Everyday Life Sounds and Speech Analysis for a Medical Telemonitoring System

Eric CASTELLI, Dan ISTRATE

Laboratory CLIPS-IMAG, Grenoble Cedex 9, France
Eric.Castelli@imag.fr, Dan.Istrate@imag.fr

Abstract

In order to improve patients’ life conditions and to reduce the costs of the long hospitalization, the medicine is more and more interested in the telemonitoring techniques. These will allow the old people or the high risk patients to stay at home, and to benefit from a remotely and automated medical supervision. We develop in collaboration with TIMC-IMAG laboratory, a system of telemonitoring in a habitat equipped with physiological sensors, position encoders of the person, and microphones. The originality of our approach consists in replacing the video camera monitoring, hardly accepted by the patients by microphones recording the sounds (speech or noises) in the apartment. The microphones carry out a multichannel sound acquisition system which, thanks to the sound information coupled with physical information, will enable us to identify a situation of distress. We describe the practical solutions chosen for the acquisition system and the recorded corpus of situations.

1. Introduction

Telemedecine consists in associating electronic techniques of monitoring with computers “intelligence” and with the speed of telecommunications (established either through network or radio connections). Telemedicine is announced as a significant reform of the medical care because it allows to improve the response time of the specialists which could be informed about a medical emergency and react as soon as the first symptoms appear without waste of time. The telemedicine also allows a significant reduction of the costs of public health by avoiding the hospitalization of the patients for long periods of time.

The system we work on is designed for the surveillance of elderly persons who can benefit from telemedicine. Its main goal is to detect serious accidents as falls or faintness (which can be characterized by a long idle period of the signals) at any place in the apartment [5]. This technique allows a medical center in charge to analyze the information gathered by the telemonitoring system and to intervene if needed [2]. We noted that the elderly had difficulties in accepting a monitoring by video camera, because they consider that their constant recording is a violation of their privacy. Thus, the originality of our approach consists in replacing the video camera by a system of multichannel sound acquisition charged to analyze in real time the sound environment of the apartment in order to detect abnormal noises (falls of objects or of the patient), calls for help or moans which could characterize a situation of distress in the habitat.

2. Presentation of the Telemonitoring System

The habitat we used for experiments is a 30 m² apartment situated in the TIMC laboratory buildings, at the Michalon hospital of Grenoble.

The patient carries a set of sensors which give information about his activity: vertical position (standing) or horizontal (lying) and sudden move (falling). The localization sensors with infra-red radiation are installed in each part of the apartment in order to establish where the person is at any moment. These sensors communicate with the acquisition system by radio waves and by bus CAN. The control of the activities sensors is ensured by a PC using a monitoring software programmed in JAVA.

Fig. 1. Position of the sensors inside the apartment

The sound sensors are represented by 8 microphones, their position is given in Figure 1. An acoustic antenna composed of 4 microphones allows to monitor both the living room and the bedroom, aiming to localize the patient inside the two rooms and also to analyze the sounds. As the other rooms are much smaller, only one microphone per room is sufficient. The microphones used are omni-directional, condenser type, of small size and low cost. A signal conditioning card, consisting of an amplifier and an anti-aliasing filter, is associated to each microphone. The acquisition system consists of a multi-channels acquisition card PCI 6034E of National Instruments, installed inside a second computer. The acquisition is made at a sampling...
rate of 16 KHz, a frequency usually used in speech applications. We programmed the entire software which controls the acquisition under LabWindows/CVI of National Instruments. After digitalization the sound data is saved in real time on the hard disk of the host PC in a temporary file. The two computers are connected between them by a conventional IP network. The general outline of the acquisition system is presented in Figure 2.

![Diagram of telemonitoring system](image)

**Fig. 2. Diagram of telemonitoring system**

3. **Acoustic Localization of the Person**

The exact geometrical location of the person in the apartment is important, as well to establish if a suspect noise which is registered indicates a distress or not, and also to indicate the location of the person to the emergency services in the event of an alarm. To locate the person, we analyze the information from the infra-red sensors and sound sensors. The information obtained through analysis of the data from the two sources is necessary to guarantee a better characterization of the patient’s situation in order to avoid a false alarm.

The 4 microphones M1 to M4, located in the small parts of the apartment (hall, toilet, shower-room and kitchen) are sufficient to evaluate the location of the noise or speech source, by a comparison of the noise level of each microphone. On the contrary, for the larger spaces (the living room and the bedroom), we use a mural acoustic antenna consisting of 4 microphones forming a 320x170mm rectangle. After evaluation of several shapes (square, T, parallelogram, cross), we chose for our acoustic antenna the shape which allows to identify most exactly the spatial position of the sound source for the needs of our application [7]. Using our antenna and a 16KHz sampling rate, we obtain a 0.2m precision for each of the space coordinates x, y, z sufficient for our application.

4. **Corpus**

In order to validate our analysis strategies by data fusion, we recorded a multichannel corpus, presenting a simultaneous and synchronised recording of the 8 signals (WAV type files), geometrical coordinates of position, position information from the infrared sensors, and data from the sensors carried by the patient. Files are recorded according to the SAM norm used for speech corpus.

The corpus is made up of a set of 20 scripts reproducing a string of events, either voluntary actions of the patient inside the apartment, or unexpected events which could characterize an abnormal or distress situation. The voluntary actions are moves of the patient inside the apartment and everyday gestures presenting natural speech signals, completed by other usual sounds (radio, telephone ring, dish noise, door noise, etc...). The unexpected events are abnormal noises (falling, glass breaking) and abnormal speech signals such as: moans, cries and helps. Every script is registered 4 times. For this first study, we limit the number of subjects to 5 persons. We present one of the scripts written in XML:

```
<Script no. A0>
<description> Walking in the house.Normal situation (no alarm) </description>
<time> 0 </time> <Position> Entry </Position> <Action> Person speaking </Action>
<time> 4 </time> <Position> Hall </Position> <Action> Person speaking </Action>
<time> 10 </time> <Position> Toilet </Position> <Action> Person speaking </Action>
```

---

**Note:** The XML script shown is a simplified representation and does not include all details provided in the document. The actual XML script would provide more detailed information about the actions and events.
The first step before the recording of the scripts of the corpus is the calibration of the sound recording system. A 1KHz rectangular signal is reproduced by a speaker situated to 1 meter from every microphone. The gains of every sound channel are then adjusted to obtain the same amplitude of the signal.

After recording, the corpus is analysed: every event of the script (nature and time) is characterised, the speech is labelled following the usual procedures used for speech corpus.

Figure 3 shows three microphone signals with their energy and their corresponding infra-red sensors. Figure 4 shows a diagram of the apartment on which the lamp type indicators are superimposed, giving the position of the person as localized by the infra-red sensors and the localisation in three dimensions of the person in the living room.

The speech signals energy was calculated using a 1 second average in order to obtain the location of the patient inside the room. These first results show that this location information, given by the microphone, is coherent with the information given by the infra-red sensors. The signal energy recorded in the hall (figure 3) has several maximums: at time 8s and 19s recorded the slapping sound of the entrance door and respectively of the toilet door. We can note that at the same moments the infra-red sensors was detected the presence of the person in the hall. However this signal present other peaks which correspond to the speech recorded inside the toilets and shower room. The energies of the toilet and the shower signals have the same peaks with different values, of course. It should be mentioned that between the rooms there is no door and in this case the microphone placed in the shower also records the signal inside the toilets and vice-versa. The energy of the toilet signal has a peak at time t=13s of a value=300, which is smaller than the value of the peak of the energy of the shower signal in the same moment, which is 396 (we can locate the person by comparing the two levels). Thus, we have two kinds of additional and coherent information, we shall use in our analysis by fusion of data.

The corpus will be used in order to develop a module of sound signals characterisation. It has two goals: first, it has to establish if the audio signal is a speech signal or a noise, then it has to characterise the noises (everyday life noises, or abnormal noises, i.e. falling) and the speech sounds (normal, cry for help or moans). This module is about to be produced and validated.

To distinct between speech and noise we shall compare the speech signal which has periodical characteristics (such as the pitch), to the spectral characteristics (large spectrum and impulse characteristics) of the noises. The characterization of noises will be made by a noise recognition system, based on a classical HMM method [4]. Speech characterization (normal or stressed) will be inspired from speaker characterization methods [1], [6]. If sounds produced by the speaker do not match the normal speech, the system will consider that the speaker produces abnormal speech, such as: calls, cries or moans.

If speech is found to be normal, a recognition system of Word Spotting type will allow to recognize about 20 words,
chosen out of a specialized vocabulary of help calls (help, aie!, etc...), in order to facilitate to the emergency services the decision to intervene. However the recognition system will not do continuous speech recognition with large vocabulary, as we consider that the semantic content of the phrases uttered by the speaker in normal situations is part of private life information and should not be recorded.

5. Conclusions and Perspectives

We presented a system of telemonitoring for an apartment equipped with audio and infra-red sensors. This is meant to replace video cameras as the patients are not very comfortable with them. The hardware of the system was set up and was validated. We started to record a corpus of scenarios.

The first results enable us to locate the person in three dimensions in the living room or to indicate the room where the patient is located (following the energy of the speech signals). The current recordings also allowed us to validate the sensors and the audio acquisition system. Information from the audio sensors is mixed with the one from the infra-red sensors to ensure the best results.

We shall exploit the corpus of situations of everyday life and of distress situations, in order to continue our research. We shall study and validate a set of monitoring strategies, based on a analysis of the fusion of the data mixing the formation from the activity sensors with the one obtained by sound environment analysis.

This study is a part of the RESIDE-HIS project, financed by IMAG Institute in collaboration with TIMC laboratory.

6. References