FIRST IMPLEMENTATION OF A SOUND/SPEECH REMOTE MONITORING REAL-TIME SYSTEM FOR HOME HEALTHCARE

M Vacher, J.-F Serignat, Pelayo Menendez-Garcia, D Istrate

FIRST IMPLEMENTATION OF A SOUND/SPEECH REMOTE MONITORING REAL-TIME SYSTEM FOR HOME HEALTHCARE

M. Vacher, J.-F. Serignat, Pelayo Menendez-Garcia

CLIPS-IMAG, Team GEOD
UMR CNRS-UJF-INPG 5524
385, rue de la Bibliothèque
BP 53, F-38 041 Grenoble cedex 9
{Michel.Vacher, Jean-Francois.Serignat}@imag.fr

D. Istrate

RMSE-ESIGETEL
1, Rue du Port des Valvins
F-77 215 Avon-Fontainebleau
email: Dan.Istrate@esigetel.fr

ABSTRACT

Medical Remote Monitoring needs human operator assistance by smart information systems. Physiological and position sensors give already numerous informations, but sound classification can give interesting additional informations about the patient and may help to the decision-making. A Real-Time implementation of a multichannel smart sound/speech system is presented in this paper, it is capable in a first step to detect and identify sound events in noisy conditions, in a second step to classify it into speech or life sound. According to that result the third step is speech recognition or sound classification. The multichannel sound processing allows us to localize the sound in the apartment and to select appropriate signal segments for identification procedure. Recognized sentences and classification results are sent to the medical remote monitoring application through Ethernet network. The system is composed of several parallel tasks: detection & channel selection, sound/speech classification, life sound classification, speech recognition and graphical interface. The event detection module is carried out for each channel in real time. The classification modules are launched in a parallel task. Speech Recognition is running as an independent application on the same computer.

1. INTRODUCTION

Nowadays the required sensors capabilities are increased very much: these sensors become more and more complex, the digital signal processing being a crucial component of them. The most difficult task in digital signal processing for sound sensor is the extraction of high-level information from an one-dimensional signal. The evolution of the man-machine interfaces involves the development of sound sensors and their use for speech or sound recognition.

The processed output signal of a sound sensor is frequently an enhanced audio signal but it could be a series of different types of sound information: words or speaker name in case of speech, sound class in case of everyday life sounds, type of noise and so on. Unlike the speech/speaker recognition domain, there are only few studies in the sound class identification field.

In this paper we describe the software implementation of a multichannel smart sound/speech system:

- sounds are detected and identified between several predefined sound classes,
- speech is analyzed and sentences are recognized.

This system is a part of a medical remote monitoring project with the aim of detecting abnormal patient behaviour at home in case of residential health care[1]. The medical monitoring system not described in this article uses this sound system and sensors to take its decision: medical sensors (oxymeter, tensiometer, thermometer and actimeter), various sensors (infrared sensors and door contacts).

Each sound produced in the apartment is characteristic of:

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

If the system has a good ability of classification for such sounds [2], it will be possible to know if the patient is needing help.

In the same way, the speech said by the patient may give precious information on the patient:

- a distress case: "Help me!", "Doctor!",
- a normal state: "Coffee is cold!", "The door is open!".

The input of the smart sound system is composed of data collected by the means of 5 to 8 microphones (one per room).

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.

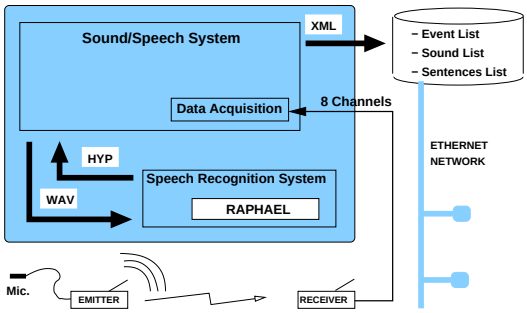

Fig. 1. Global System Organisation

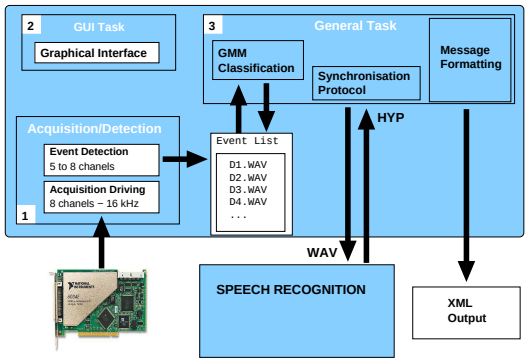

Fig. 2. Real-Time Flow-Chart

2. SYSTEM ORGANISATION

The general organisation of the software is shown in Figure 1. The sound/speech system and the speech recognition system are running as independent applications on the same computer, they are synchronized through a file exchange protocol.

For a real time working purpose, the sound/speech analysis is divided in three modules : the graphical user interface, the sound event detection for each sound channel and the general sound analysis. The event detection has the highest priority and can not be interrupted because signal frames can never be lost.

The extracted information is sent through the network using XML format and if it is needed, the recorded sound can be transferred for latter analysis. Useful informations are: time and date, name of the room, classification result (sound or speech, class of the life sounds, recognized sentences).

Here is an example of results in case of speech recognition in the kitchen, log. likelihood was -20.2 for "No Speech", -17.2 for "Speech":

```
<appli:segmentation description="appli audio "
type="son">
<piece>Cuisine</piece>
<horodate>1-12-2005 a 15:19:20</horodate>
<resultat>parole</resultat>
<information>Probabilite de son=-20.2018, Proba
bilite de parole=-17.2258</information>
</appli:segmentation>
<appli:reconnaissance description="appli audio"
type="son">
<piece>Cuisine</piece>
<horodate>1-12-2005 a 15:19:20</horodate>
<resultat>un docteur vite</resultat>
</appli:reconnaissance>
```

Therefore the sound event was classified as *"parole"* (speech) and the recognized sentence was in French *"un docteur vite"* (a doctor quickly).

3. SOUND SYSTEM IMPLEMENTATION

The software is developed with LabWindows/CVI, including card drivers. It is running on a single PC with 1Gb RAM, the clock frequency of the mono-processor is 3 GHz. For sound sample acquisition, low-level functions are used in order to drive the card in real time. The detected events are saved temporarily on PC hard-disk. The sound part is composed of 5 up to 8 HF microphones, an acquisition card (National Instruments PCI-6034E) plugged in the PC what is in charge with the sound/speech analysis software.

HF microphones (SENNHEISER eW500) are used because of their small dimensions and of their omnidirectional characteristics. Each receiver (32MHz frequency band) is connected to a channel of the acquisition card. Acquisition card has 8 differential inputs and a maximal sampling rate of 200 ksamples/s. The sampling frequency was fixed at 16 kHz. This value is usual in speech recognition.

4. REAL-TIME SOUND ARCHITECTURE

4.1. Acquisition and detection

The sound analysis system has been divided in three tasks as shown in Figure 2.

The acquisition and detection modules are making up the *First Parallel Task*. The first task modules are processed on each channel in order to acquire and detect a sound event and to extract it from signal flow. This task is initiated by an high priority interrupt signal as soon as the active half buffer of the acquisition card is full. At this time the other half buffer is activated for acquisition, data of the 8 channels are copied into memory and analysed by the detection module. In case of simultaneous events and sound identification, they are all analysed and the energy level is stored in the XML data for later analysis.

There are many techniques for sound detection: energy threshold, statistical model [3], energy processing [4] or wavelet processing [5]. Unlike Fast Fourier Transform, Wavelet Transform is well adapted to signals that have very localized features in the time-frequency space. This transform is frequently used for signal detection [6] and audio processing. A wavelet based event detection algorithm has been proposed [7]. We have chosen Daubechies wavelets with 6 vanishing moments to compute DWT on windows of 2,048 samples. The size of each half buffer used for acquiring the signal is 1,024 samples. After acquisition these data are stored in a

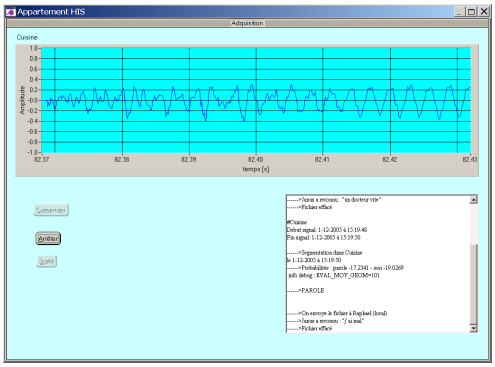

Fig. 3. Acquiring Front Panel

circular buffer. The detection analysis is completed as soon as 2 half buffers are stored in the circular buffer. The dyadic properties of DWT and the analysis of wavelet trees of 32 milliseconds each allows us a good time resolution to obtain start and stop time of the signal event.

4.2. Graphical Interface

The graphical user interface is the *Second Parallel Task*. It is used for setting up the application and to display signal and results on the screen as they are available (see Figure 3): channel calibration, wave visualisation, choice of features, detection/classification results and recognized sentences.

4.3. General Task

The *Third Parallel Task* has the ability to send its results through the network to the medical monitoring system in XML format. This task is receiving the extracted sound event (output of *First Parallel Task*) and is composed of several modules in charge of:

- Segmentation between "Speech" and "No Speech": this module aims to identify the kind of sound event between these 2 classes because further processings are different. This module is extracting the acoustical features from the recorded event (16 LFCC, Linear Frequencies Cepstral Coefficients), then it is using a Gaussian Mixture Model (GMM) method[8]. The training step is not involved in this real-time application, the corpus used for training is described in section 5. The Gaussian model number is 24 and was chosen after a BIC (Bayesian Information Criterion) study. LFCC where chosen after statistical studies on the corpus (FDR, Fisher Discriminant Ratio).
- Control of the Speech Recognizer : this module is initiated when sound was classified as speech. It is in charge of sending a new file if there is no work in progress and in charge of reading the hypothesis answer. Only one wave file can be processed at once.
- Classification of life sounds into 7 predefined classes : this module is initiated when the sound event was classified as life sound, it aims to identify the event sound between 7 classes which are useful for our application (door lock, door slap, glass breaking, ringing phone, step sounds, screams, dishes sounds). This module is processing the acoustical features of the recorded event, then it is using a GMM method like in the segmentation module. The used features are 16 MFCC (Mel Frequency Cepstral Coefficients), Zero Crossing Rate, Roll-Off Point and Centroid; the number of Gaussian models is 4. These values are different of the values used in case of segmentation because of the differences in the corresponding corpus.

5. TRAINING CORPUS

In order to train, to test and to validate the system we have generated a *life sound corpus* and recorded an adapted *speech corpus*. With these two corpus we have generated a noised corpus with 4 signal to noise ratio (SNR=0dB, +10dB, +20dB, +40dB), noise was recorded in our experimental test apartment [9]. This noised corpus was used for evaluation of detection and classification modules.

5.1. Life sound corpus

The everyday life sounds are divided into 7 classes related to 2 categories: *normal* sounds related to an usual activity of the patient (door clapping, phone ringing, step sounds, dishes sounds, door lock), *abnormal* sounds related to a distress situation (breaking glasses, screams). This corpus contains recordings made at the CLIPS laboratory (15%), files of "Sound Scene Database in Real Acoustical Environment"[10] (70%) and files from a commercial CD (15%). 20 types of sounds were selected with 10 to 300 repetitions per type.

5.2. Speech corpus

This corpus has been recorded at the CLIPS laboratory by 21 speakers (11 men and 10 women) between 20 and 65 years old. It is composed of 126 sentences in French: 66 are characteristic of a normal situation for the patient: "Bonjour" (Hello), "O ù est le sel" (Where is the salt)... and 60 are distress sentences: "Au secours" (Help), "Un m édecin vite" (A doctor quick)... This corpus has a total duration of 38 minutes and is constituted by 2,646 audio files.

6. SPEECH RECOGNITION SYSTEM

The autonomous Speech Recognition System RAPHAEL is used [11]. The language model of this system is a medium vocable statistical model (around 11,000 words). This model is achieved through textual informations extracted from the

WEB as described by D. Vaufreydaz [12] and optimized for the distress sentences of our corpus. The training of the acoustical models of RAPHAEL are made with large corpora recoded with a large number of French speakers.

The synchronisation with the sound/speech system is achieved by a file exchange protocol. As soon the requested wave file has been analysed by RAPHAEL, it is erased and the found hypothesis is stored in a hypothesis file. An other wave file may then be analysed.

7. EVALUATION AND FIRST RESULTS

The acquiring module was evaluated from Receiver Operating Curves giving *missed detection rate* as function of *false detection rate*. The Equal Error Rate (EER) is 0% below +10dB of SNR and 6.5% at 0dB. The segmentation between speech and sound was evaluated with a "cross-validation" protocol. The Error Segmentation Rate is 5% below +10dB and 17% at 0dB. The classification of the everyday life sounds was evaluated with a "leave one out protocol". The Error Classification Rate is 13% below +20dB, 27% at +10dB.

It is very important that key words related to a distress situation will be well recognized. The speech recognition system has been evaluated with the sentences of 5 speakers of our corpus (630 tests). For normal sentences and in 6% of the cases, an unexpected distress key word is introduced by the system and leads a *False Sentence Alarm*. For distress sentences and in 16% of the cases, the distress key word is not recognized and then missed: that leads a *Missed Sentence Alarm*. It often occurs in isolated words like "A ïe" (Ouch) or "SOS" or in French syntactically incorrect expressions like "Ça va pas bien" (I am not feeling very well). The language model has to be best optimized and the dictionary completed to obtain lower error rate.

8. CONCLUSIONS AND PERSPECTIVES

In this paper we have presented the Real-Time implementation of a multichannel sound processing system which detects sound events, identifies sounds among 7 predefined sound classes and recognize French speech sentences in order to detect distress key words. Thanks to the acquiring part of the system, it can run on an operating system without real-time capacity. It may be used under realistic condition with moderate noise: +10dB SNR. This system as the ability of extracting new additional informations from sound which may be very useful for the medical monitoring system to take a decision in a distress case.

9. REFERENCES

[1] V. Rialle, J.B. Lamy, N. Noury, and L. Bajolle, "Telemonitoring of patients at home: A software agent approach," *Computer Methods and Programs in Biomedicine*, vol. 72, no. 3, pp. 257–268, 2003.

[2] M. Vacher, D. Istrate, L. Besacier, J.F. Serignat, and E. Castelli, "Sound detection and classification for medical telesurvey," in *Proc. 2nd Conference on Biomedical Engineering*, Calgary ACTA Press, Ed., Innsbruck, Austria, Feb. 2004, ISBN 0-88986-379-2, pp. 395–398.

[3] 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.

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

[5] L. Daudet, *Représentations structurelles de signaux audiophoniques - Méthodes hybrides pour des applications à la compression*, Ph.D. thesis, Université de Provence, Marseille, 2000.

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

[7] M. Vacher, D. Istrate, and J.F. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," in *Proc. 12th European Signal Processing Conference*, Suvisoft LTD, Ed., Vienna, Austria, Sep. 2004, pp. 1171–1174.

[8] 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.

[9] 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.

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

[11] M. Akbar and J. Caelen, "Parole et traduction automatique : le module de reconnaissance raphael," in *Proc. COLING-ACL'98*, Montréal, Quebec, 1998, vol. 2, pp. 36–40.

[12] D. Vaufreydaz, J. Rouillard, and M. Akbar, "Internet documents: a rich source for spoken language modeling," in *Proc. IEEE Workshop ASRU'99*, Keystone-Colorado, USA, Dec. 1999, pp. 277–281.