Speech and Sound Use in a Remote Monitoring System for Health Care

Michel Vacher, Jean-François Serignat, Stéphane Chaillol, Dan Istrate, Vladimir Popescu

Speech and Sound Use in a Remote Monitoring System for Health Care

Michel Vacher, Jean-François Serignat, Stéphane Chaillol, Dan Istrate, and Vladimir Popescu

CLIPS-IMAG, UMR CNRS-UJF-INPG 5524,
BP53, 38041 GRENOBLE Cedex9, France,
Michel.Vacher@imag.fr ,
WWW home page: http://www-clips.imag.fr/geod/User/michel.vacher

Abstract. Ageing affects the economic and social foundations of societies at world level. Health care has to respond to the challenge that population ageing presents. Medical remote monitoring needs human operator to be assisted by means of smart information systems. Physiological and position sensors give numerous data, but speech analysis and sound classification can give interesting additional information about the patient and may help in decision-making. The entire analysis system is composed of parallel tasks: signal detection & channel selection, sound/speech classification, life sound classification and speech recognition. The multichannel sound processing allows us to localize the source of sound in the apartment and to select appropriate signal segments for analysis. Recognized key words indicative of a distress situation are extracted from sentences. Key words and classification results are sent to the medical remote monitoring application through network. An adapted speech corpus was recorded in French and used for evaluation purposes.

1 Introduction

It is well known that ageing is emerging as an important concern for the most developed countries, but in the 21st century rapid ageing will progressively become a global phenomenon [1]. In the developed countries, older people will constitute 33% of their population -but 37% in Europe- in 2050 as opposed to 19% today, and the median age will increase by 9 years, reaching 46 years in 2050. 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, meeting the growing demand for health and care services, related to the significant growth of 80 years and over old people. Progress in aids and assistive technologies might be a cost-efficient way to support the supply of informal care and care provisions.

Therefore, effective medical monitoring from a remote location, requires that a smart information system is used to alert a human operator of patient distress. Presently, physiological and position sensors give a variety of data, but do not take into account distress calls or fall sounds. According to that, speech analysis

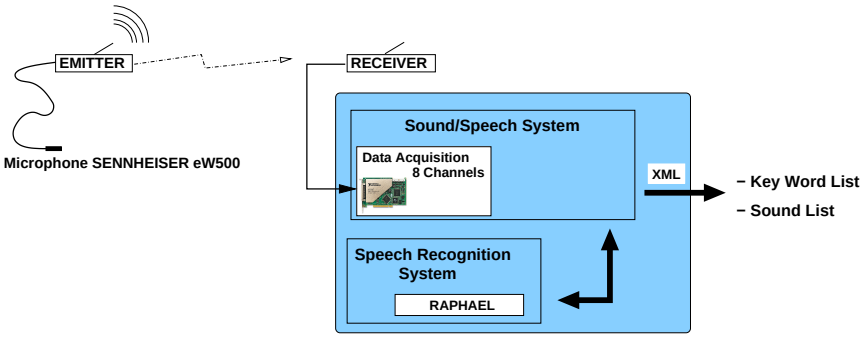

Fig. 1. Global System Organization

and sound classification can give interesting additional information about the patient and may help the decision-making.

In this paper we describe the speech/sound analysis part of a multichannel smart system:

- speech is analyzed in order to extract informative key words,
- sounds are detected and identified among several predefined sound classes.

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 [2]. The medical monitoring system not described in this article uses the sound system and sensors to take its decision: medical sensors (oxymeter, tensiometer, thermometer and actimeter), various sensors (infrared sensors and door contacts). The input of the smart sound system is composed of data collected by the means of 5 to 8 microphones (one per room).

The sentence uttered by the patient may give valuable information on the patient:

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

In the same way, each sound produced in the apartment is indicative 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 to recognize speech and to classify such sounds, it will be possible to know if the patient needs help.

2 System Organization

The general organization of the system is shown in Figure 1. High Frequency microphones (SENNHEISER eW500) are used because of their small dimensions

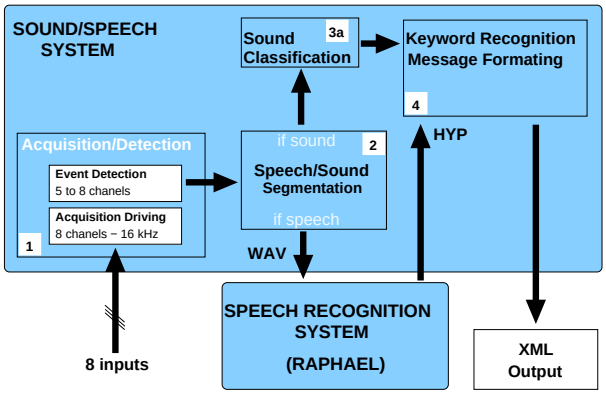

Fig. 2. Sound Flow-Chart

and of their omnidirectional characteristics. Each receiver (32 MHz frequency band) is connected to a channel of the acquisition card. The sound synchronisation system is made up by the *sound/speech system* and the *speech recognition system* as shown in Figure 2. These two components are running as independent applications on the same computer, they are synchronized through a file exchange protocol. The first stage of the *sound/speech system* is the acquisition and detection module, it is not describe in this article. The 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. The acquisition module was evaluated via Receiver Operating Curves giving *missed detection rate* as function of *false detection rate*. The Equal Error Rate (EER) is 0% above +10 dB of SNR and 6.5% at 0 dB. After the detection step, the signal has to be sent only to the appropriate stage: *speech recognition system* or sound classifier. This decision-making, speech/sound segmentation, is achieved by the second stage. It is very important to send non linguistic sound to the classifier and not to the recognizer.

At the end the useful extracted information is sent by the keyword recognition and message formatting stage through the network using XML format. Useful information is: time and date, name of the room, classification result (sound or speech, recognized keywords, class of the life sounds). Figure 3 shows an example of results coded in XML format in case of speech recognition in the kitchen: the logarithmic likelihood was -20.2 for "No Speech", -17.2 for "Speech", therefore the sound event was classified as *"parole"* (speech) and sent to the speech recognizer RAPHAEL. The recognized sentence was in French *"un docteur vite"* (a doctor quickly).

3 Corpus for Training and Evaluation

In order to train, test and validate the system we have recorded an adapted *speech corpus* in French and a *life sound corpus*. With these two corpora we

```
<appli:segmentation description="appli audio">
<pièce>Cuisine</pièce>
<horodate>1-12-2005 à 15:19:20</horodate>
<résultat>parole</résultat>
<information>Probabilité de son=-20.2018, Probabilité de parole=-17.2258</information>
</appli:segmentation>
<appli:reconnaissance description="appli audio">
<pièce>Cuisine</pièce>
<horodate>1-12-2005 à 15:19:20</horodate>
<résultat>un docteur vite</résultat>
</appli:reconnaissance>
```

Fig. 3. Result Sample in the case of Speech

have generated a noised corpus with 4 levels of signal to noise ratio (SNR=0 dB, +10 dB, +20 dB, +40 dB). The HIS (*Habitat Intelligent pour la Santé*) noise was recorded in an experimental test apartment [3]. This noised corpus was used for evaluation of detection and classification modules.

Speech corpus This corpus has been recorded at 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), "Où est le sel" (Where is the salt)... and 60 are distress sentences: "Au secours" (Help), "Un médecin vite" (A doctor quickly)... This corpus has a total duration of 38 minutes and is constituted by 2,646 audio files in wave format.

Life sound corpus The every day life sounds are divided into 7 classes related to 2 categories: *normal* sounds related to usual activities of the patient (door clapping, phone ringing, step sounds, dishes sounds, door lock), *abnormal* sounds related to distress situations (breaking glasses, screams). This corpus contains recordings made at CLIPS laboratory (15%), files of "Sound Scene Database in Real Acoustical Environment" [4] (70%) and files from a commercial CD (15%). 20 types of sounds were selected with 10 to 300 repetitions per type.

4 Speech/Sound Segmentation System

4.1 Segmentation

A Gaussian Mixture Model (GMM) method is used in order to classify sounds into speech or life sounds [5], [6]. There are other possibilities : Hidden Markov Model (HMM) [8], Bayesian method, etc. GMM has been chosen because with other methods similar results have been obtained, however at the cost of higher complexity. The segmentation system can not use 1 s windows like in [6], [7] because the audio signal can be as short as 36 ms. A preliminary step before signal classification is the extraction of acoustic parameters. LFCC (Linear Frequency Cepstral Coefficients) and normalised energy as additional parameter are used. The bandwidth of filters is constant and lead to a good resolution in high frequencies; life sounds are better discriminated with LFCC from speech

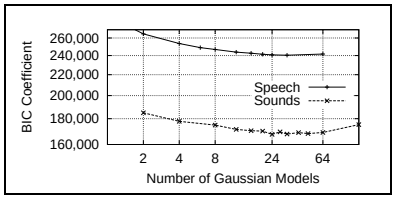

Fig. 4. BIC Coefficient Evolution as Function of Gaussian Model Number

by using high frequency components than with MFCC (MEL scale). For each frame, the energy is normalised with the average of energies of the frames of the complete signal and may be in this way less dependent on experimental recording conditions.

The classification with a GMM method supposes that the distribution of acoustic parameters for a sound class may be modeled with a sum of Gaussian models after a training step (K-means followed by Expectation Maximisation in 20 steps). The BIC (Bayesian Information Criterion) is used in this paper in order to determinate the optimal number of Gaussian models. It selects the model through the maximization of integrated likelihood: $BIC_{m,\kappa} = -2L_{m,\kappa} + \nu_{m,\kappa} \ln(n)$, where $L_{m,\kappa}$ is the logarithmic maximum of likelihood, equal to $f(x|m, \kappa, \hat{\theta})$ (f is integrated likelihood), m is the model and κ the component number of the model, $\nu_{m,\kappa}$ is the number of free parameters of model m and n is the number of frames. The minimum of BIC indicates the best model.

The BIC has been calculated for the speech class and for the life sound class with various number of Gaussian models. Results are given on figure 4 in the case of speech class (continuous line) and life sound class (dashed line): 24 Gaussian models seems to be a good choice.

4.2 Evaluation

The analysis window was set to 16 ms (2^8 samples, sample rate fixed to 16 kHz) with an overlap of 8 ms. The segmentation between speech and sound was evaluated with a "cross-validation" protocol: training is achieved with 80% of the corpus, and each file of the 20% remaining is tested according to the models. Training is performed with pure sounds and testing with sounds mixed with HIS noise at 0, +10, +20 and +40 dB levels. Cutting is made to insure that no test is done on a model trained with the same speaker or the same sentence.

Sound classification performances are evaluated through the segmentation error rate (SER) which represents the ratio between the wrong classified sounds and the total number of sounds to be classified. In table 1 the segmentation results are presented for LFCC parameters coupled or not with normalized energy. The best results are achieved for LFCC with energy: the Segmentation Error Rate is 4% below +10 dB and 14.5% at 0 dB.

Detection and Segmentation stages For global system evaluation we have built a test set containing a mixture between real noise (recorded in the test apartment),

Signal to Noise Ratio	0 dB	+10 dB	+20 dB	+40 dB
16 LFCC alone	17.3%	5.1%	3.8%	3.6%
16 LFCC with normalized energy	14.5%	3.9%	3.9%	4%

Table 1. Segmentation Error Rate (24 Gaussian models), cross-validation on the whole corpus (Speech & Life Sounds).

22 sentences (2 different speakers) and 23 life sounds. The noise level is constant and SNR is equally distributed between +10 dB and +40 dB by signal level variation. The duration of the test file is 3,600 s (1 hour).

The detection stage has extracted 44 audio signals, whereas 1 signal has been missed. The classification results for the 44 extracted signals are given in table 2. Speakers, sentences and life sounds of the test set are put out for training like in the cross-validation protocol.

The best results are encountered with LFCC parameters. Normalized energy is not relevant because this feature remains too dependent on experimental recording conditions, although it is less dependent than non normalized energy. Loss of performances for SNR between +40 dB and +10 dB is about 2%.

Features	16 LFCC alone	16 LFCC with normalised energy
Error Rate	6.7%	8.9%

Table 2. Global Segmentation Error Rate (24 Gaussian models), detection and segmentation on 45 audio signals.

5 Speech Recognition System

5.1 RAPHAEL

For Speech Recognition, the autonomous system RAPHAEL is used [9]. The language model of this system is a medium vocabulary statistical model (around 11,000 words). This model is obtained through textual information extracted from the Internet as described in [10] and from "Le Monde" corpora. It is then optimized for the distress sentences of our corpus. In order to insure a good speaker independence, the training of the acoustic models of RAPHAEL has been made with large corpora recorded with near 300 French speakers: respectively 80 speakers (BREF80 corpus), 120 (BREF120 [11]) and 100 (BRAF100 [12]).

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

5.2 Evaluation and first results

It is very important that the key words related to a distress situation will be well recognized. The speech recognition system has been evaluated with the sentences from 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 missed: this leads a *Missed Sentence Alarm*. It often occurs in isolated words like "Aïe" (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 of the speech recognizer to be completed in order to obtain lower error rate.

6 Sound Classification

GMM and HMM are well adapted to sound classification [6], [8]. For this framework the life sound corpus has been supplemented with a new class (object falls). The preliminary results are given in table 3. The analysis window was set to 16 ms with an overlap of 8 ms. The classification was achieved with a "cross validation-protocol". We used LFCC features with first and second derivatives. The optimal number of Gaussian models was determined using the BIC criterion: a number of 12 Gaussian models is appropriate for the GMM method with the 8 sound classes of our corpus. For the HMM method, we use 3 state HMM, each state of a class is described by 12 Gaussian models.

HMM seems to give best results for SNR$\geq$+10 dB and we are working to improve these first results.

Signal to Noise Ratio	0 dB	+10 dB	+20 dB	+40 dB	$\geq$+50 dB
GMM	23.6	15.4	16.5	10.2	3.2
HMM	29.7	12.6	10.8	9	2

Table 3. Classification Error Rate (%) using 12 Gaussian models.

7 Conclusions and Perspectives

In this paper we have presented a sound processing system designed to work in the framework of a medical remote monitoring application. An adapted French speech corpus with distress and normal sentences has been recorded and used for evaluation purpose. The system detects sound events, identifies the type of sounds among speech and life sounds. This step may be used under realistic condition with moderate noise: +10 dB SNR. In the case of speech, a French

speech recognizer is initiated, then distress key words or call for help may be extracted from the recognized sentences. Therefore, this system is able to extract new additional information from speech and sounds: the behaviour of the patient is best known and that may be very useful for the medical monitoring system to make a decision in a distress case.

Acknowledgements: 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.

References

1. European Commission: Europe's response to World Ageing. Promoting economic and social progress in an ageing world. A contribution of the European Commission to the Second World Assembly on Ageing. 18 March (2002)
2. V. Rialle, J.B. Lamy, N. Noury, L. Bajolle: Remote monitoring of patients at home: A Software Agent approach. Computer Methods and Programs in Biomedicine. **72**, 3 (2003) 257–268
3. G.Virone, N.Noury and J.Demongeot: A System for Automatic Measurement of Circadian Activity in Telemedicine. Proc. IEEE Transactions on Biomedical Engineering, **49** (2002) 1463–1469
4. Real World Computing Partnership: CD - Sound Scene Database in Real Acoustical Environments (1998-2001)
5. D. Reynolds: Speaker Identification and Verification using Gaussian Mixture Speaker Models. Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland (1994) 27–30
6. J. Pinquier, C. Senac and R. Andre-Obrecht: Speech and music classification in audio documents. proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, **4** (2002) 4164
7. L. Lu, H.J. Zhang and H. Jiang: Content analysis for audio classification using feature extraction matrix. proc. IEEE Transaction on Speech and Audio Processing, **10**, 7 (2002) 504–516
8. T. Yamada and N. Watanabe: Voice Activity Detection using non-Speech Models and HMM Composition. Workshop on Hands-free Speech Communication, Tokyo, Japan (2001)
9. M. Akbar et al.: Parole et traduction automatique : le module de reconnaissance RAPHAEL. Proc. COLING-ACL'98, Montréal, Quebec, **2** (1998) 36–40
10. D. Vaufreydaz et al.: Internet Documents - a Rich Source for Spoken Language Modeling. Proc. IEEE Workshop ASRU'99, Keystone-Colorado, USA (1999) 277–281
11. J.L. Gauvain, L.F. Lamel, M. Eskenazi: Design considerations and text selection for BREF, a large French read-speech corpus Proc. ICSLP'90, Kobe, Japan (1990) 1097–1100
12. D. Vaufreydaz et al.: A New Methodology for Speech Corpora Definition from Internet Documents Proc. LREC'2000, 2nd Int. Conf. on Language Ressources and Evaluation, Athens, Greece (2000) 423–426