignat, S. Chaillol, D. Istrate and V. Popescu, *Speech and Sound Use in a Remote Monitoring System for Health Care*, Lecture Notes in Computer Science, Artificial Intelligence, Text Speech and Dialogue, vol. 4188, pp.711-718, April 2006.