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, B Lecouteux

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.

Reference of the published paper of this draft:
@articleVacher:2019:MEC:3340326.3310132,
author = Vacher, Michel and Aman, Frédéric and Rossato, Solange and Portet, François and Lecouteux, Benjamin,
title = Making Emergency Calls More Accessible to Older Adults Through a Hands-free Speech Interface in the House,
journal = ACM Trans. Access. Comput.,
issue_date = June 2019,
volume = 12,
number = 2,
month = jun,
year = 2019,
issn = 1936-7228,
pages = 8:1–8:25,
articleno = 8,
numpages = 25,
url = http://doi.acm.org/10.1145/3310132,
doi = 10.1145/3310132,
acmid = 3310132,
publisher = ACM,
address = New York, NY, USA,
keywords = Specific voice recognition, ambient assisted living, assistive technology, emergency call,
}

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 *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)

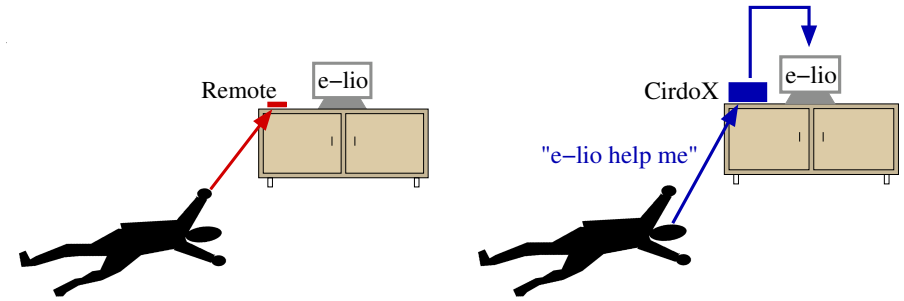
nearly 90 percent of USA citizens over age 65 want to stay in their residence for as long as possible and 80% believe their current residence is where they will always live. This is corroborated by a new study (CSA, 2010) showing that 87% of family members and relatives plan to ensure that the person 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[2] is the specizalized device making emergency calls which is used preferentially in the framework of the Cirdo project.

[2]http://www.technosens.fr/



(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 degrades 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 can not 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 interoperated 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).

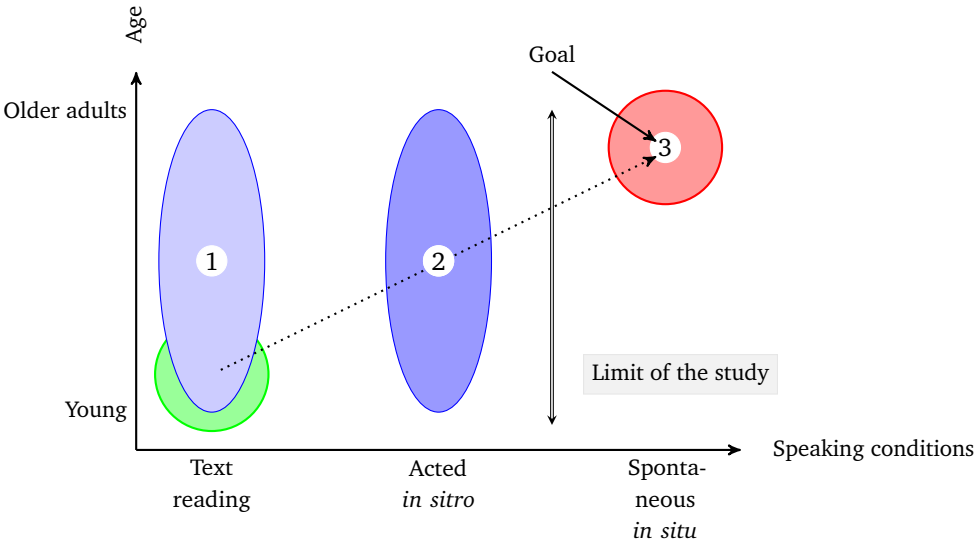


Figure 2: Experimental method for problem solving.

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. *in sitro* emergency by performing a voluntary fall on the carpet,
3. spontaneous emergency *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.

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 sitro*, 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.

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

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

Spontaneous Emergency Sentence	*Formulated Emergency Sentence*
(Ouch Ouch Ouch)	(Call someone e-lio)
(Ouch)	(e-lio call someone)
(Shit)	(e-lio call an ambulance)
(I falled)	(e-lio phone to an ambulance)
(I can't get up)	(e-lio call for help)
(What happened)	(e-lio call for help)
Ouch I have pain	e-lio call my daughter
Ouch I bleed I am injured	e-lio call for assistance

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.

[3]The 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: $\text{WER} = \frac{\text{S+D+I}}{\text{N}}$, 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. $\mathscr{L}$ represent the set of identified calls for help H, $\mathscr{L} = \{H^l, 1 \leq l \leq L\}$ 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 $H^l = \{h^l_j, 1 \leq j \leq m_l\} \in \mathscr{L}$ 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)

Identified emergency call: H^l		Example
Appelez le SAMU	(call an ambulance)	a p ø l e l ə s a m y
Qu'est-ce qu'il m'arrive	(what happened)	k ɛ s ə k i l m a r i v ə
Aïe aïe aïe	(ouch ouch ouch)	a j a j a j
Je suis tombé	(I falled)	ʒ ə s ɥ i t ɔ̃ b e

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:

$$\text{Se} = \text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (1)$$

$$= \frac{\text{Detected Call}}{\text{Detected Call} + \text{Missed Call}} \quad (2)$$

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:

$$\text{FPR} = 1 - \text{Spc} \quad (3)$$

$$= \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \quad (4)$$

$$= \frac{\text{False Alarm}}{\text{False Alarm} + \text{Detected Colloquial Sentence}} \quad (5)$$

Thus, the Call Error Rate is defined as:

$$\text{CER} = \frac{\text{Missed Call} + \text{False Alarm}}{\text{Missed Call} + \text{Detected Call}} = \frac{\text{False Negative} + \text{False Positive}}{\text{False Negative} + \text{True Positive}} \tag{6}$$

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

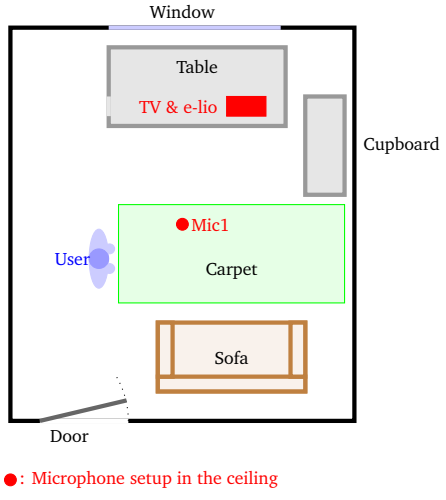
$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

$$\text{where} \begin{cases} \text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \\ \text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \end{cases}$$

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



(a) Experimental room

(b) Worn simulator

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[4] in charge of alert management. *CirdoX* is able to process on the fly the data obtained by several microphones.

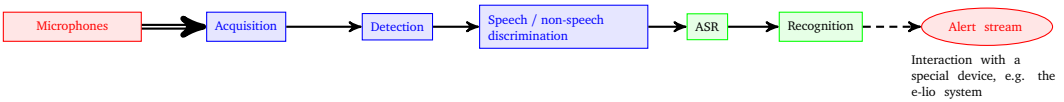


Figure 4: Global organisation of CirdoX.

[4]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)

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

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 *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 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 5: Corpora used for CirdoX training.

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

[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.

Acoustic model	*BREF120*	*BREF120_SWEET-HOME_G*
WER calls for help	80.46%	49.32%
WER other (colloquial and non-speech)	107.14%	102.04%

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

	BREF120_SWEET-HOME_G
WER	49.3%
Se	78.0%
Spc	72.5%
FAR	6.5%

Table 8: Confusion matrix for speech filtering.

Threshold=1.364	d<threshold	d≥threshold
Help calls	TP=159	FN=45
Other (colloquial and non-speech)	FP=11	TN=29

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_SWEET-HOME_G*) models.

Model	Sensitivity	Specificity	Precision	F-measure	FAR	CER
BREF120	53%	99.5%	94.2%	68%	0.54%	32%
BREF120_SWEETHOME_G	57%	99.3%	93.5%	71%	0.66%	**27%**

As shown Table 9, best results are achieved with the adapted acoustic model (*BREF120_SWEET-HOME_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.

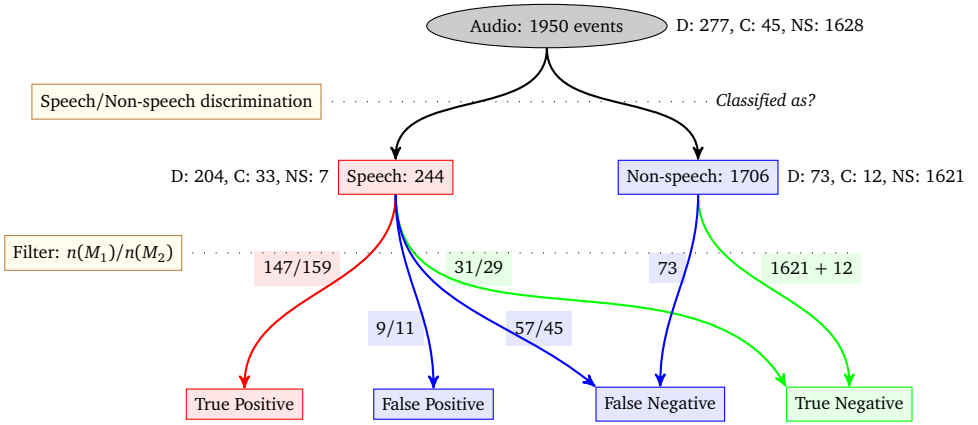


Figure 5: Global view of experimentation in the DOMUS Living lab for two acoustic models, M_1=*BREF120* and M_2=*BREF120_SWEETHOME_G* (Distresg calls: D, Colloquial sentence: C, Non-Speech: NS).

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.

Type	Number	Classified as speech	Classified as non-speech
Clean speech	227	196 (86.3%)	31 (13.7%)
Noisy speech	95	41 (43.2%)	54 (56.8%)
All speech	322	237 (73.6%)	85 (26.4%)

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*).

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

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):

$$\begin{cases} p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \Sigma_i), \\ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \\ w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}, \end{cases}$$

where $\mathbf{x}$ 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 $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi} and mixture weights, w_{jmi} and it has a shared number of Gaussians, I. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices Σ_i (i.e, the globally shared parameters $\breve{\Sigma}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \Sigma_i\}$) 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.

Young Speaker	WER (%) *All*	WER (%) *Emergency*	CER (%)
S01	45.0	39.1	27.8
S03	41.4	44.4	40.0
S05	19.1	15.4	14.3
S08	61.8	50.8	20.0
S09	49.4	41.2	33.3
S10	24.5	22.4	14.3
S11	21.3	17.0	16.7
S12	30.8	25.0	25.0
S14	67.0	54.8	50.0
S15	21.5	19.5	5.3
S16	14.9	11.76	7.4
S17	21.4	22.4	19.0
S18	57.7	44.9	71.4
Young (all)	34±18	30±15	**23±13**
All	36±16	32±14	**24±12**

Older Speaker	WER (%) *All*	WER (%) *Emergency*	CER (%)
S04	51.9	49.6	34.0
S06	39.2	34.3	26.3
S07	21.2	20.3	28.6
S13	45.9	43.6	23.8
Older (all)	40±13	37±13	**28±4**

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.

References

Frédéric Aman, Véronique Auberge, and Michel Vacher. 2013a. How affects can perturbe the automatic speech recognition of domotic interactions. In *Proceedings of WASSS 2013, Satellite workshop of Interspeech 2013*. ISCA, Grenoble, France, 1–5.

Frédéric Aman, Véronique Aubergé, and Michel Vacher. 2016a. Influence of expressive speech on ASR performances: application to elderly assistance in smart home. In *Proceedings of the 19th International Conference on Text, Speech, and Dialogue (TSD 2016)*, Petr Sojka, Ales Horak, Ivan Kopecek, and Karel Pala (Eds.). Lecture Notes in Computer Science, Vol. 9924. Springer International Publishing, Brno, Czech Republic, 522–530. `https://doi.org/10.1007/978-3-319-45510-5_60`

Frédéric Aman, Michel Vacher, François Portet, William Duclot, and Benjamin Lecouteux. 2016b. CirdoX: an on/off-line multisource speech and sound analysis software. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Portorož, Slovenia, 1978–1985.

Frédéric Aman, Michel Vacher, Solange Rossato, and François Portet. 2013b. Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences. In *Proceedings of the 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*. IEEE, Cluj-Napoca, Romania, 177–184.

Gretchen Anderson. 2010. Loneliness Among Older Adults: A National Survey of Adults 45+. AARP Research. `https://doi.org/10.26419/res.00064.001` Accessed: 2018/02/28.

Akira Baba, Shinichi Yoshizawa, Miichi Yamada, Akinobu Lee, and Kiyohiro Shikano. 2004. Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan, Part 2* 87 (2004), 49–57.

Jaco Badenhorst and Febe de Wet. 2017. The limitations of data perturbation for ASR of learner data in under-resourced languages. In *Proceedings of the 2017 PRASA-RobMech International Conference*. Pattern Recognition Association of South Africa, Bloemfontein, South Africa, 44–49.

Fabio Bagalà, Clemens Becker, Angello Cappello, Lorenzo Chiari, Kamiar Aminian, Jeffrey M. Hausdorf, Wiebren Zijlstra, and Jochen Klenk. 2012. Evaluation of Accelerometer-Based Fall Detection Algorithms on Real-World Falls. *PLoS ONE* 7, 5 (2012), 1–9. doi:10.1371/journal.pone.0037062.

Frédéric Bloch, V. Gautier, Norbert Noury, Jean Éric Lundy, J. Poujaud, Y.E. Claessens, and Anne-Sophie Rigaud. 2011. Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects. *Annals of Physical and Rehabilitation Medicine* 54 (2011), 391–398.

Marc-Eric Bobillier Chaumon, Salima Bekkadja, Florence Cros, and Bruno Cuvillier. 2014. The user-centered design of an ambient technology for preventing falls at home. *Gerontechnology* 13, 2 (2014), 169.

Marc-Eric Bobillier-Chaumon, Bruno Cuvillier, Salima Body, and Florence Cros. 2016. Detecting Falls at Home: User-centered Design of a Pervasive Technology. *Human Technology* 12, 2 (2016), 165–192. https://doi.org/10.17011/ht/urn.201611174654

Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier. 2005. ALIZE, a free toolkit for speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Vol. 1. IEEE, Philadelphia, PA, USA, 737–740. https://doi.org/10.1109/ICASSP.2005.1415219

Saïda Bouakaz, Michel Vacher, Marc-Eric Bobillier-Chaumon, Frédéric Aman, Salima Bekkadja, François Portet, Erwan Guillou, Solange Rossato, Elodie Desserée, Pierre Traineau, Jean-Paul Vimon, and Thierry Chevalier. 2014. CIRDO: Smart companion for helping elderly to live at home for longer. *IRBM* 35, 2 (March 2014), 101–108. https://doi.org/10.1016/j.irbm.2014.02.011

John Brown, Bonifaz Kaufmann, Florian Bacher, Christophe Sourisse, and Martin Hitz. 2013. "Oh, I Say, Jeeves!" A Calm Approach to Smart Home Input. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Lecture Notes in Computer Science, Vol. 7947. Springer, Maribor, Slovenia, 265–274. https://doi.org/10.1007/978-3-642-39146-0_23

Zoraida Callejas and Ramôn López-Cózar. 2009. Designing Smart Home Interfaces for the Elderly. *SIGACCESS Newsletter* 95 (2009), 10–16.

Inigo Casanueva, Heidi Christensen, Thomas Hain, and Philip Green. 2014. Adaptive Speech Recognition and Dialogue Management for Users with Speech Disorders. In *Proceedings of Interspeech 2014*. ISCA, Singapore, 1033–1037.

Pedro Chahuara, Anthony Fleury, François Portet, and Michel Vacher. 2016. On-line Human Activity Recognition from Audio and Home Automation Sensors: comparison of sequential and non-sequential models in realistic Smart Homes. *Journal of Ambient Intelligence and Smart Environments, Human-Centric Computing and Intelligent Environments* 8, 4 (2016), 399–422.

Marie Chan, Daniel Estève, Christophe Escriba, and Eric Campo. 2008. A review of smart homes– Present state and future challenges. *Computer Methods and Programs in Biomedicine* 91, 1 (2008), 55–81.

Jean-Michel Charpin and Cécile Tlili. 2011. *Perspectives démographique et financières de la dépendance, Rapport du groupe de travail sur la prise en charge de la dépendance*. Technical Report. French Government Health Ministry. Retrieved from https://www.ladocumentationfrancaise.fr/rapports-publics/114000333/index.shtml.

CSA. 2003. Les Français et la dépendance. http://www.csa.eu/fr/s26/nos-\-sondages-\-publies.aspx. Accessed: 2013/03/12.

CSA. 2010. Les français face à la dépendance des personnes agées. http://www.csa.eu/fr/s26/nos-sondages-publies.aspx. Accessed: 2015/11/10.

George Demiris, Marilyn Rantz, Myra Aud, Karen Marek, Hary Tyrer, Marjorie Skubic, and Ali Hussam. 2004. Older adults' attitudes towards and perceptions of "smart home" technologies: a pilot study. *Medical Informatics and the Internet in Medicine* 29, 2 (2004), 87–94.

Michel Duée and Cyril Rebillard. 2006. La dépendance des personnes âgées : une projection en 2040. In *Données sociales - La société française*. INSEE, Paris, France, 613–619.

Anthony Fleury, Michel Vacher, and Norbert Noury. 2010. SVM-Based Multi-Modal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms and First Experimental Results. *IEEE TITB* 14, 2 (2010), 274 –283.

Daniel Forman, Aaron Berman, Carolyn McCabe, Donald Baim, and Jeanne Wei. 1992. PTCA in the elderly: The "young-old" versus the "old-old"". *Journal of the American Geriatrics Society* 40, 1 (1992), 19–22.

Mark Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language* 12, 2 (1998), 75–98.

Sylvain Galliano, Edouard Geoffrois, Guillaume Gravier, Jean-François Bonastre, Djamel Mostefa, and Khalid Choukri. 2006. Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*. ELDA, Genoa, Italy, 315–320.

Jort Gemmeke, Bart Ons, Netsanet Tessema, Hugo Van hamme, Janneke van de Loo, Guy De Pauw, Walter Daelemans, Jonathan Huyghe, Jan Derboven, Lode Vuegen, Bert Van Den Broeck, Peter Karsmakers, and Bart Vanrumste. 2013. Self-taught assistive vocal interfaces: an overview of the ALADIN project. In *Proceedings of Interspeech 2013*. ISCA, Lyon, France, 2039–2043.

Kallirroi Georgila, Maria Wolters, Johanna Moore, and Robert Logie. 2010. The MATCH Corpus: A Corpus of Older and Younger Users' Interactions with Spoken Dialogue Systems. *Language Resources and Evaluation* 44, 3 (March 2010), 221–261. `https://doi.org/10.1007/s10579-010-9118-8`

Matteo Gerosa, Giuliani Diego, and Fabio Brugnara. 2009. Towards age-independent acoustic modeling. *Speech Communication* 51(6) (2009), 499–509.

Mary Gorham-Rowan and Jacqueline Laures-Gore. 2006. Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders* 39 (2006), 171–184.

Melinda Hamill, Vicky Young, Jennifer Boger, and Alex Mihailidis. 2009. Development of an automated speech recognition interface for personal emergency response systems. *Journal of NeuroEngineering and Rehabilitation* 6, 1 (2009), 26.

B. Heinbüchner, Martin Hautzinger, Clemens Becker, and Klaus Pfeiffer. 2010. Satisfaction and use of personal emergency response systems. *Zeitschrift für Gerontologie und Geriatrie* 43, 4 (2010), 219–223.

Virginia Hessels, Glenn Le Prell, and William Mann. 2011. Advances in Personal Emergency Response and Detection Systems. *Assistive Technology* 23, 3 (2011), 152–161.

Institut National pour la Prévention et l'Education à la Santé. 2006. Mieux prévenir les chutes chez les personnes âgées [Preventing falls of the elderly people]. *La Santé de l'homme* 381 (2006), 22–44.

Dan Istrate, Eric Castelli, Michel Vacher, Laurent Besacier, and Jean-François Serignat. 2006. Information Extraction From Sound for Medical Telemonitoring. *Information Technology in Biomedicine, IEEE Transactions* 10, 2 (April 2006), 264–274.

Min-Soo Kang, Kyung Mi Kim, and Hee-Cheol Kim. 2006. A Questionnaire Study for the Design of Smart Homes for the Elderly. In *Proceedings of Healthcom 2006*. IEEE, New Delhi, India, 265–268.

Maarit Kangas, Irene Vikman, Lars Nyberg, Raija Korpelainen, J. Lindblom, and Timo Jämsä. 2012. Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects. *Gait & Posture* 35, 3 (2012), 500–505.

Tiiu Koskela and Kaisa Väänänen-Vainio-Mattila. 2004. Evolution towards smart home environments: empirical evaluation of three user interfaces. *Personal and Ubiquitous Computing* 8 (2004), 234–240.

Lori Lamel, Jean-Luc Gauvain, and Maxine Eskénazi. 1991. BREF, a Large Vocabulary Spoken Corpus for French. In *Proceedings of EUROSPEECH 91*, Vol. 2. ISCA, Geneva, Switzerland, 505–508.

Benjamin Lecouteux, Michel Vacher, and François Portet. 2011. Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions. In *Proceedings of InterSpeech 2011*. ISCA, Florence, Italy, 2273–2276.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics-Doklady* 10 (1966), 707–710.

William Mann, Patrícia Belchior, Machiko Tomita, and Bryan Kemp. 2005. Use of Personal Emergency Response Systems by Older Individuals With Disabilities. *Assistive Technology* 17, 1 (2005), 82–88.

Andrew Cameron Morris, Viktoria Maier, and Philip Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.. In *Proceedings of Interspeech 2004/ICLSP*. ISCA, Jeju Island, Korea, 1–4.

Mumtaz Begum Mustafa, Fadhilah Rosdi, Siti Salwah Salim, and Muhammad Umair Mughal. 2015. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Systems with Applications* 42, 8 (2015), 3924 – 3932. https://doi.org/10.1016/j.eswa.2015.01.033

Frans Mäyrä, Anne Soronen, Jukka Vanhala, Jussi Mikkonen, Mari Zakrzewski, Ilpo Koskinen, and Kristo Kuusela. 2006. Probing a Proactive Home: Challenges in Researching and Designing Everyday Smart Environments. *Human Technology* 2 (2006), 158–186.

Samuel Nyman and Christina Victor. 2014. Use of personal call alarms among community-dwelling older people. *Ageing & Society* 34, 1 (2014), 67–89.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proceedings of the 4th Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing*. Reykjavik, Island, 24–30.

OHCHR. 2018. Human rights of older persons. http://www.ohchr.org/EN/Issues/OlderPersons/. Accessed: 2018/03/09.

Kirsten Peetoom, Monique Lexis, Manuela Joore, Carmen Dirksen, and Luc De Witte. 2015. Literature review on monitoring technologies and their outcomes in independently living elderly people. *Disability and Rehabilitation: Assistive Technology* 10, 4 (2015), 271–294. https://doi.org/10.3109/17483107.2014.961179 arXiv:https://doi.org/10.3109/17483107.2014.961179

François Portet, Heidi Christensen, Frank Rudzicz, and Jan Alexandersson. 2015. Perspectives on Speech and Language Interaction for Daily Assistive Technology: Overall Introduction to the Special Issue Part 3. *TACCESS* 7, 2 (2015), 4:1–4:8.

François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. 2013. Design and evaluation of a smart home voice interface for the elderly – Acceptability and objection aspects. *Personal and Ubiquitous Computing* 17, 1 (2013), 127–144. https://doi.org/10.1007/s00779-011-0470-5

Daniel Povey, Lukáš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas. 2011a. The subspace Gaussian mixture model—A structured model for speech recognition. *Computer Speech & Language* 25, 2 (2011), 404 – 439. https://doi.org/10.1016/j.csl.2010.06.003

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukás Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Peter Schwarz, et al. 2011b. The Kaldi Speech Recognition Toolkit. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*. IEEE, Hawaii, USA, 1–4.

Emanuele Principi, Stefano Squartini, Roberto Bonfigli, Giacomo Ferroni, and Francesco Piazza. 2015. An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Systems with Applications* 42, 13 (2015), 5668 – 5683. https://doi.org/10.1016/j.eswa.2015.02.036

Mirco Ravanelli and Maurizio Omologo. 2014. On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In *Proceedings of Interspeech 2014*. ISCA, Singapore, 1028–1032.

Carole-Anne Rivière and Amandine Brugière. 2010. *Bien vieillir grâce au numérique - Qualité de vie, autonomie, lien social*. FYP éditions, France.

Royal Society for the Prevention of Accidents RoSPA. 2016. Older People Safety. https://www.rospa.com/home-safety/advice/older-people. Accessed: 2018/02/28.

Rudy Rotili, Emanuele Principi, Stefano Squartini, and Björn Schuller. 2013. A Real-Time Speech Enhancement Framework in Noisy and Reverberated Acoustic Scenarios. *Cognitive Computation* 5, 4 (2013), 504–516.

Klaus Scherer, Tom Johnstone, and Gundrun Klasmeyer. 2003. *Vocal expression of emotion*. Oxford University Press, U.K., 433–456.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication* 53, 9-10 (Nov. 2011), 1062–1087. https://doi.org/10.1016/j.specom.2011.01.011

Mohammed Sehili, Dan Istrate, Bernadette Dorizzi, and Jerome Boudy. 2012. Daily sound recognition using a combination of GMM and SVM for home automation. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO 2012)*. EURASIP, Bucharest, Romania, 1673–1677.

Paul Simpson, Jason Bendall, Anne Tiedemann, Stephen Lord, and Jacqueline Close. 2014. Epidemiology of Emergency Medical Service Responses to Older People Who Have Fallen: A Prospective Cohort Study. *Prehospital Emergency Care* 18, 2 (2014), 185–194.

Randi Stokke. 2016. The Personal Emergency Response System as a Technology Innovation in Primary Health Care Services: An Integrative Review. *Journal of Medical Internet Research* 18, 7 (2016), e187.

Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark Plumbley. 2015. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia* 17, 10 (2015), 1733–1746.

Friederike Thilo, Barbara Hürlimann, Sabine Hahn, Selina Bilger, Jos Schols, and Ruud Halfens. 2016. Involvement of older people in the development of fall detection systems: a scoping review. *BMC Geriatrics* 16, 42 (2016), 1–9.

Michel Vacher, Saïda Bouakaz, Marc-Eric Bobillier Chaumon, Frédéric Aman, Rezza Khan, and Salima Bekkadja. 2016. The CIRDO corpus: comprehensive audio/video database of domestic falls of elderly people. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Portorož, Slovenia, 1389–1396.

Michel Vacher, Sybille Caffiau, François Portet, Brigitte Meillon, Camille Roux, Elena Elias, Benjamin Lecouteux, and Pedro Chahuara. 2015a. Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing* 7, issue 2 (2015), 5:1–5:36. `https://doi.org/10.1145/2738047`

Michel Vacher, Benjamin Lecouteux, Frédéric Aman, Solange Rossato, and François Portet. 2015b. Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home. In *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies*. SIG-SLPAT, ACL/ISCA, Dresden, Germany, 1–7.

Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon, and Nicolas Bonnefond. 2014. The Sweet-Home speech and multimodal corpus for home automation interaction. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*. ELRA, Reykjavik, Iceland, 4499–4506.

Michel Vacher, Benjamin Lecouteux, and François Portet. 2012. Recognition of Voice Commands by Multisource ASR and Noise Cancellation in a Smart Home Environment. In *Proceedings of the European Signal Processing Conference (EUSIPCO 2012)*. EURASIP, Bucarest, Romania, 1663–1667.

Michel Vacher, François Portet, Anthony Fleury, and Norbert Noury. 2011. Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges. *International journal of E-Health and medical communications* 2, 1 (2011), 35–54.

Michel Vacher, Jean-François Serignat, Stéphane Chaillol, Dan Istrate, and Vladimir Popescu. 2006. Speech and Sound Use in a Remote Monitoring System for Healthcare. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD 2006)*. Lecture Notes in Computer Science, Vol. 4188. Springer-Verlag, Brno, Czech Republic, 711–718. https://doi.org/10.1007/11846406_89

Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. 2013. The second CHiME Speech Separation and Recognition Challenge: Datasets, tasks and baselines. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*. IEEE, Vancouver, Canada, 126–130.

Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. 2017. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech and Language* 46 (2017), 535–557.

Ravichander Vipperla, Steve Renals, and Joe Frankel. 2008. Longitudinal study of ASR performance on ageing Voices. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*. ISCA, Brisbane, Australia, 2550–2553.

Bogdan Vlasenko, Dmytro Prylipko, and Andreas Wendemuth. 2012. Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. In *Proceedings of the 35th Annual German Conference on AI, Advances in Artificial Intelligence (KI 2012)*, Antonio Krüger Birte Glimm (Ed.). Lecture Notes in Computer Science, Vol. 7526. Springer Heidelberg, Saarbrücken, Germany, 103–107.

Matthias Woelfel and John McDonough. 2009. *Distant Speech Recognition*. John Wiley & Sons, U.K.

Victoria Young and Alex Mihailidis. 2013. The CARES corpus: a database of older adult actor simulated emergency dialogue for developing a personal emergency response system. *International Journal of Speech Technology* 16, 1 (01 Mar 2013), 55–73.

Victoria Young, Elizabeth Rochon, and Alex Mihailidis. 2016. Exploratory analysis of real personal emergency response call conversations: considerations for personal emergency response spoken dialogue systems. *Journal of NeuroEngineering and Rehabilitation* 13, 1 (14 Nov 2016), 97.

Leila Zouari and Gérard Chollet. 2006. Efficient Gaussian Mixture for Speech Recognition. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, Vol. 4. IAPR, Hong Kong, 294–297. https://doi.org/10.1109/ICPR.2006.475