Making emergency calls more accessible to older adults through a hands-free speech interface in the house
Michel Vacher, Frederic Aman, Solange Rossato, François Portet, B Lecouteux

To cite this version:
Michel Vacher, Frederic Aman, Solange Rossato, François Portet, B Lecouteux. Making emergency calls more accessible to older adults through a hands-free speech interface in the house. ACM Transactions on Accessible Computing, ACM New York, NY, USA 2019, 12 (2), pp.8:1-8:25. 10.1145/3310132. hal-02009828

HAL Id: hal-02009828
https://hal.archives-ouvertes.fr/hal-02009828
Submitted on 28 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Making emergency calls more accessible to older adults through a hands-free speech interface in the house

MICHEL VACHER, FRÉDÉRIC AMAN, SOLANGE ROSSATO, FRANÇOIS PORTET, BENJAMIN LECOUTEUX
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France

ABSTRACT

Wearable personable emergency response (PER) systems are the mainstream solution for allowing frail and isolated individuals to call for help in an emergency. However, these devices are not well adapted to all users and are often not worn all the time, meaning they are not available when needed. This paper presents a Voice User Interface system for emergency call recognition. The interface is designed to permit hands-free interaction using natural language. Crucially, this allows a call for help to be registered without necessitating physical proximity to the system. The system is based on an ASR engine and is tested on a corpus collected to simulate realistic situations. The corpus contains French speech from 4 older adults and 13 younger people wearing an old-age simulator to hamper their mobility, vision and hearing. On-line evaluation of the preliminary system showed an emergency call error rate of 27%. Subsequent off-line experimentation improved the results (call error rate 24%), demonstrating that emergency call recognition in the home is achievable. Another contribution of this work is the corpus, which is made available for research with the hope that it will facilitate related research and quicker development of robust methods for automatic emergency call recognition in the home.

KEYWORDS:

CCS Concepts: • Human-centered computing → Natural language interfaces; • Social and professional topics → People with disabilities; • Computing methodologies → Speech recognition;

Additional Key Words and Phrases: Specific voice recognition, assistive technology, emergency call, Ambient Assisted Living.

This study was funded by the National Agency for Research under the project CIRDO “Un compagnon Intelligent Réagissant au Doigt et à l’Oeil” (ANR-2010-TECS-012).

Author’s addresses: Laboratoire d’Informatique de Grenoble - 700 avenue Centrale - Bâtiment IMAG - Domaine Universitaire - 38401 St Martin d’Hères - France.
Current author’s addresses: Michel.Vacher@imag.fr, Frederic.Aman@imag.fr, Solange.Rossato@imag.fr, Francois.Portet@imag.fr, Benjamin.Lecouteux@imag.fr, Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France.
1 Introduction

Life expectancy has increased in all countries of the European Union over the last decade. In early 2013, only 9% of people in France were at least 75 years old. By 2040, according to INSEE\(^1\) institute, the number of dependent older adults will increase by 50% (Duée and Rebillard, 2006). Dependency refers to the restriction of the activities of daily living, and the need for help or assistance of someone in performing regular elementary activities due to the alteration of physical, sensory and cognitive functions (Charpin and Tlili, 2011). Falls are the leading cause of dependence in older adults. Indeed, according to (Institut National pour la Prévention et l’Education à la Santé, 2006), about 25% of French people aged between 65 to 75 years old had fallen in the 12 months before their study and these falls represent about 80% of daily accidents. Some of these falls have serious consequences primarily when the person stays on the floor for one hour or more. Those, called “long-lies”, often result in higher mortality rates and hospital admissions (Simpson et al., 2014). The \textit{de facto} course of action after a fall is to transfer the individual to a nursing home. Apart from falls, other main risks are bruising or crushing, cuts, wounds resulting from piercing and straining or twisting a part of the body (RoSPA, 2016). However, a survey shows that 80% of people above 65 years old would prefer to stay living at home and not lose their autonomy (CSA, 2003) while the study of (Anderson, 2010) reports that

\(^1\)Institut National de la Statistique et des Études Économiques (French National Institute of Statistics and Economic Studies)
nearly 90 percent of USA citizens over age 65 want to stay in their residence for as long as possible and 80% believe their current residence is where they will always live. This is corroborated by a new study (CSA, 2010) showing that 87% of family members and relatives plan to ensure that the person continues to live at home by visiting as often as possible, and enlisting professional in-home help. However, to make this possible, falls must be detected early and solutions must be found to let older adults access help as quickly as possible in case of emergency (Thilo et al., 2016).

To address this need, personal emergency response (PER) systems were originally developed. These systems consist of a wearable button attached to the neck or the wrist that must be constantly worn. In case of emergency, the wearer pushes the button and is connected to a professional operator at any time of day or night (Mann et al., 2005; Hessels et al., 2011). While PER systems can reduce anxiety and health care costs (Mann et al., 2005), some older adults are reluctant to adopt them (Heinbüchner et al., 2010; Hessels et al., 2011; Nyman and Victor, 2014). Despite PER systems being widely used, research studies have described that many users acquired the alarm, but hardly ever wear or activate it even in true situations of emergency (Heinbüchner et al., 2010). This seems to be related to a lack of ability to use the alarm but also because of a large set of circumstances such as finding other solutions for being safe, not wanting to bother or be bothered, finding the PERS stigmatizing, forgetting to wear it, not remembering that one is wearing it, or being unable to let helpers inside (Stokke, 2016).

To overcome these challenges, better ways to provide emergency response functionality without hampering daily living while respecting privacy are needed. With advances in Information and Communication Technology (ICT), a new generation of assistive technology is being developed. These assistive technologies can be embedded into smart devices or environments such as smart phones, robots, or smart homes (Hessels et al., 2011) to provide natural communication capabilities (Portet et al., 2015). In particular, Voice-User Interfaces (VUIs) may be more appropriate than “push-button” systems for making emergency calls for help because natural language interaction is intuitive (Vacher et al., 2015a; Young et al., 2016) and some people naturally call for help when stuck on the floor after a fall or express vocal cues during the event of a fall (Bobillier-Chaumon et al., 2016). Indeed, VUIs are appropriate for people with reduced mobility or visual impairment (Vacher et al., 2015a). Moreover, with a hands-free VUI a device does not need to be constantly worn on the neck or wrist and would thus be available at any time. Furthermore, since speech interaction systems are not visible and are of common usage, they should not raise stigmatisation issues.

In this paper, we propose to study a VUI system for emergency call recognition in the home called CirdoX as part of the Cirdo project. Our aim is to include a dedicated system including Automatic Speech Recognition (ASR) into the living environment of dependant and isolated older adults. This autonomous system should recognise phone calls to relatives or caregivers and detect emergency calls uttered by the person in order to send an alert to the appropriate service. This system would thus provide hands-free accessible help from anywhere in the home at anytime. An illustration of a use-case is given in Figure 1a. It shows a person has fallen to the floor and cannot access the alarm switch because it is on the cupboard. By contrast, Figure 1b shows the same situation, but the call for help “e-lio help me” is recognised by the CirdoX system and emergency assistance can be provided. e-lio is the specialized device making emergency calls which is used preferentially in the framework of the Cirdo project.

http://www.technosens.fr/
1. INTRODUCTION

M. Vacher et al.

Remote e-li-o

(a) The person can't warn emergency services: she can't access the remote.

(b) CirdoX recognizes the call for help, and e-lio can warn emergency services.

Figure 1: Warning emergency services in the usual case with a remote and using a system based on speech technologies.

A few existing systems have been designed to provide assistance in the home through voice interfaces such as the HELPER system (Young et al., 2016) and Sweet-Home (Vacher et al., 2015a). However, these systems face several challenges. First, the conditions for acoustic processing in these contexts are difficult. For example, ASR performances degrade for older adults. To assist this problem, it is necessary to adapt the ASR to the particular characteristics of older adults' voices. Moreover, the uttered sentences must be automatically extracted from the signal of the microphone before speech recognition and call detection. Without a wearable microphone, we cannot expect a favourable, near condition voice signal. On the contrary, in most cases in which the person will not be near the microphone, the system will operate in distant speech conditions. Moreover, speech will be uttered spontaneously with the speaker potentially under emotional stress. Second, the important challenge for VUI systems is to prove their robustness. Indeed, although some systems have been tested with real users, at the time of writing none are ready to be deployed to the market. Furthermore, testing such a system with real users in real situations is highly difficult given that reproducing artificially stressful situations is particularly risky with this kind of frail population and that opportunistic sensing is insufficiently accurate for gathering development data (Peetoom et al., 2015).

Given these two challenges, this paper will contribute to the problem by:

- providing a VUI that respects privacy and permits distant speech and hands-free interaction, all in a real-time manner (i.e., reacting in a time frame acceptable to the user);
- assessing the system in an experiment that simulates older adults' behaviour studied in the field in a smart environment.

This contribution is first positioned with respect to the state of the art which is described in Section 2, with a focus on ASR use in smart homes, the effect of ageing on voice and its consequences for ASR performance, as well as the existing corpora in that domain. The methods used to produce the VUI system are then detailed in Section 3. In particular this section describes a methodology that can be used to address, in the near future, the very challenging mixed condition of aged voices and emotion in real emergency situations. In section 4, an experiment involving aged and non-aged persons allowed us to evaluate the complete system in realistic conditions. This section shows that addressing such difficult situations is already challenging. Section 5
presents an off-line experiment in which performance of the system dramatically improved. A discussion in section 6 draws conclusions from this study and sketches future work.

2 State of the art

2.1 Health Smart Home

A Smart Home can be seen as an interconnected collection of smart objects that work together to provide services to inhabitants. The development of ICT has allowed for a profusion of new services in the home not seen since the introduction of electricity. However, at the time of the writing, the developments of this kind of home automation have led to only marginal improvements in quality of life. Moreover, durability, usability, security, privacy and trustworthiness have greatly impeded the development of smart homes.

Several studies aimed to identify the needs of older adults for a system to help them in their everyday lives (Koskela and Väänänen-Vainio-Mattila, 2004; Mäyrä et al., 2006; Demiris et al., 2004; Kang et al., 2006; Callejas and López-Cózar, 2009; Portet et al., 2013). The proposed systems must provide assistance in 3 main areas:

- monitoring of a person’s state of health and of the evolution of their loss of autonomy over time (Fleury et al., 2010; Chahuara et al., 2016);
- security by detecting emergency situations (e.g., fall (Bloch et al., 2011));
- comfort.

The need to communicate with relatives or the outside world was ignored by these studies, although outside communication is paramount for an isolated person at home in order to maintain social links (Rivière and Brugière, 2010). Further challenges were identified by (Chan et al., 2008): the reliability and efficiency of the sensory and data processing systems, the standardization of communication systems, the cost effectiveness, the social impact and ethical and legal issues.

2.2 ASR performances with aged voice and expressive speech

Some authors (Baba et al., 2004) have reported that classical ASR systems exhibit poor performances for older adults’ voice. The MATCH corpus is a collection of dialogues in which older and younger users interacted with a spoken dialogue system in English (Georgila et al., 2010). Analysis made thanks to the MATCH corpus confirmed that representative corpora of human-machine interactions need to contain a substantial sample of older adults.

A more general study of Gorham-Rowan and Laures-Gore (Gorham-Rowan and Laures-Gore, 2006) highlights the effects of ageing on speech utterances and the subsequent consequences for speech recognition. Experiments carried out in automatic speech recognition have shown a performance degradation for certain population groups such as children or older adults (Vipperla et al., 2008; Gerosa et al., 2009) and have shown the promise of adaptation to these target populations (Gerosa et al., 2009). Adapting speech recognition to the voice of older adults is still
an under-explored area of research. A very interesting study (Vipperla et al., 2008) analyzed recordings of speeches delivered in the Supreme Court of the United States over the course of a decade. These recordings are particularly interesting because they were used to study the evolution of recognition performance on the same person over 7-8 years. A limitation of this study, however, is that it relates to people with good diction and with experience in public speaking. Nevertheless, these studies show that the performance of recognition systems decreases steadily with age, and also that special adaptation to each speaker can produce scores similar to those obtained from the youngest speakers without adaptation.

Our purpose is to detect emergency calls like “I fell”, “I can’t stand up” or “Help me”. These sentences may be charged in emotion and prosodic modifications influencing voice quality (Vlasenko et al., 2012) which could be perceived by the interlocutor. (Scherer et al., 2003) compare effects of different emotions on prosodic parameters with respect to a neutral voice. Instead of detecting emergency states, a more immediate and reachable task is simply making ASR more robust in presence of emotion, but we have found only one study in this domain (Vlasenko et al., 2012). They observed a performance degradation with expressive speech when ASRs were trained with a neutral voice. These results were corroborated by Aman et al. (Aman et al., 2016a) who recorded a Voix-détresse (Distress Voice) corpus which is made of emergency sentences uttered in expressive and neutral manner in an elicitation protocol. Improved performance was achieved by using an adapted acoustic model.

2.3 ASR use in smart homes

Speech recording performed in smart homes is done at a distance (Woelfel and McDonough, 2009). ASR systems are challenged by these recording conditions and several studies were thus conducted (Ravanelli and Omologo, 2014; Brown et al., 2013; Principi et al., 2015) to take into account reverberation, noise, distance and orientation of the speaker with regard to the microphones or acoustic antenna.

Voice interaction of older adults in a home-setting has been studied in (Portet et al., 2013) in which older adults commanded a smart home voice control system simulated by a Wizard of Oz (WoZ). The participants expressed high interest in using voice commands for controlling the environment, but also expressed concern regarding security and the potential negative effects of such technology driving them towards a lazy life style. These studies showed that the audio channel shows promise towards providing security, comfort and health related assistance in smart homes (Vacher et al., 2011), but it remains relatively unexplored compared to classical and mobile physical interfaces such as switches, remote control and mobile phones.

Regarding the experimental settings, few experiments have actually been conducted with end users within realistic homes and even fewer within the participants’ own homes, with the notable exception of (Mäyrä et al., 2006). Furthermore, In home studies are typically run with fewer participants. For example, in (Callejas and López-Cózar, 2009), 200 Spanish people between 50 and 80 years old were questioned about different features of a smart home, but these people were not confronted with a prototype system, whereas in (Hamill et al., 2009) the developed Personal Emergency Response System (PERS) was tested with only 9 healthy young people. In an experiment conducted during the SWEET-HOME project, an experiment involving 6 older adults and 5 visually impaired people was conducted (Vacher et al., 2015a). The participants played out scenarios of everyday life and interopereated with a home automation system in a Living lab.
This experiment highlighted some weaknesses in ASR due to online analysis and distant speech conditions which must be addressed, as well as the need for better adaptation to the user and the environment. Despite these problems, the system was viewed favourably by participants with respect to diminishing most fears related to loss of autonomy.

Other recent studies addressed the issue of vocal interfaces for people with speech disorders. Indeed, the accuracy of general purpose speech recognizers is significantly affected in this case. For example, the case of patients with dysarthria was studied by (Casanueva et al., 2014), (Mustafa et al., 2015) and in the framework of the ALADIN project by (Gemmeke et al., 2013). This is a very specific case and, for this study, we restrict focus to older adults without voice disorders.

2.4 Existing suited corpora

The use of corpora is essential at all steps of the investigations and particularly during the model training and the evaluation. There are several French corpora that are commonly used in automatic speech recognition like BREF120 (Lamel et al., 1991), ESTER (Galliano et al., 2006) and Quaero (Névéol et al., 2014). They present the advantage of being recorded by a large amount of speakers. For example, BREF120 is made of more than 100 hours of recordings produced by 120 speakers. However these corpora are not applicable to our study because they do not include calls for help. Furthermore, they were all acquired from younger speakers and were either recorded in studio by speakers reading texts or extracted from broadcast news.

Corpora with more relevant content to our field of study include Anodin-détresse (AD80), recorded by 95 speakers between 18 and 94 years old, and of 2 hours and 18 minutes in duration. This corpus contains emergency sentences but was recorded in studio conditions (Vacher et al., 2006). Corpora with appropriate recording conditions (i.e., no reading and distant conditions in a smart home) include the SWEET-HOME (Vacher et al., 2014) corpus made of short sentences (vocal commands for home automation) uttered by persons interacting with a home automation system. Unfortunately, this corpus may only be used for system adaptation purposes because no emergency call were included in it. It is worth noting that a Canadian English corpus called CARES (Canadian Adult Regular and Emergency Speech) was collected in Toronto (Young and Mihailidis, 2013). This corpus is made of voices of 40 people from 23—91 years enacting emotional or stressed speech and emergency type dialogue. Although this corpus cannot be directly used in a French study, it does emphasize the difficulty of the task and the lack of resources in the community.

3 Method

Our aim is to develop a system able to automatically detect calls for help to relatives or carers through ASR. This service is targeted at older adults living alone at home. This section describes the fundamental constraints on the system and the chosen development approach adapted to the application.
3.1 From “in vitro” to “in situ” experiments

The aim of our work is to develop an emergency call recognition system to be implemented in older users’ homes. In this context, emergency calls are statements uttered spontaneously, and it is well known in the speech community that automatic recognition of spontaneous speech is very challenging. The ASR performance is hampered on one hand by the presence of emotion in the voice in emergency situations, and on the other hand by the acoustic characteristics of older adults’ voices. The focus of this study differs drastically from most studies in automatic speech recognition, which focus on young speakers reading aloud (i.e., non spontaneously). Old age is not a definite biological stage, it varies culturally and historically. There is no consensus, some studies distinguishes the young old (60 to 69), the middle old (70 to 79), and the very old (80+) (Forman et al., 1992). In our study, we consider that a person over 60 years old is aged. This value was chosen by the United Nations in the Charter of Rights of Older Persons (OHCHR, 2018).

![Figure 2: Experimental method for problem solving.](image-url)

Figure 2 shows the data for different experimental conditions for 2 age groups of speakers (young and older adults) :

1. emergency calls or sentences of everyday life read without elicitation of emotion (classic case of the speech corpora)
2. \textit{in situ} emergency by performing a voluntary fall on the carpet,
3. spontaneous emergency \textit{in situ}, i.e. corresponding to a real event that happened to the person.

These experimental conditions progress from typical ASR conditions to real conditions envisaged for the application.
3. METHOD

The green zone corresponding to case #1 in Figure 2 represents the acquired data with young people who performed a reading task in a studio or in a smart home. This is the furthest area from the orange area corresponding to case #3 and representing the goal, namely the record of spontaneous emergency call uttered by older adults at home which is the ultimate dimension of the project.

The data that are the closest to the origin of the graph (readings with young people) are the easiest to acquire. The further we move away from the origin, the older the users are, and the closer to the spontaneous situation, so data are more difficult to acquire and are thus in smaller quantities. Seniors can easily perform reading tasks, but it is difficult to play emergency scenarios including, for example, falls. Finally, the acquisition of real emergency data from older individuals requires continuous recording over a long term period in their homes. However, these situations are still infrequent, as testified by fall detection from wearable sensors (Bagalà et al., 2012; Kangas et al., 2012). That is why we restrict our study to #2 in Figure 2. We consider, in the future, getting data for #3 data from a commercially used product or through industrial collaboration backed by an ethical agreement. Therefore, the collection of data corresponding to area #3 was not conducted, this remains as future work.

In the online experiment (area #2 in Figure 2) presented in section 4, we observed, in situ, the performances of ASR systems thanks to emergency scenarios played by the subjects in a Living Lab.

3.2 Constraints and adaptation of the ASR

3.2.1 An ubiquitous system

The system must operate in the background and must be constantly and transparently attentive to the sound environment of the living space in order to recognize and filter any occurrence of speech, especially speech emitted during an emergency situation.

For this, the hands-free system uses sound activity detection to automatically select continuous sound segments containing relevant vocal signal without the need of physical intervention. That is to say, it begins voice recognition without the need to press an on/off button. To detect sound occurrences, an algorithm based on the signal energy is used (decomposition using a 3-depth wavelet tree) with an adaptive threshold (Istrate et al., 2006).

Also, in order to preserve the privacy of older adults, anything that is not an emergency call and call to caregivers should be ignored. For that, we used a language model with an ASR grammar reduced to automation orders and emergency calls. In addition, filtering is performed to retain only the emergency calls and not all spoken sentences.

As a result of the age of the potential users, their possible physical disabilities, and their possible unfamiliarity with new technologies, the system must be designed to operate fully independently in everyday life, without the need of any operation by the older adult or an operator other than during the startup, setup and initial configuration. Of course, the processing of data must run in pseudo real-time so as to trigger an alarm in as short a time as possible after detecting a help call.
Table 1: Examples of identified calls for help in French, where ⋆ denotes a sentence identified during the ethnological study by M.E. Bobillier Chaumont et al.

<table>
<thead>
<tr>
<th>Spontaneous Emergency Sentence</th>
<th>Formulated Emergency Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aïe aïe aïe ⋆</td>
<td>Appelle quelqu’un e-lio ⋆</td>
</tr>
<tr>
<td>Oh là ⋆</td>
<td>e-lio, appelle quelqu’un ⋆</td>
</tr>
<tr>
<td>Merde ⋆</td>
<td>e-lio tu peux appeler une ambulance</td>
</tr>
<tr>
<td>Je suis tombé ⋆</td>
<td>e-lio tu peux téléphoner au SAMU</td>
</tr>
<tr>
<td>Je peux pas me relever ⋆</td>
<td>e-lio, appelle du secours</td>
</tr>
<tr>
<td>Qu’est-ce qu’il m’arrive ⋆</td>
<td>e-lio appelle les secours</td>
</tr>
<tr>
<td>Aïe ! J’ai mal ⋆</td>
<td>e-lio appelle ma fille</td>
</tr>
<tr>
<td>Oh là ! Je saigne ! Je me suis blessé ⋆</td>
<td>e-lio appelle les secours</td>
</tr>
</tbody>
</table>

Table 2: English translation of the examples of calls given in Table 1

<table>
<thead>
<tr>
<th>Spontaneous Emergency Sentence</th>
<th>Formulated Emergency Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Ouch Ouch Ouch)</td>
<td>(Call someone e-lio)</td>
</tr>
<tr>
<td>(Ouch)</td>
<td>(e-lio call someone)</td>
</tr>
<tr>
<td>(Shit)</td>
<td>(e-lio call an ambulance)</td>
</tr>
<tr>
<td>(I failed)</td>
<td>(e-lio phone to an ambulance)</td>
</tr>
<tr>
<td>(I can’t get up)</td>
<td>(e-lio call for help)</td>
</tr>
<tr>
<td>(What happened)</td>
<td>(e-lio call for help)</td>
</tr>
<tr>
<td>Ouch I have pain</td>
<td>e-lio call my daughter</td>
</tr>
<tr>
<td>Ouch I bleed I am injured</td>
<td>e-lio call for assistance</td>
</tr>
</tbody>
</table>

3.2.2 A system adapted to the application

First of all, acoustic models of the ASR must be adapted to the acoustic environment and to the person and her situation. For this purpose, speech corpora adapted to this context must be used for model training. Secondly, the set of possible emergency calls must be determined. This set should be composed of the sentences a person could utter during an emergency situation, for instance, when she/he has fallen. The determination of a list of these calls is a challenging task because the person could utter sentences that are difficult to anticipate. Therefore, our list was defined in collaboration with the GRePS laboratory after an ethnological study (Bobillier Chaumon et al., 2014) and as an extension of previous studies (Vacher et al., 2011). Some example are presented in Table 1, with the first category representing spontaneous calls and the second representing vocal commands intended for the e-lio system. Table 2 gives the English translation of these sentences.

Only the emergency calls should be identified, the colloquial sentences must be rejected. To this end, a specialized language model and a generic language model are considered. The generic model is estimated from French news wire collected data. The specialized language model is used to reduce the linguistic variability, and is learnt from the sentence list we developed.

The final Language Model (LM) is a combination of the generic LM (with a 10% weight) and the specialized LM (with 90% weight). This weighting has been shown to lead to low Word Error Rate (WER)³ (Morris et al., 2004) for domain specific applications (Lecouteux et al., 2011). The intent of such a combination is to bias the recognition towards the domain LM, but when the speaker deviates from the domain, the general LM helps avoid the incorrect recognition of sentences leading to “false-positive” detection.

³The WER is the more commonly used metric for comparing different ASR systems as well as for evaluating improvements within one system. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. WER can then be computed as: \( WER = \frac{S+D+I}{N} \), where \( S \) is the number of substitutions, \( D \) the number of deletions, \( I \) the number of insertions, \( N \) the number of words of the reference.
3.3 Call for help recognition

We propose to transcribe each identified call and ASR output into a phoneme graph in which each path corresponds to a variant of pronunciation. Table 3 shows the phoneme representation of some predefined emergency calls. \( \mathcal{L} \) represent the set of identified calls for help \( \mathcal{L} = \{H_l, 1 \leq l \leq L\} \) at the phonetic level. For each ASR output \( O \), the corresponding phonetic transcription \( T = \{t_i, 1 \leq i \leq n\} \) is identified and every call \( H_l = \{h_j, 1 \leq j \leq m_l\} \in \mathcal{L} \) is aligned to \( T \) using Levenshtein distance (Levenshtein, 1966). The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into another. If this distance from a particular \( H_l \) after normalisation with its phoneme number is above a predefined threshold, the ASR output \( O \) is ignored.

### Table 3: Phoneme coding of some predefined emergency calls \( H_l \) (IPA –International Phonetic Alphabet– code)

<table>
<thead>
<tr>
<th>Identified emergency call: ( H_l )</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appelez le SAMU</td>
<td>a p ô l e s a m y</td>
</tr>
<tr>
<td>Qu’est-ce qu’il m’arrive</td>
<td>k e s o k i l m a r i v o</td>
</tr>
<tr>
<td>Aïe aïe aïe</td>
<td>a j a j a j</td>
</tr>
<tr>
<td>Je suis tombé</td>
<td>ʒ ô s ô i t ô b e</td>
</tr>
</tbody>
</table>

This approach takes into account some recognition errors like word endings or light variations in syntax or orthography. Moreover, in many cases, an improperly decoded word is phonetically close to the true one (due to the close pronunciation).

If the normalized distance between a sentence and an identified emergency call is under a specified threshold, this call is detected; this threshold is empirically defined and is a function of the number of aligned phonemes. If the corresponding sentence is a colloquial sentence, it is a False Positive; if the corresponding sentence is an emergency call, it is a True Positive (benefits). In the opposite case, a colloquial sentence identified as an emergency call is a True Negative and an emergency call identified as a colloquial sentence is a False Negative (costs). From this the Sensitivity (Se) or True Positive Rate (TPR) is defined as:

\[
Se = TPR = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}
\]

\[
Se = \frac{\text{Detected Call}}{\text{Detected Call} + \text{Missed Call}}
\]

Defining a Colloquial sentence a sentence that does not represent a Call for Help, the False Positive Rate (FPR) and the Specificity (Spc) are defined as:

\[
FPR = 1 - Spc
\]

\[
FPR = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}
\]

\[
FPR = \frac{\text{False Alarm}}{\text{False Alarm} + \text{Detected Colloquial Sentence}}
\]
Thus, the Call Error Rate is defined as:

$$\text{CER} = \frac{\text{Missed Call} + \text{False Alarm}}{\text{Missed Call} + \text{Detected Call}} = \frac{\text{False Negative} + \text{False Positive}}{\text{False Negative} + \text{True Positive}}$$

(6)

F-measure gives system ability to find all relevant results and reject others:

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

(7)

where

- **Precision** = \( \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \)
- **Recall** = \( \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \)

4 Online experiments

This section describes an experiment in Living lab conditions where participants were asked to enact risky situations (corresponding to Speaking conditions #2 in Figure 2) (Vacher et al., 2016) and is briefly summarised next. The experiment was made with a prototype audio processing system to evaluate its online processing capability. In this section, "online" refers to the complete processing chain, from signal to the detected emergency call. The acoustic signal is treated in a streaming mode.

4.1 Play of scenarios in Living Lab: experimental framework

![Diagram of experimental room and equipment](image-url)

Figure 3: Experimental room and equipment.
Some scenarios including calls for help were played out in a Living lab (Vacher et al., 2016). The room in which the experiment was performed is displayed in Figure 3a. The choice of this room and its equipment was made according to studies made by the GRePS laboratory (Bobillier Chaumon et al., 2014). Indeed, emergency situations occur often in the living room (Bobillier-Chaumon et al., 2016; RoSPA, 2016) where older adults spend a great part of their time.

Only the setup related to audio recordings is presented. For details related to video analysis and fall detection, the reader is referred to (Bouakaz et al., 2014). The experimental room is equipped like a living room with a sofa, a carpet, a cupboard and a table. An HF microphone Sennheiser EW-300-G2 was setup in the ceiling for audio recordings (2.5m height). The special device e-lio and a TV was placed on the table in front of the sofa, in the usual place for this equipment.

The participants played out five types of risky situations: slip, stumble, a fall in a stationary position and being unable to rise from the sofa due to a blocked hip. These situations were selected because they were representative falls in a domestic environment and could be played safely. Two other scenarios, called “true-negative”, were added for providing situations that look like emergency situations but which are not. Emergency calls were included in each scenario except the true negative ones, with sentences coming from AD80.

### 4.2 Real-time audio analysis: CirdoX system

In order to process the uttered sentences on the fly, we have developed a real-time sound analysis system, CirdoX, able to inter-operate with an ASR system. The goal of CirdoX is to process, on-line, older persons' emergency calls. The architecture of the system is provided Figure 4. A complete description of CirdoX is given in (Aman et al., 2016b). CirdoX performs the following tasks:

1. the acquisition of the audio signal;
2. the detection of sound events (speech or non-speech sound);
3. differential treatment of speech and non-speech with a call to the ASR module in the case of speech and a call to a sound classifier in the case of non-speech;
4. and finally filtering of keywords corresponding to emergency calls.

In case of call for help recognition, the appropriate service is initiated using a specialised device such as e-lio in charge of alert management. CirdoX is able to process on the fly the data obtained by several microphones.

![Figure 4: Global organisation of CirdoX.](http://www.technosens.fr/1-4612-e-lio-au-domicile.php)
4.3 Data set record

As stated in section 3.1, the age at which a person is considered old is not clearly defined, we consider here that a person over 60 years of age is aged.

Ideally falls and other scenarios should be played out by older adults. However, it is difficult to find seniors willing and able to play such scenarios. To record realistic data (but not necessarily played out by older adults), 13 people under 60 were recruited. These younger volunteers were instructed to wear equipment (i.e., an old age simulator) which hampered mobility and reduced vision and hearing. The voice is not affected. A participant wearing this simulator is shown Figure 3b.

Overall 17 participants were recruited (9 men and 8 women). Table 4 summarises the content of the corresponding corpus: the Cirdo-set corpus.

Table 4: Cirdo-set: audio data produced during the experiment in DOMUS, participants are older adults or people wearing an old age simulator (Vacher et al., 2016)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Gender</th>
<th>Age</th>
<th>Wav length</th>
<th>SNR average</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>M</td>
<td>30</td>
<td>8min 40s</td>
<td>25.6dB</td>
</tr>
<tr>
<td>S03</td>
<td>F</td>
<td>24</td>
<td>4min 35s</td>
<td>19.9dB</td>
</tr>
<tr>
<td>S05</td>
<td>M</td>
<td>29</td>
<td>6min 30s</td>
<td>22.7dB</td>
</tr>
<tr>
<td>S08</td>
<td>M</td>
<td>44</td>
<td>5min 54s</td>
<td>22.2dB</td>
</tr>
<tr>
<td>S09</td>
<td>F</td>
<td>16</td>
<td>8min 50s</td>
<td>17.8dB</td>
</tr>
<tr>
<td>S10</td>
<td>M</td>
<td>16</td>
<td>5min 07s</td>
<td>25.3dB</td>
</tr>
<tr>
<td>S11</td>
<td>M</td>
<td>52</td>
<td>5min 17s</td>
<td>23.7dB</td>
</tr>
<tr>
<td>S12</td>
<td>M</td>
<td>28</td>
<td>7min 04s</td>
<td>21.1dB</td>
</tr>
<tr>
<td>S14</td>
<td>F</td>
<td>52</td>
<td>6min 48s</td>
<td>17.3dB</td>
</tr>
<tr>
<td>S15</td>
<td>M</td>
<td>23</td>
<td>5min 50s</td>
<td>22.0dB</td>
</tr>
<tr>
<td>S16</td>
<td>F</td>
<td>40</td>
<td>7min 31s</td>
<td>23.3dB</td>
</tr>
<tr>
<td>S17</td>
<td>F</td>
<td>40</td>
<td>8min 01s</td>
<td>21.0dB</td>
</tr>
<tr>
<td>S18</td>
<td>F</td>
<td>25</td>
<td>5min 54s</td>
<td>24.7dB</td>
</tr>
<tr>
<td>S04</td>
<td>F</td>
<td>83</td>
<td>9min 07</td>
<td>20.0dB</td>
</tr>
<tr>
<td>S06</td>
<td>F</td>
<td>64</td>
<td>6min 31</td>
<td>19.1dB</td>
</tr>
<tr>
<td>S07</td>
<td>M</td>
<td>61</td>
<td>6min 00</td>
<td>19.5dB</td>
</tr>
<tr>
<td>S13</td>
<td>M</td>
<td>66</td>
<td>7min 16</td>
<td>17.6dB</td>
</tr>
</tbody>
</table>

4.4 Speech extraction from recorded audio data

CirdoX was used in order to process the data and to extract online the sentences uttered during the experiment. The \textit{Speech/Non-speech Discrimination} stage was performed using Gaussian Mixture Models (GMMs) which were trained using the ALIZE software (Bonastre et al., 2005). The GMM approach is based on the fact that the probabilistic distribution of a (multivariate) random variable (here speech and non-speech) can be modelled as the (weighted) sum (i.e., the mixture) of several Gaussian distributions. Training data were composed of distant speech records from the \textit{SWEET-HOME} corpus (Vacher et al., 2014): 9,147 speech files (voice commands or emergency calls) were used to train the speech model (i.e., 4 hours 53 minutes), 13,448 sound files (i.e., 5 hours 10 minutes) were used to train the non-speech model.
The Detection module extracted 1,950 sound events from the data recorded by the 17 participants (Cirdo-set). The *Speech/Non-speech Discrimination* module had an overall accuracy of 95.3%. It classified 244 of the events as speech: 237 were really speech, 7 were non-speech events classified as speech; in the same way, 85 speech events were incorrectly classified as non-speech. The *Speech/Non-speech Discrimination* module classifies with high specificity \((Spc = 0.996)\) (non-speech events classified as non-speech). However, the classification sensitivity, or the classification rate of speech events as speech, is low \((Se = 0.736)\). A large part of the speech events were emergency calls but a small number were colloquial sentences spontaneously uttered by participants or experimenters.

### 4.5 ASR training and adaptation

Automatic speech recognition was performed using Sphinx3 with 2 different acoustic models. The first is *BREF120* which was built from the BREF120 corpus described in Section 2.4 and *BREF120_SWEET-HOME_G* the second, is an adaptation of the *BREF120* acoustic model to older adults’ voice and to distant speech conditions using Maximum Likelihood Linear Regression (MLLR) (Gales, 1998). Given an initial GMM and a new dataset, MLLR is a method to re-estimate the means of the Gaussians of the acoustic model so that it maximises the probability of the model to generate the new dataset. When few target data is available, this is a way to learn models from large amount of non-target data to adapt to the target task thanks to the few target data at hand. MLLR was applied using the “User Specific Interaction” subset of the *SWEET-HOME* corpus (Vacher et al., 2014) composed of 337 sentences (9min 30s of signal) uttered in the DOMUS flat by older adults or visually impaired people (average age: 72 years).

Table 5 describes the characteristics of corpora used for acoustic and language model training; GIGAWORD\(^5\) and AD80 are used as text corpus.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Role</th>
<th>Number of speakers</th>
<th>Age</th>
<th>Size</th>
<th>Recording conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWEET-HOME (Multimodal subset)</td>
<td>Discrimination model</td>
<td>21</td>
<td>22-63</td>
<td>speech: 9.147 files non-speech: 13.448 files</td>
<td>Distant speech in a smart home</td>
</tr>
<tr>
<td>BREF120</td>
<td>Generic acoustic model</td>
<td>120</td>
<td>20-65</td>
<td>speech: 100h</td>
<td>Read text (recording studio)</td>
</tr>
<tr>
<td>SWEET-HOME (User specific interaction subset)</td>
<td>Acoustic model adaptation</td>
<td>11</td>
<td>49-91</td>
<td>speech: 9mn 30s</td>
<td>Interaction in a smart home (distant speech)</td>
</tr>
<tr>
<td>GIGAWORD(^4)</td>
<td>Generic LM</td>
<td>-</td>
<td>-</td>
<td>text: 13.304 words</td>
<td>Text corpus</td>
</tr>
<tr>
<td>ANODIN/DÉTRESSE (AD80)</td>
<td>Specific LM</td>
<td>95</td>
<td>18-94</td>
<td>text: 99 words speech: 2h 18mn</td>
<td>Read text</td>
</tr>
</tbody>
</table>

\(^5\)http://catalog.ldc.upenn.edu/LDC2006T17
4.6 WER results

Input to the ASR are the decoded audio events from the previous stages of Detection and Speech/Non-speech Discrimination by CirdoX. That is 244 audio events were decoded: 204 of these are emergency calls, 33 are colloquial sentences and 7 are non-speech signals. Table 6 presents the decoding results. Regardless of the acoustic model, WER is very high for colloquial sentences. This is due to the language model which is non adapted to this task and is actually desirable: for privacy reason, it is not appropriate that these kinds of sentences be well recognized. For emergency calls, WER is high using the generic model BREF120 (80.46%). We hypothesize that this is due to recording conditions being very different between BREF120 corpus (reading of speech close to the microphone) and our experimental conditions (distant speech, microphone setup in the ceiling and utterance during a scenario difficult to play). Using a model adapted to our recording conditions and to older adults’ voices provides a significant improvement. With the BREF120_SWEET-HOME_G model, WER is reduced to 49.32%, an improvement in performance of 38%.

Table 6: WER for sentences of the Cirdo-set audio data.

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>BREF120</th>
<th>BREF120_SWEET-HOME_G</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER calls for help</td>
<td>80.46%</td>
<td>49.32%</td>
</tr>
<tr>
<td>WER other (colloquial and non-speech)</td>
<td>107.14%</td>
<td>102.04%</td>
</tr>
</tbody>
</table>

4.7 Results of help call recognition

Speech recognition hypothesis is then filtered by CirdoX (as described in Section 3.3). The distance is normalized by the number of candidate phonemes in the ASR hypothesis before applying a threshold. If the normalized distance is lower than the threshold, the sentence is considered to be an emergency call. The threshold was fixed using a ROC curve analysis (for each acoustic model). The confusion matrix for the two acoustic models is presented in Table 8. From these results, sensitivity (Se), specificity (Spc) and False Alarm Rate (FAR) were evaluated with the corresponding values presented in Table 7. Despite the important gain in WER performance, we can observe only a small increase in sensitivity and a corresponding small decrease in specificity. With the adapted model, about 78% of the calls could be detected, but the false alarm rate is 6.5%.

Table 7: Sensitivity, specificity and false alarm rate of the filtering stage from automatically detected speech

<table>
<thead>
<tr>
<th></th>
<th>BREF120_SWEET-HOME_G</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER</td>
<td>49.3%</td>
</tr>
<tr>
<td>Se</td>
<td>78.0%</td>
</tr>
<tr>
<td>Spc</td>
<td>72.5%</td>
</tr>
<tr>
<td>FAR</td>
<td>6.5%</td>
</tr>
</tbody>
</table>
Table 8: Confusion matrix for speech filtering.

<table>
<thead>
<tr>
<th>Threshold=1.364</th>
<th>d&lt;threshold</th>
<th>d≥threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help calls</td>
<td>TP=159</td>
<td>FN=45</td>
</tr>
<tr>
<td>Other</td>
<td>FP=11</td>
<td>TN=29</td>
</tr>
</tbody>
</table>

(colloquial and non-speech)

4.8 Overall performances

Results on the Cirdo-set corpus using CirdoX are summarized in Figure 5. 1950 audio events (277 emergency calls, 45 colloquial sentences, 1628 non-speech) were extracted by the Detection stage. The Speech/Non-speech Discrimination stage had an accuracy of 95.3%: 7 non-speech were misclassified as speech and 85 speech utterances were misclassified as non-speech. Non-speech events were not sent to the Filtering stage. Thus, 1621 non-speech events and 12 colloquial sentences are identified as True Negatives, 73 emergency calls as False Negative (i.e., missed alarms).

Speech events were sent to the ASR and then to the Filtering stage. For the ASR, the results of the acoustic models are reported. With BREF120 model, 31 colloquial sentences and non-speech events were rejected (True Negatives), 57 emergency call were rejected by mistake (False Negatives), 9 colloquial sentences and non-speech events were kept by mistake (False Positives), 147 emergency calls were properly identified (True Positives). In the case of the BREF120_SWEETHOME_G adapted model, 159 emergency calls were properly recognised. This improvement is due to the better recognition of calls for help with this model as reported Table 6 (WER = 49.32% versus 80.48% for BREF120 model).

Table 9: Global performances of the Filtering stage for the BREF120 and adapted (BREF120_SWEETHOME_G) models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F-measure</th>
<th>FAR</th>
<th>CER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BREF120</td>
<td>53%</td>
<td>99.5%</td>
<td>94.2%</td>
<td>68%</td>
<td>0.54%</td>
<td>32%</td>
</tr>
<tr>
<td>BREF120_SWEETHOME_G</td>
<td>57%</td>
<td>99.3%</td>
<td>93.5%</td>
<td>71%</td>
<td>0.66%</td>
<td>27%</td>
</tr>
</tbody>
</table>

As shown Table 9, best results are achieved with the adapted acoustic model (BREF120_SWEETHOME_G). Recall is 57%, which is still relatively low for use in real-world applications. This recall value means only 57% of the emergency calls send an alert to the appropriate service. The False Alarm Rate (FAR) is extremely low (0.66%). This is due to the high number of non-speech events and their effective classification as non-speech. Moreover a part of the non-speech events misclassified as speech are rejected by the Filtering stage. As shown in the previous section, considering only the Filtering stage, FAR is dramatically different, 6.5%, because at this step, only events classified as speech are taken into account. All the same, this higher value (6.5%) must be taken into consideration and it is too large for use in the real world. In the same way, the Call Error Rate (CER: c.f. Equation 6) which is the most suitable criterion for evaluating the overall performance of the system is 32% with the BREF120 model; it is reduced to 27% with the adapted model which represents a significant improvement in performance.
4.9 Analysis of the results

Figure 5 summarises the results of the different stages of the analysis. The Speech/Non-speech Discrimination stage classified 85 speech events as non-speech. 73 of these are emergency calls, and due to this classification error, they are missed. One explanation could be the presence of noise in the speech. To verify this, we manually listened to each speech event and analyzed it. We concluded that in fact 95 speech events are not pure speech. In most cases they are a superposition of speech and of the “noise” made by the person during the fall (occurring when the person calls during the fall and not before or after). Discrimination results concerning all speech events are presented in Table 10. Only 43.2% of speech events are classified as speech when they are noisy, versus 86.3% for clean speech. It is also not surprising, all depend on the proportion of noise related to speech in terms of energy or time.

It is now necessary to determine whether noisy sentences could be well recognized by the ASR. Table 11 displays WER for the different categories of emergency call speech events: clean/noisy, and these discriminated as speech/non-speech. With the adapted model, WER for noisy calls is approximately 100% and is 50% for clean speech. Noise has a strong impact on performance. Moreover, calls classified as non-speech are very badly recognized (WER=101%), whether noisy speech (WER=109%) or clean speech (WER=87.7%). Therefore, the Speech/Non-speech

Table 10: Speech/Non-speech discrimination of speech events in presence or absence of noise.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Classified as speech</th>
<th>Classified as non-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>227</td>
<td>196 (86.3%)</td>
<td>31 (13.7%)</td>
</tr>
<tr>
<td>Noisy speech</td>
<td>95</td>
<td>41 (43.2%)</td>
<td>54 (56.8%)</td>
</tr>
<tr>
<td>All speech</td>
<td>322</td>
<td>237 (73.6%)</td>
<td>85 (26.4%)</td>
</tr>
</tbody>
</table>
Discrimination stage would not improve performance and this stage seems to be a good filter for bad quality speech events. In the matter of the clean calls classified as non-speech, 25 are very short sentences like “oh la”, “aïe” and “j’ai mal”. What is worthy of note, however, is that this kind of sentence is very challenging for the language model of the ASR because of its small number of words.

Table 11: WER for emergency call speech event categories and the two language models: generic (BREF120) and adapted (BREF120_SWEETHOME_G).

<table>
<thead>
<tr>
<th>Help calls</th>
<th>Number</th>
<th>WER (Generic model)</th>
<th>WER (Adapted model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>203</td>
<td>81.4%</td>
<td>49.4%</td>
</tr>
<tr>
<td>Noisy</td>
<td>74</td>
<td>107%</td>
<td>95.5%</td>
</tr>
<tr>
<td>Classified as:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- speech</td>
<td>204</td>
<td>80.5%</td>
<td>49.3%</td>
</tr>
<tr>
<td>- non-speech</td>
<td>73</td>
<td>114%</td>
<td>101%</td>
</tr>
<tr>
<td>Clean call classified as</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- speech</td>
<td>175</td>
<td>78.8%</td>
<td>45.5%</td>
</tr>
<tr>
<td>- non-speech</td>
<td>28</td>
<td>108%</td>
<td>87.7%</td>
</tr>
<tr>
<td>Noisy call classified as</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- speech</td>
<td>29</td>
<td>93%</td>
<td>77.9%</td>
</tr>
<tr>
<td>- non-speech</td>
<td>45</td>
<td>117%</td>
<td>109%</td>
</tr>
<tr>
<td>All help calls</td>
<td>277</td>
<td>87.4%</td>
<td>59.4%</td>
</tr>
</tbody>
</table>

5 Improvement of the system

Since the observed performance described above was insufficient, we set out to improve it. To do so, we chose to use the Kaldi speech recognition tool-kit (Povey et al., 2011b) to build a system suitable for older adults’ and expressive voices. Kaldi is an open-source state-of-the-art ASR system with a high number of tools and a strong support from the community. This new built system (Vacher et al., 2015b) was evaluated off-line using Cirdo-set in the same manner as in section 4. The challenge was to determine whether the use of SGMM-based acoustic models would improve performance through adaptation to the environment and users. SGMMs were chosen over Deep Neural Network (DNN) models because SGMM is more adapted to situation where a low amount of adaptation data is available (Badenhorst and de Wet, 2017). This section is an extension of our previous work in (Vacher et al., 2015b).

5.1 Subspace GMM Acoustic Modelling

The GMM and Subspace GMM (SGMM) both model the emission probability of each HMM state using a Gaussian mixture model. However, in the SGMM approach, the Gaussian means and the mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections.
The SGMM model is described in the following equations (Povey et al., 2011a):

\[
\begin{align*}
    p(x|j) &= \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} N(x; \mu_{jmi}, \sigma_i), \\
    \mu_{jmi} &= M_i v_{jm}, \\
    w_{jmi} &= \frac{\exp w_{jmi}^T v_{jm}}{\sum_{i'} \exp w_{jmi}^T v_{jm}},
\end{align*}
\]

where \(x\) denotes the feature vector, \(j \in \{1..J\}\) is the HMM state, \(i\) is the Gaussian index, \(m\) is the substate and \(c_{jm}\) is the substate weight. Each state \(j\) is associated with a vector \(v_{jm} \in \mathbb{R}^S\) (\(S\) is the phonetic subspace dimension) which derives the means, \(\mu_{jmi}\) and mixture weights, \(w_{jmi}\) and it has a shared number of Gaussians, \(I\). The phonetic subspace \(M_i\), weight projections \(w_{jmi}^T\) and covariance matrices \(\sigma_i\) (i.e, the globally shared parameters \(\tilde{\gamma}_i = \{M_i, w_{jmi}^T, \sigma_i\}\)) are common across all states. These parameters can be shared and estimated over multiple record conditions.

A generic mixture of \(I\) gaussians, denoted as Universal Background Model (UBM), models all the speech training data for the initialization of the SGMM. Povey et al. (Povey et al., 2011a) showed that the model is also effective with large amounts of training data. The acoustic models were trained on 500 hours of transcribed French speech composed of the ESTER 1&2 (broadcast news and conversational speech recorded on the radio) and REPere (TV news and talk-shows) challenges as well as from 7 hours of transcribed French speech of the SWEET-HOME corpus (Vacher et al., 2014) which consists of records of 60 speakers interacting in a smart home and from 28 minutes of the Voix-détresse corpus (Aman et al., 2016a) which is made of recordings of speakers eliciting distress.

Our aim is to bias the acoustic model to the smart home and expressive speech conditions, therefore three UBMs were trained separately. The generic UBM was obtained from clean speech data from ESTER and REPere, the smart-home UBM from SWEET-HOME corpus and the expressive UBM from Voix-détresse. These tree UBMs contained 1K gaussians and were merged into a single one mixed down to 1K gaussian (closest Gaussians pairs were merged (Zouari and Chollet, 2006)).

5.2 ASR adapted to the task and the person

Acoustic model In our experiments, context-dependent classical three-state left-right HMMs were used as acoustic models. Acoustic features were based on Mel-Frequency Cepstral Coefficients (MFCC). 13 MFCC-features coefficients were first extracted and then expanded with delta and double delta features and energy (40 features). The SGMM model presented in Section 5.1 was used as acoustic model, it was composed of 11,000 context-dependent states and 150,000 Gaussians. State tying is performed using a decision tree based on tree-clustering of the phones. In addition, off-line fMLLR linear transformation acoustic adaptation was performed.

Language model In the same way, the language model was adapted to the task. Firstly, a generic language model (LM) was estimated from French newswire collected in the Gigaword corpus. This model was 1-gram model with 13,304 words. Secondly, to reduce the linguistic variability, a 3-gram domain language model, the specialized language model was learnt from the sentences used during the corpus collection described in Section 4.3, with 99 1-gram, 225 2-gram
Table 12: SGMM system: Word and Call Error Rate for the two categories of participants: 13 young participants wearing an age simulator and 4 older adults.

<table>
<thead>
<tr>
<th>Young Speaker</th>
<th>WER (%)</th>
<th>Emergency</th>
<th>CER (%)</th>
<th>Older Speaker</th>
<th>WER (%)</th>
<th>Emergency</th>
<th>CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>45.0</td>
<td>39.1</td>
<td>27.8</td>
<td>S04</td>
<td>51.9</td>
<td>49.6</td>
<td>34.0</td>
</tr>
<tr>
<td>S03</td>
<td>41.4</td>
<td>44.4</td>
<td>40.0</td>
<td>S06</td>
<td>39.2</td>
<td>34.3</td>
<td>26.3</td>
</tr>
<tr>
<td>S05</td>
<td>19.1</td>
<td>15.4</td>
<td>14.3</td>
<td>S07</td>
<td>21.2</td>
<td>20.3</td>
<td>28.6</td>
</tr>
<tr>
<td>S08</td>
<td>61.8</td>
<td>50.8</td>
<td>20.0</td>
<td>S13</td>
<td>45.9</td>
<td>43.6</td>
<td>23.8</td>
</tr>
<tr>
<td>S09</td>
<td>49.4</td>
<td>41.2</td>
<td>33.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S10</td>
<td>24.5</td>
<td>22.4</td>
<td>14.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S11</td>
<td>21.3</td>
<td>17.0</td>
<td>16.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>30.8</td>
<td>25.0</td>
<td>25.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S14</td>
<td>67.0</td>
<td>54.8</td>
<td>50.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>21.5</td>
<td>19.5</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S16</td>
<td>14.9</td>
<td>11.76</td>
<td>7.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S17</td>
<td>21.4</td>
<td>22.4</td>
<td>19.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S18</td>
<td>57.7</td>
<td>44.9</td>
<td>71.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Young (all)</strong></td>
<td><strong>34±18</strong></td>
<td><strong>30±15</strong></td>
<td><strong>23±13</strong></td>
<td><strong>Older (all)</strong></td>
<td><strong>40±13</strong></td>
<td><strong>37±13</strong></td>
<td><strong>28±4</strong></td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>36±16</strong></td>
<td><strong>32±14</strong></td>
<td><strong>24±12</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and 273 3-gram models. Finally, the language model was a 3-gram-type which resulted from the combination of the *generic LM* (with a 10% weight) and the *specialized LM* (with 90% weight).

5.3 Off line experiments

This adapted system was evaluated using the Cirdo-set corpus presented in Section 4.3 according to the methods presented in Section 3. Results on manually annotated data are given in Table 12. The most important performance measures are the Word Error Rate (WER) of the overall decoded speech and those of the specific emergency calls as well as the Call Error Rate (CER: c.f. Equation 6). Considering emergency calls only, the average WER is 32.0% (young: 30% - older: 37%), whereas it is 36% (young: 34% - older -40%) when all interjections and sentences are taken into account. These results show a significant improvement over the previous system given in Table 6 but this does not suffice to ensure satisfactory performances.

As a matter of fact, CER is the most suitable criterion for evaluating the overall performance of the system. On average, CER is equal to 24% (young: 23% - older: 28%) with significant disparity between the speakers. CER is greater for older adults, this reflects the fact that less data specific to older adults has been used to train the acoustic model, particularly with respect to expressive voice. Furthermore, the *Voix-détresse* corpus was recorded by persons sitting on a chair and eliciting emotions. These conditions are quite different to those of Cirdo-set: some participants called out loudly and almost screamed (S18), others called while sighing (S04).

Excepting for one speaker (CER=71.4%), CER is always below 50% and consequently more
than 50% of the calls were recognized. For 6 speakers, CER was below 20%. This suggests that an emergency call could be detected if the speaker is able to repeat his call two or three times. However, if the system did not identify the first emergency call because the person's voice is altered by stress, it is likely that this person will have more and more stress and, as a consequence, future calls would be even more difficult to identify.

Previous studies based on the AD80 corpus showed recall, precision and F-measure equal to 88.4%, 86.9% and 87.2% (Aman et al., 2013b). However, this corpus was recorded in very different conditions (text reading in a studio, no expressive speech) in contrast to those of Cirdo-set.

6 Discussion

The results of the experiments presented in the previous sections shed light on the two research questions presented in the introduction:

1. Is it possible to adapt an ASR system to older adults in order to detect predefined emergency calls?
2. Is an ASR system, implemented in a real-time and running on-line, able to detect predefined emergency calls in a realistic environment?

The results concerning only the speech recognition for the on-line experiment (Section 4.6) show a very poor performance of 80.5% WER for the emergency calls in the case of the baseline system. This is due to the distant speech conditions of the recording and the noise made by the participants during the experiment. However, the performance can be highly improved when the acoustic models are adapted through MLLR to the acoustic condition. Hence, the same experiment with the adapted BREF120_SWEET-HOME_G model demonstrate an average WER of 49.32% using only 9 minutes of speech for adaptation. When the off-line experiment is considered, the use of more sophisticated models such as SGMM bring an average improvement to a WER of 37% for the emergency calls of the seniors. It must be emphasised that the SGMM was adapted on a larger corpus than the previous model. Although these results were obtained on a small test corpus, it is in line with other voice command experiments performed in a smart home (Vacher et al., 2015a), but in quiet environment, where the obtained WER was 43% in average. Our application faces more difficult conditions, due to differences in the environment (noise generated by a fall or a device) and feelings of emergency in the person. The difference in WER between older adults’ voice and the youngest voice is substantial which is again in line with the literature but also explainable by the fact that the SNR of older adults’ speech signal was amongst the lowest in the corpus and so the hardest to process.

We have thus showed that it is possible to adapt an ASR system to older adults and distant speech to improve greatly the ASR performances. However, the performance still need to be improved by increasing the robustness to noise and by considering a larger dataset for adaptation.

The real-time aspect of the ASR cannot be analysed without taking into consideration the whole processing chain. As reported, a fair amount of emergency calls were filtered out by the Speech/Non-speech discrimination module. Although the accuracy is 95.3% overall, 73 emergency calls over 277 (26%) were classed as Non-speech and then were not analysed further.
Some of the missed emergency calls were altered with background noise. Two methods exist to handle this problem:
1) noise cancellation when the noise source is known and can be isolated (e.g., TV or radio) which is not possible in our case (Vacher et al., 2012) or 2) source separation techniques which have been recently applied to detection of keywords in a noisy environment (Vincent et al., 2013). The related works in this domain (Rotili et al., 2013; Ravanelli and Omologo, 2014; Vincent et al., 2017) show that although it is still an open problem, there are ways to perform robust ASR in noisy conditions with good localisation of the noise sources. However, it is unclear to what extent this can be performed in real time.

The best average CER obtained in older adults’ case was 28% while it was 23% for the youngest participants. This means that in practice every 4 emergency calls, the person should repeat it to have a chance to be heard. Surprisingly the CER is only slightly related to the WER in emergency situations. Although, the lower the WER, the lower the CER overall, a WER of 50.8% can lead to CER=20% (S08) and a WER of 44.9% can lead to a CER of 71.4% (S18). This shows that there is a clear need for other measures to be able to relate the CER performance to the ASR performance. This also shows that the Levenshtein distance, despite its simplicity, is a robust measure. Despite the improvements to be made, this work shows that it is possible to detect predefined emergency calls in a realistic environment in real-time but that more experiments are needed to analyse deeply the factors that could permit to predict the performance of such system.

As we have just outlined, the main problem of ASR systems, in the smart home context, is resistance to speech signal degradation caused by the environment (long distance capture, mixing with home noises, etc.) and the effects of ageing on voice production. However, it is clear that ASR can be challenged in a real use case, because some commands will be expressive, especially, for emergency calls where sentences are motivated by strong emotion. It can be affirmed that the more a command is related to something important for the user, the more the speech signal is expressive. However, while many studies focus on the automatic recognition of emotions/social affects (Schuller et al., 2011), few evaluations of ASR performance comparing specifically expressive vs. non expressive natural speech could be found in the field on automatic spontaneous speech recognition. (Aman et al., 2013a) observed that there is significant variation in ASR performance depending on how the system is trained. They concluded that for real smart home applications, especially for weakened users, this problem must be taken into account. It is an open question whether the ASR could be helped by an automatic emotion recognition processing system (that is not directly required for the overall task).

In this work, the non-speech sounds were not used in the emergency call recognition. However, non-linguistic information such as scream, fall of objects, groaning, etc could be identified as supplementary evidence for inferring an emergency situation. Sound classification is a domain which has seen a regain of interest as exemplified by the D-CASE challenge (Stowell et al., 2015) which aims at classifying environmental acoustic scenes. However, sound classification in the home is very difficult given the large number of sound events that can occur (Sehili et al., 2012) and the difficulty to disambiguate them (a large number of sounds can sound like a fall). This is even more an issue since corpora including acoustic recording of real falls (not simulated ones) are nonexistent or not publicly available. Nevertheless, work must be undertaken to study the acoustic context of an emergency call to estimate whether non-linguistic evidence can be fused with speech utterances in order to improve emergency call recognition.
7 Conclusion

This paper reports a study of a VUI system for emergency calls recognition with the aim of helping older isolated people to live longer at home, by recognising and detecting calls for help and connecting them to helpers or caregivers. The on/off-line multi-source speech and sound analysis software, CirdoX, was developed for this purpose. This software was used during an experiment involving participants who called for help after falling on a carpet or when being unable to get up from a sofa because of a blocked hip. The recognition of emergency calls was insufficient (CER 27%) but more sophisticated ASR approaches improved these results (call error rate 24%). The most important limitation was the noise produced by a fall when the person is speaking while falling. Nevertheless, emergency call recognition from ASR hypothesis using acoustic model adapted to older adults' voices might be a promising way to help older adults to live at home in an independent manner and could be used with other sensors such as in (Vacher et al., 2015a).

Another contribution of this work was the corpus of audio recordings of people falling and calling for help in a Living lab environment. The participants were 4 older adults and 13 younger people wearing an old age simulator which hampered mobility, reduced vision and hearing. When they played out the scenarios, some participants produced sighs, grunts, coughs, cries, groans, panting or throat clearings. Overall, each speaker uttered between 10 and 65 short sentences or interjections (“ah”, “oh”, “aïe”, “je peux pas me relever”, etc.). This corpus is made available for research (Vacher et al., 2016).

This experiment had some limitations and the scenarios examined were very constrained, but it still constitutes a necessary step towards the development of automatic speech recognition applications for individuals living in isolation in their homes. Before making these systems available to the public, datasets and evaluation methods must be designed and shared among the community for the quick replication, evaluation and development of technological advances. Next, improvements must be performed at the acoustic level to produce more adapted ASR to ageing voices, emotion (Aman et al., 2013a; Schuller et al., 2011) and real-life settings, which will require improved methodology for dataset collection. Finally, the system can be seen as a complementary service to automatic fall detectors which could work together in order to detect emergency situations when the person is not able to talk (in case of a fall where the person becomes unconscious).

8 Acknowledgments

The authors would like to thank the participants who volunteered for the experiments. Thanks are extended to Stefania Raimondo for her proof reading of this article.


