Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment

Michel Vacher, Anthony Fleury, J.-F. Serignat, Norbert Noury, Hubert Glasson

Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment

Michel Vacher[1], *Anthony Fleury*[2], *Jean-François Serignat*[1], *Norbert Noury*[2], *Hubert Glasson*[1]

[1]Laboratory LIG, UMR CNRS/INPG/UJF 5217, Team GETALP, Grenoble, France.
[2]Laboratory TIMC-IMAG, UMR CNRS/UJF 5525, Team AFIRM, Grenoble, France.
{firstname.lastname}@imag.fr

Abstract

Improvements in medicine increase life expectancy and the number of elderly persons, but the institutions able to welcome them are not sufficient. A lot of projects work on ways allowing elderly persons to stay at home. This article describes the implementation of a sound classification and speech recognition system equipping a real flat. This system has been evaluated in uncontrolled conditions for distinguishing normal sentences from distress ones; these sentences are uttered by heterogeneous speakers. The detected signals are first classified as sound and speech. The sounds are clustered in eight classes (object fall, doors clap, phone ringing, steps, dishes, doors lock, screams and glass breaking). As for speech signals, an input utterance (in French) is recognized and a subsequent process classifies it in normal or distress, by analysing the presence of distress key words. In the same way, some sound classes are related to a possible distress situation. An experimental protocol was defined and tested in real conditions inside the flat. Finally, we discuss the results of this experiment, where ten subjects were involved.

Index Terms: ASR, Linear-Frequencies Cepstral Coefficients (LFCCs), Noisy Conditions, Sound Classification.

1. Introduction

The constant growing of life expectancy in the world yields a lack of places and workers in institutions able to take care of elderly people. Researcher teams all over the world try to tackle this issue by working on ways to maintain elderly people in their own home as long as possible. Geriatrics is thus in great need for sensors in order to assess the evolution of the person in her environment and to detect early the appropriate moment for admitting that person in an institution.

Abnormal situations in the behaviour of the person should be detected by smart sensors and "smart houses" [1]. Smart houses have demonstrated that measuring the activity of a person at home can be relevant [2], and also that this monitoring is useful for people with cognitive impairments [3]. A few systems have sound recognition capabilities [4][5].

A fully functional flat has been fitted with numerous sensors, chosen for classifying the different activities of a person's everyday life. This flat, shown in Fig. 1, is fitted with: -Infrared presence sensors (IPR) for locating the subject, -large angle webcams to save, analyse and time-stamp every action made by the person, -a weather station that give an information on temperature and hygrometry, -open/close detectors placed on communication doors, fridge... -an embedded kinematic sensor, -and, finally, eight microphones that cover the entire flat; these microphones are in the focus of this paper.

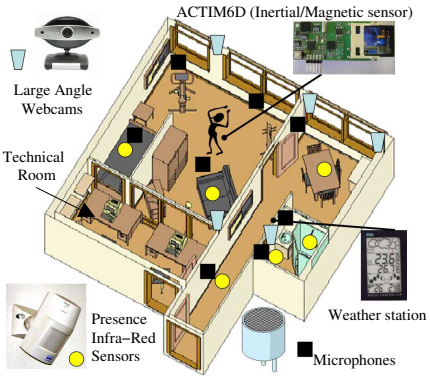

Figure 1: A Smart Home environment

Data from these sensors are acquired and processed on four computers disposed in the technical room. These data are used inputs to off-line data fusion algorithms, for detecting and classifying daily activities. The features of the sensors (i.e., sensitivity and specificity) are important constraints for these algorithms. This paper presents the sound and speech detection and classification system, as well as the results of an experiment made in the flat, in order to assess its performances out of "laboratory conditions" (results for these conditions are given in section 2). The sentences uttered by the subject give valuable information on her or his usual activities, or on a distress situation.

2. Architecture of the Sound Analysis System

2.1. Overview of the System

The general organization of the sound analysis system is shown in Figure 2. Each microphone is connected to an analog channel of the acquisition board (National Instrument PCI-6034E). The global system is composed of the *analysis system* and the *autonomous speech recognizer*, running in real time as independent applications on the same computer, under GNU/Linux. These two applications are synchronized through a file exchange protocol. The analysis system is set up through a dedicated module, while other modules run as independent threads and are synchronized by a scheduler.

The "Acquisition and First Analysis" module is in charge of data acquisition on the 8 analog channels simultaneously, at a sampling rate of 16 kHz. Noise level is evaluated by this mod-

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

Table 1: Segmentation Error Rate between Speech and Sound, 16LFCC, GMM, 24 Gaussian models, Sound and speech corpora, 4,631 tests per SNR

Signal to Noise Ratio	0 dB	+10 dB	+20 dB	+40 dB
GMM, 24 LFCC	36.6 %	21.3 %	13 %	9.3 %
HMM, 24 LFCC	29.8 %	16.3 %	6.6 %	5.9 %

Table 2: Classification Error Rate between 8 Sound classes, 24LFCC, 12 Gaussian models, Life sound corpus, 2,646 tests per SNR

ule, in order to allow the Signal to Noise Ratio analysis. The SNR of each signal event is very important for the data fusion system in order to estimate the reliability of the outputs provided by the analysis modules. The "Detection" module is in charge of signal extraction, also detecting the beginning and the end of the speech, or of the everyday life sound. This module was evaluated through Receiver Operating Curves giving the *missed detection rate* as a function of the *false detection rate*. The Equal error Rate is 0 % for a SNR above +10 dB, and 6.5 % at a SNR of 0 dB.

2.2. Corpora and Sound Analysis

In order to train and validate the system, two adapted corpora were recorded: the *normal/distress speech corpus* in French and the *everyday life sound corpus*. They are both needed for training the "Segmentation" module, the sound corpus for classification training and the speech corpus for speech recognition evaluation. The *normal/distress speech corpus* was recorded at CLIPS laboratory by 21 speakers (11 men and 10 women) between 20 and 65 years old. This corpus has a total duration of 38 minutes and is constituted by 2,646 audio files in wave format, each file containing one utterance. The *everyday life sound corpus* contains 8 sound classes of two types: normal sounds, related to usual activities of the patient (door clapping, phone ringing, step sounds, dishes sounds, door lock), and abnormal sounds, related to distress situations (breaking glasses, falling objects, screams). This corpus contains records made at LIG laboratory (61% of the files) using eW500 Sennheiser microphones. This corpus also contains files extracted from previous recording sessions, performed at the time of former studies in the CLIPS laboratory; finally, the corpus also contains some files obtained from the Web. The corpus is constituted of 1,985 audio files and its total duration is of 35 min and 38 s, each file containing one sound.

Then, the detected signal is transferred by the "Segmentation" module to the "Speech Recognition System" in case of speech, or to the "Sound Classifier" in case of everyday life sounds. Signal segmentation is achieved through a Gaussian Mixture Model (GMM) classifier, trained with the everyday life sound corpus, and the normal/distress speech corpus recorded in the LIG laboratory. Acoustical features are Linear-Frequency Cepstral Coefficients (LFCC) with 16 filter banks; the classifier uses 24 Gaussian models. These features are used because life sounds are better discriminated from speech with constant bandwidth filters, than with Mel-Frequency Cepstral Coefficients (MFCC) , on a logarithmic Mel scale. Frame width is of 16 ms, with an overlap of 50 %.

In order to validate the segmentation and classifications stages, the sound and speech corpora were mixed with noise recorded in the smart home at 4 different Signal to Noise Ratios (SNR=0 dB, +10 dB, +20 dB, +40 dB), whereas training was achieved with pure sounds. Segmentation performances are evaluated through the segmentation error rate (SER), which represents the ratio between the misclassified files and the total number of files to be classified. Results are presented in Table 1. SER remains quite constant with a 5 % value above +10 dB.

Everyday life sounds are classified with a GMM or Hidden Markov Model (HMM) classifier; the classifier is chosen before the beginning of the experiment. These models were trained with the corpus containing the eight classes of everyday life sounds, using LFCC features (24 filter banks) and 12 Gaussian models. Classification performances are evaluated through the classification error rate (CER). Results are presented in Table 2. These results are highly influenced by the SNR.

2.3. Speech analysis

The autonomous speech recognizer RAPHAEL [6] is running as an independent application and analyzes the speech events resulting from the segmentation module, through a file exchange protocol. As soon as an input file is analyzed, it is deleted, and the 5 best hypotheses are stored in a hypotheses file. This event allows the scheduler to send an other file to the recognizer. The language model of this system is a medium vocabulary statistical system (9,958 words in French). This model is obtained by extraction of textual information from the Internet and from the French journal "Le Monde". Then, it is optimized using textual information of a current conversation corpus in French. This conversation corpus contains the sentences in the *normal/distress speech corpus*, along with 253 sentences currently uttered during a telephone conversation: "Allo oui", "A demain", "J'ai bu ma tisane", "Au revoir"... The *normal/distress speech corpus* is composed of 126 sentences in French: 66 are typical for a normal situation for the patient: "Bonjour" (Hello), "Où est le sel" (Where is the salt)... , 60 are typical for a distress situation: "Aouh", "Aïe", Au secours" (Help), "Un médecin vite" (Call a doctor hurry) along with syntactically incorrect French expressions like "Ça va pas bien" (I don't feel good)... Our main requirement is the correct detection of a possible distress situation through keyword detection, without understanding the patient's conversation. For speech recognition, the training of the acoustic models was made with large corpora in order to ensure a good speaker independence. These corpora were recorded by 300 French speakers in the CLIPS (BRAF100) and LIMSI laboratories (BREF80 and BREF120) [7].

3. Speech Recognition Evaluation

The speech recognition system has been evaluated using the sentences from all the speakers in the *normal/distress speech* corpus (2,646 tests); see Table 3. In 0.5 % of the cases, for normal sentences, an unexpected distress keyword is detected by the system thus leading to a *False Alarm Sentence*. In 22 % of the cases, for distress sentences, the distress keyword is not recognized (missed): this leads to a *Missed Alarm Sentence*. This often occurs with isolated words like "Aouh", "Aïe" (Ouch) or "SOS", or in sentences like "Ça va pas bien" recognized as "Ça

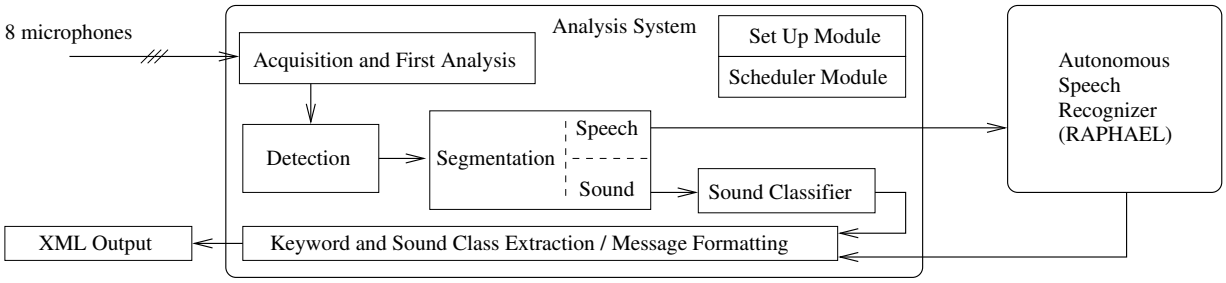

Figure 2: Sound Analysis System

Corpus Part	Keyword Detection Error	Recognition Error
(1) Normal	False Alarm: 6	0.5 %
(2) Distress	Missed Alarm: 282	22 %

Table 3: Speech Recognition Error Rate, Normal/distress speech corpus, 2,646 tests

va bien", where the negation mark "pas" is missed. It is more difficult to recognize isolated words, because of the great number of phonetical variants and of the ineffectiveness of the language model: for example "Aouh" (Cry in pain, distress expression) has the same probability as "Ah oui" (Normal expression). Thus, the global Distress Keyword Recognition Rate is 11 %.

4. Experiments and Results

4.1. Experimental Protocol

To validate the system in uncontrolled conditions, we designed a scenario where every subject has to utter 45 sentences (20 distress sentences, 10 normal sentences and 3 phone conversations of 5 sentences each). For this experiment, 10 subjects volunteered, 3 women and 7 men (age: 37.2 ± 14 $years$, weight: 69 ± 12 kgs, height: 1.72 ± 0.08 m). The number of sounds collected in this experiment was $3,164$ ($2,019$ of them were not segmented because their SNR was less that 5 dB), with an SNR of 12.65 ± 5.6 dB. After classification, we kept $1,008$ sounds with a mean SNR of 14.4 ± 6.5 dB.

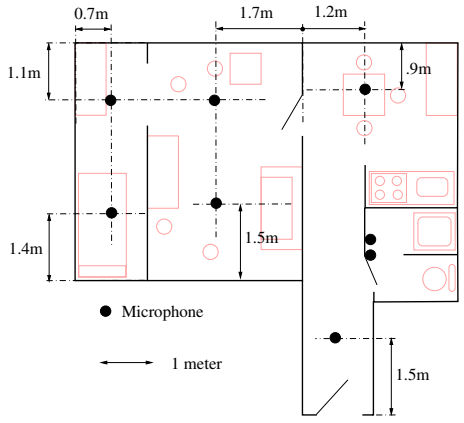

Figure 3: Microphone setting in the flat

The experiment took place during daytime – hence we did not control the environmental conditions of the experimental session (such as noises occurring in the hall). The sentences were uttered in the flat, with the subject sat down or stood up. The subjects were situated between 1 and 10 meters away from the microphones and have no instructions concerning their orientation with respect to the microphones (They could choose to turn their back to the microphone direction). Microphones are set on the ceiling and directed vertically to the floor as shown on Fig. 3. The phone was placed on a table in the living room.

The protocol was quite simple. The subjects were asked to first go in the flat and close the door, and then to act a little scenario (close the toilet door, make a noise with a cup and a spoon, let a box fall on the floor and scream "Aïe"). This whole scenario was repeated 3 times for each subject. Then, the subjects have first to go to the living room and close the door and then to go to the bed room and read the first half of one of the five successions of sentences, out of 10 normal and 20 distress sentences. Afterwards, they had to go to the living room and utter the second half of the set of sentences. Each subject was finally called 3 times and had to answer the phone and read the phone conversation given (5 sentences each). To realize these successions of sentences, we chose 30 typical sentences and 5 phone conversations, and then we scrambled the sentences five times, and we randomly chose 3 of the 5 conversations.

4.2. Data Processing

Every audio signal is recorded by the application, analyzed on the fly and finally stored on the hard disk drive of a computer. For each detected signal, it is first segmented (as sound or speech), and then classified (as one of the eight classes), or, in case of speech, the 5 more probable sentences are stored. For each sound, a XML file is generated, containing all the important information. Afterwards, distress keywords are extracted from the complete sentences, and these collected data are processed using Matlab™. They are classified using the two methods.

The first one (named M1) selects, out of several simultaneous signals, the one that has the highest SNR. After this selection, two classification methods are applied. The first one, named C1, considers only the most probable sentence acquired via the selected microphone and extracts the distress keyword from it. The second one, named C2, takes the three most probable sentences, extracts the distress keywords from them and allots a weight of 1, 0.75 and 0.5, respectively, to the decision from each of the three sentences (for instance, if we have a normal sentence as the first one, and two distress sentences after, we will classify it as distress – because of the score of $0.75+0.5$ for distress and 1 for normal).

The second sound classification method (named M2) will take the sound with the best SNR (named x), and keep all the

microphones that acquire sounds having an SNR greater that $0.8 * x$. We will make our decision with a vote between these different decisions, with two rules : (1) if a distress speech is detected, we will keep this decision and (2) in case of equality with another decision , different from distress speech, we keep the decision of the microphone that has the highest SNR. This classification method is referred to as C3.

	S1	S2	C1	C2	C3
Global	8.3 %	6 %	33.4 %	34.5 %	30.5 %
Normal	9.6 %	6.9 %	10.4 %	10 %	9.6 %
Distress	7 %	4.3 %	60.1 %	63.1 %	54.8 %

Table 4: Segmentation/Classification error rate for the distress/normal sentence recognition.

4.3. Sound and Speech Segmentation

The two first stages of the algorithm are the detection of the sound, and its segmentation (to know if it is a sound or a speech sample). The adaptive threshold allows the system to miss no event, this is the reason why we have 0 % error on the detection part. Since the mean SNR of the signals during the experimental session is $14.4 \pm 6.5\ dB$, we have relatively acceptable rates with about 8.3 % of segmentation error in the cases C1 and C2, and 6 % with C3. Table 4 shows in detail the segmentation performances of these algorithms. S1 refers to the segmentation made with only one microphone (method M1) and S2 to the segmentation made with a fusion between the different microphones that have a sufficient SNR (method M2). In laboratory conditions with an equivalent SNR, the segmentation error rate is between 3.8 % and 5.1 % (See Table 1). This underlines the difficulty of working in real conditions. The sounds are far from being perfect and the segmentation gives us, in the first stage, an error greater than obtained in laboratory conditions.

4.4. Normal/distress Sentences Recognition

During the experimental sessions, 446 sentences were uttered by the subjects, out of which 206 were distress ones. Table 4 shows the results for the three different classification processes (C1, C2, C3, see section 4.2). It is worth noticing that experimental recording conditions are critical. For example, in the living and bed rooms, reverberation between windows (70 % of wall area) and technical room glasses (100 %) is very high; hence, it was necessary to partially close the curtains to reduce its effect. These results are shown as a function of the speaker on Fig. 4. For 3 speakers, the missed alarm rate is more than 70 %; on the contrary, 3 of them are under 40 %. This can be caused by a different pronunciation due to a regional accent. We can conclude that we have to improve the acoustic models and to add more phonetical variants to the phonetic dictionary, but these results may also be explained by the lower SNR (14 dB), compared to the studio conditions for corpus recording (more than 30 dB). Results in Table 4 demonstrate that the classification of normal sentences is better that for distress ones. The comparison between the three algorithms demonstrates that the third is the best one. It improves the missed alarm rate without changing significantly the false alarm rate.

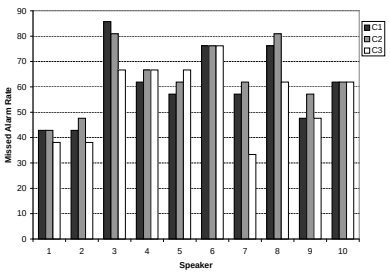

Figure 4: Distress sentences: missed alarm rate per speaker

5. Conclusion and Perspectives

This paper has presented the results of an experimental protocol where French speakers had to utter normal and distress sentences in a real flat, in uncontrolled conditions. The sentences were uttered in the flat; no conditions were imposed to the subjects who were located between 1 and 10 meters away from the microphones and not necessary in front of them. The results show that the segmentation and the detection were acceptable, and the false alarm rate was not too high (10 % with the best classification algorithm). But it also showed us that we have to work to improve the missed alarm rate. The results obtained in laboratory are far from those obtained in real conditions. The different classification processes and the improvements brought by taking into account the different significant microphones allow to reduce the segmentation error and the false alarm rates, but as far as the missed alarm rate is concerned, the results are not satisfactory for using the system in real conditions with these models.

For the largest part of the sentences, errors may be caused by the noise present in the flat during the recording sessions, and not by the speaker dependency. The collected sounds will allow us to improve the acoustic models for the silent HMM state. Another part of our current work is to validate noise suppression techniques and to work on a better language model for French.

6. References

[1] C. N. Scanaill, S. Carew, P. Barralon, N. Noury, D. Lyons, and G. M. Lyons, "A Review of Approaches to Mobility Telemonitoring of the Elderly in their Living Environment," *Annals of Biomedical Engineering*, vol. 34, pp.547–563, Apr. 2006.

[2] G. LeBellego, N. Noury, G. Virone, M. Mousseau, and J. Demongeot, "A Model for the Measurement of Patient Activity in a Hospital Suite," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 10 (1), pp. 92–99, Jan. 2006.

[3] B. Bouchard, A. Bouzouane, and S. Giroux, "A Smart Home Agent for Plan Recognition of Cognitivly-Impaired Patients," *Journal of Computers,* vol. 1 (5), pp. 53–62, Aug. 2006.

[4] M. Stäger, P. Lukowicz, G. Tröster, "Power and accuracy tradeoffs in sound-based context recognition systems," *Pervasive and Mobile Computing,* vol. 3 (3), pp. 300–327, 2007.

[5] J. C. Wang, H. P. Lee, J. F. Wang, and C. B. Lin, "Robust Environmental Sound Recognition for Home Automation", *Automation Science and Engineering, IEEE Transactions on*, vol. 5 (1), pp. 25–31, Jan. 2008.

[6] M. Akbar, and J. Caelen, "Parole et Traduction Automatique : le Module de Reconnaissance RAPHAEL." in *Proc. COLING-ACL'98,* Montréal, Quebec, pp. 36–40, Aug. 10-14. 1998.

[7] J.L. Gauvain, L.F. Lamel, M. Eskenazi, "Design Considerations and Text Selection for BREF, a large French read-speech corpus," in *Proc. ICSLP'90,* Kobe, Japan, pp. 1097–1100, Nov. 18-22. 1990.