Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home

Michel Vacher, Benjamin Lecouteux, Frédéric Aman, Solange Rossato, François Portet

Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home

Michel Vacher[1], *Benjamin Lecouteux*[2], *Frédéric Aman*[1],
Solange Rossato[2], *François Portet*[2]

[1]CNRS, LIG, F-38000 Grenoble, France
[2]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
41 rue Mathématiques, BP 53, 38041 Grenoble cedex9, France
Michel.Vacher@imag.fr, Benjamin.Lecouteux@imag.fr, Frederic.Aman@imag.fr,
Solange.Rossato@imag.fr, Francois.Portet@imag.fr

Abstract

This paper presents a system to recognize distress speech in the home of seniors to provide reassurance and assistance. The system is aiming at being integrated into a larger system for Ambient Assisted Living (AAL) using only one microphone with a fix position in a non-intimate room. The paper presents the details of the automatic speech recognition system which must work under distant speech condition and with expressive speech. Moreover, privacy is ensured by running the decoding on-site and not on a remote server. Furthermore the system was biased to recognize only set of sentences defined after a user study. The system has been evaluated in a smart space reproducing a typical living room where 17 participants played scenarios including falls during which they uttered distress calls. The results showed a promising error rate of 29% while emphasizing the challenges of the task.

Index Terms: Smart home, Vocal distress call, Applications of speech technology for Ambient Assisted Living

1. Introduction

Life expectancy has increased in all countries of the European Union in the last decade. Therefore the part of the people who are at least 75 years old has strongly increased and solutions are needed to satisfy the wishes of elderly people to live as long as possible in their own homes. Ageing can cause functional limitations that –if not compensated by technical assistance or environmental management– lead to activity restriction [1][2]. Smart homes are a promising way to help elderly people to live independently at their own home, they are housings equipped with sensors and actuators [3][4][1][5]. Another aspect is the increasing risk of distress, among which falling is one of the main fear and lethal risk, but also blocking hip or fainting. The most common solution is the use of kinematic sensors worn by the person [6] but this imposes some constraints in the everyday life and worn sensors are not always a good solution because some persons can forget or refuse to wear it. Nowadays, one of the best suited interfaces is the voice-user interface (VUI), whose technology has reached maturity and is avoiding the use of worn sensors thanks to microphones set up in the home and allowing hands-free and distant interaction [7]. It was demonstrated that VUI is useful for system integrating speech commands [8].

The use of speech technologies in home environment requires to address particular challenges due to this specific environment [9]. There is a rising number of smart home projects considering speech processing in their design. They are related to wheelchair command [10], vocal command for people with dysarthria [11][8], companion robot [12], vocal control of appliances and devices [13]. Due to the experimental constraints, few systems were validated with real users in realistic situation condition like in the SWEET-HOME project [14] during which a dedicated voice based home automation system was able to drive a smart home thanks to vocal commands with typical people [15] and with elderly and visually impaired people [16].

In this paper we present an approach to provide assistance in a smart home for seniors in case of distress situation in which they can't move but can talk. The challenge is due to expressive speech which is different from standard speech: is it possible to use state of the art ASR techniques to recognize expressive speech? In our approach, we address the problem by using the microphone of a home automation and social system placed in the living room with ASR decoding and voice call matching. In this way, the user must be able to command the environment without having to wear a specific device for fall detection or for physical interaction (e.g., a remote control too far from the user when needed). Though microphones in a home is a real breach of privacy, by contrast to current smart-phones, we address the problem using an in-home ASR engine rather than a cloud based one (private conversations do not go outside the home). Moreover, the limited vocabulary ensures that only speech relevant to the command of the home is correctly decoded. Finally, another strength of the approach is to have been evaluated in realistic conditions. The paper is organised as follow. Section 2 presents the method for speech acquisition and recognition in the home. Section 3, presents the experimentation and the results which are discussed in Section 5.

2. Method

The distress call recognition is to be performed in the context of a smart home which is equipped with e-lio[1], a dedicated system for connecting elderly people with their relatives as shown in Figure 1. e-lio is equipped with one microphone for video conferencing. The typical setting and the distress situations were determined after a sociological study conducted by the GRePS laboratory [17] in which a representative set of seniors were included.

From this sociological study, it appears that this equipment

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

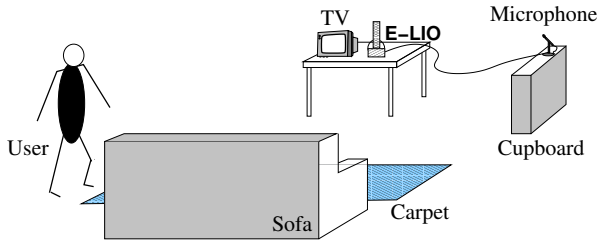

Figure 1: Microphone position in the smart home

is set on a table in the living room in font of the sofa. In this way, an alert could be given if the person falls due to the carpet or if it can't stand up from the sofa. This paper presents only the audio part of the study, for more details about the global audio and video system, the reader is referred to [18].

2.1. Speech analysis system

The audio processing was performed by the software CIRDOX[19] whose architecture is shown in Figure 2. The microphone stream is continuously acquired and sound events are detected on the fly by using a wavelet decomposition and an adaptive thresholding strategy [20]. Sound events are then classified as noise or speech and, in the latter case, sent to an ASR system. The result of the ASR is then sent to the last stage which is in charge of recognizing distress calls.

In this paper, we focus on the ASR system and present different strategies to improve the recognition rate of the calls. The remaining of this section presents the methods employed at the acoustic and decoding level.

2.2. Acoustic modeling

The Kaldi speech recognition tool-kit [21] was chosen as ASR system. Kaldi is an open-source state-of-the-art ASR system with a high number of tools and a strong support from the community. In the experiments, the acoustic models were context-dependent classical three-state left-right HMMs. Acoustic features were based on Mel-frequency cepstral coefficients, 13 MFCC-features coefficients were first extracted and then expanded with delta and double delta features and energy (40 features). Acoustic models were composed of 11,000 context-dependent states and 150,000 Gaussians. The state tying is performed using a decision tree based on a tree-clustering of the phones. In addition, off-line fMLLR linear transformation acoustic adaptation was performed.

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 SH corpus (SWEET-HOME) [22] which consists of records of 60 speakers interacting in the smart home and from 28 minutes of the Voix-détresse corpus [23] which is made of records of speakers eliciting a distress emotion.

2.2.1. Subspace GMM Acoustic Modelling

The GMM and Subspace GMM (SGMM) both model emission probability of each HMM state with a Gaussian mixture model, but 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 [24] is described in the following equations:

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

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 to 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 $\mathbf{\Sigma}_i$ i.e; the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ 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.

Our experiments aims at obtaining SGMM shared parameters using both SWEET-HOME data (7h), Voix-détresse (28mn) and clean data (ESTER+REPERE 500h). Regarding the GMM part, the three training data set are just merged in a single one. [24] showed that the model is also effective with large amounts of training data. Therefore, three UBMs were trained respectively on SWEET-HOME data, Voix-détresse and clean data. These tree UBMs contained 1K gaussians and were merged into a single one mixed down to 1K gaussian (closest Gaussians pairs were merged [25]). The aim is to bias specifically the acoustic model with the smart home and expressive speech conditions.

2.3. Recognition of distress calls

The recognition of distress calls consists in computing the phonetic distance of an hypothesis to a list of predefined distress calls. Each ASR hypothesis H_i is phonetized, every voice commands T_j is aligned to H_i using Levenshtein distance. The deletion, insertion and substitution costs were computed empirically while the cumulative distance $\gamma(i,j)$ between H_j and T_i is given by Equation 1.

$$\gamma(i,j) = d(T_i, H_j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (1)$$

The decision to select or not a detected sentence is then taken according a detection threshold on the aligned symbol score (phonems) of each identified call. This approach takes into account some recognition errors like word endings or light variations. Moreover, in a lot of cases, a miss-decoded word is phonetically close to the good one (due to the close pronunciation). From this the CER (Call Error Rate i.e., distress call error rate) is defined as:

$$\text{CER} = \frac{\text{Number of missed calls}}{\text{Number of calls}} \quad (2)$$

This measure was chosen because of the content of the corpus Cirdo-set used in this study. Indeed, this corpus is made of sentences and interjections. All sentences are calls for help, without any other kind of sentences like home automation orders or colloquial sentences, and therefore it is not possible to determine a false alarm rate in this framework.

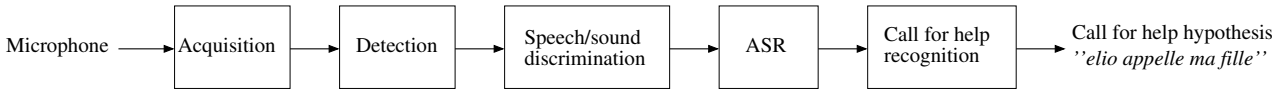

Figure 2: Architecture of the analysis system CIRDOX

Figure 3: A young participant playing a fall scenario

3. Experimentation and results

3.1. Live Experiment

An experiment was run in the experimental platform of the LIG laboratory in a room whose setting corresponds to Figure 1 and equipped with a sofa, a carpet, 2 chairs, a table and e-lio. A Sennheiser SKM 300 G2 ME2 omnidirectional microphone was set on the cupboard. In these conditions, the microphone was at a distance of above 2 meters from the speaker (Distant speech conditions). The audio analysis system consisted in the CIRDOX software presented in Section 2 which was continuously recording and analysing the audio streams to detect the calls.

3.1.1. Scenarios and experimental protocol

The scenarios were elaborated after field studies made by the GRePS laboratory [17]. These studies allowed to specify the fall context, the movements during the fall as well as the person's reaction once on the floor. Phrases uttered during and after the fall were also identified "*Blast! What's happening to me? Oh shit, shit!*". The protocol was as follows [18]. Each participant was introduced to the context of the research and was invited to sign a consent form. The participants played four scenarios of fall, one blocked hip scenario and two other scenarios called "true-false" added to challenge the automatic detection of falls by the video analysis system. If the participant's age was under 60, he wore a simulator which hampered his mobility and reduced his vision and hearing to simulate aged physical conditions. Figure 3 shows a young participant wearing the simulator at the end of a fall scenario. The average experiment duration of an experiment was 2h 30min per person. This experiment was very tiring for the participants and it was necessary to include rehearsals before starting the recordings so that the participant felt comfortable and was able to fall securely.

3.1.2. Voice commands and distress calls

The sentences of the AD80 corpus [19] served as basis to develop the language model used by our system. This corpus was recorded by 43 elderly people and 52 non-aged pepole in our laboratory and in a nursing home to study the automatic recognition of speech uttered by aged speakers. This corpus is made of 81 casual sentences, 31 vocal commands for home automation and 58 distress sentences. An excerpt of these sentences in French is given Table 2, the distress sentences identified in the field study reported in section 3.1.1 were included in the corresponding part of AD80.

The utterance of some of these distress sentences were integrated into the scenarios with the exception of the two "true-false" scenarios.

3.1.3. Acquired data: Cirdo-set

In this paper we focus on the detection of the distress calls, therefore we don't consider the audio event detected and analyzed on the fly but only the full records of each scenario. These data sets were transcribed manually using transcriber [26] and the speech segments were then extracted for analysis.

The targeted participants were elderly people that were still able to play the fall scenarios securely. However, the recruitment of such kind of population was very difficult and a part of the participants was composed of people under 60 years old but they were invited to wear a special suit [18] which hampered their mobility and reduced their vision but without any effect on speech production. Overall, 17 participants were recruited (9 men and 8 women). Among them, 13 participants were under 60 and worn the simulator. The aged participants were between 61 and 83 years old.

When they played the scenarios, some participants produced sighs, grunts, coughs, cries, groans, pantings or throat clearings. These sounds were not considered during the annotation process. In the same way, speeches mixed with sound produced by the fall were ignored. At the end, each speaker uttered between 10 and 65 short sentences or interjections ("*ah*", "*oh*", "*aïe*", etc.) as shown Table 1.

Sentences were often close of those identified during the field studies ("*je peux pas me relever* - I can't get up", "*e-lio appelle du secours* - e-lio call for help", etc.), some were different ("*oh bein on est bien là tiens* - oh I am in a sticky situation"). In practice, participants cut some sentences (i.e., inserted a delay between "*e-lio*" and "*appelle ma fille* - call my daughter"), uttered some spontaneous sentences, interjections or non-verbal sounds (i.e., groan).

3.2. Off line experiments

The methods presented in Section 2 were run on the Cirdo-set corpus presented in Section 3.1.3.

The SGMM model presented in Section 2.2 was used as acoutic model. The *generic language model* (LM) was estimated from French newswire collected in the Gigaword corpus. It was 1- gram with 13,304 words. Moreover, 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 3.1.1, with 99 1-gram, 225 2-gram and 273 3-gram models. Finally, the lan-

Distress Sentence	*Home Automation Command*	*Casual Sentence*
Aïe aïe aïe ⋆	Appelle quelqu'un e-lio ⋆	Bonjour madame
Oh là ⋆	e-lio, appelle quelqu'un ⋆	Ça va très bien
Merde ⋆	e-lio tu peux appeler une ambulance	Où sont mes lunettes
Je suis tombé ⋆	e-lio tu peux téléphoner au SAMU	Le café est brûlant
Je peux pas me relever ⋆	e-lio, appelle du secours	J'ai ouvert la porte
Qu'est-ce qu'il m'arrive ⋆	e-lio appelle les secours	Je me suis endormi tout de suite
Aïe ! J'ai mal ⋆	e-lio appelle ma fille	Il fait soleil
Oh là ! Je saigne ! Je me suis blessé ⋆	e-lio appelle les secours	Ce livre est intéressant
Aidez-moi	e-lio appelle le SAMU !	Je dois prendre mon médicament
Au secours	e-lio appelle les pompiers !	J'allume la lumière

Table 2: Examples of sentences of the AD80 corpus (⋆ denotes a sentence identified during the sociological study)

Spk.	Age	Sex	Nb. of interjections or short sentences	
			All	*Distress*
S01	30	M	22	14
S02	-	-	-	
S03	24	F	16	15
S04	83	F	65	53
S05	29	M	24	21
S06	64	F	23	19
S07	61	M	23	21
S08	44	M	25	15
S09	16	M	32	21
S10	16	M	19	15
S11	52	M	12	12
S12	28	M	15	12
S13	66	M	24	21
S14	52	F	23	21
S15	23	M	20	19
S16	40	F	29	27
S17	40	F	24	21
S18	25	F	17	14
Total	40.76		413	341

Table 1: Composition of the audio corpus Cirdo-set

guage 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). This combination has been shown as leading to the best WER for domain specific application [27]. The interest of such combination is to bias the recognition towards the domain LM but when the speaker deviates from the domain, the general LM makes it possible to avoid the recognition of sentences leading to "false-positive" detection.

Results on manually annotated data are given Table 3. The most important performance measures are the Word Error Rate (WER) of the overall decoded speech and those of the specific distress calls as well as the Call Error Rate (CER: c.f. equation 2). Considering distress calls only, the average WER is 34.0% whereas it a 39.3% when all interjections and sentences are taken into account.

Unfortunately and as mentionned above, the used corpus doesn't allow the détermine a False Alarm Rate. Previous studies based on the AD80 corpus showed recall, precision and F-measure equal to 88.4%, 86.9% and 87.2% [19]. Nevertheless, this corpus was recorded in very different conditions, text reading in a studio, in contrary of those of Cirdo-set.

Spk.	WER (%)		CER (%)	Spk.	WER (%)		CER (%)
	All	*Distress*			*All*	*Distress*	
S01	45.0	39.1	27.8	S11	21.3	17.0	16.7
S03	41.4	44.4	40.0	S12	30.8	25.0	25.0
S04	51.9	49.6	34.0	S13	45.9	43.6	23.8
S05	19.1	15.4	14.3	S14	67.0	54.8	50.0
S06	39.2	34.3	26.3	S15	21.5	19.5	5.3
S07	21.2	20.3	28.6	S16	14.9	11.76	7.4
S08	61.8	50.8	20.0	S17	21.4	22.4	19.0
S09	49.4	41.2	33.3	S18	57.7	44.9	71.4
S10	24.5	22.4	14.3	All	39.3	34.0	26.8

Table 3: Word and Call Error Rate for each participant

On average, CER is equal to 26.8% with an important disparity between the speakers.

4. Discussion

These results are quite different from those obtained with the AD80 corpus (with aged speakers and speaker adaptation): WER was 14.5% [19]. There are important differences between the recording conditions used for AD80 and for the Cirdo-set corpus used in our study that can explain this performance gap:

- AD80 is made of readings by speakers sitting in comfortable position in front of a PC and the microphone ;
- AD80 was recorded in nearest conditions in comparison with distant setting for Cirdo-set ;
- Cirdo-set was recorded by participants who fell on the floor or that are blocked on the sofa. They were encouraged to speak in the same way that they would speak if they would be really put in these situations. Obviously, we obtained expressive speech, but there is no evidence that the pronunciation would be the same as in real conditions of a fall or a blocked hip.

Regarding the CER, its global value 26.8% shows that 74.2% of the calls were correctly recognized ; furthermore, at the exception of one speaker (CER=71.4%), CER is always below 50% consequently more than 50% of the calls were recognized. For 6 speakers, CER was below 20%. This suggests that a distress 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 distress call because the person's voice is altered by the stress, it is likely that this person will fill more and more

stress and as a consequence future calls would be more difficult to identify. In a same way, our corpus was recorded in realistic conditions but not in real conditions and frail elderly people may not be adequately simulated by healthy human adults. A relatively small number of missed distress calls could render the system unacceptable for use amongst the potential user and therefore some efforts in this regard would need to be pursued.

5. Conclusion and perspectives

This study is focused on the framework of automatic speech recognition applications in smart homes, that is in distant speech conditions and especially in realistic conditions very different from those of corpus recording when the speaker is reading a text.

Indeed in this paper, we presented the Cirdo-set corpus made of distress calls recorded in distant speech conditions and in realistic conditions in case of fall or blocked hip. The WER obtained at the output of the dedicated ASR was 36.3% for the distress calls. Thanks to a filtering of the ASR hypothesis at phonetic level, more than 70% of the calls were detected.

These results obtained in realistic conditions gives a fairly accurate idea of the performances that can be achieved with state of the art ASR systems for end user and specific applications. They were obtained in the particular case of the recognition of distress calls but they can be extended to other applications in which expressive speech may be considered because it is inherently present.

As stated above, obtained results are not sufficient to allow the system use in real conditions and two research ideas can be considered. Firstly, speech recognition performances may be improved thanks to acoustic models adapted to expressive speech. This may be achieved to the record of corpora in real conditions but this is a very difficult task. Secondly, it may be possible to recognize the repetition, at regular intervals, of speech events that are phonetically similar. This last method does not request the good recognition of the speech. Our future studies will address this problem.

6. Acknowledgements

This work is part of the CIRDOproject founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-10-TECS-012). The authors would like to thank the persons who agreed to participate in the recordings.

7. References

[1] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. De Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, pp. 1–24, 2014.

[2] L. C. D. Silva, C. Morikowa, and I. M. Petra, "State of the art of smart homes," *Engineering Applications of Artificial Intelligence*, no. 25, pp. 1313–1321, 2012.

[3] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes- present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[4] L. De Silva, C. Morikawa, and I. Petra, "State of the art of smart homes," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 7, pp. 1313–1321, 2012.

[5] Q. Ni, A. B. García Hernando, and I. P. de la Cruz, "The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," *Sensors*, vol. 15, no. 5, pp. 11 312–11 362, 2015.

[6] F. Bloch, V. Gautier, N. Noury, J. Lundy, J. Poujaud, Y. Claessens, and A. Rigaud, "Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects," *Annals of Physical and Rehabilitation Medicine*, vol. 54, pp. 391–398, 2011.

[7] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[8] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition," in *4th Workshop on Speech and Language Processing for Assistive Technologies*, 2014.

[9] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges," *International Journal of E-Health and Medical Communications*, vol. 2, no. 1, pp. 35–54, 2011.

[10] W. Li, J. Glass, N. Roy, and S. Teller, "Probabilistic dialogue modeling for speech-enabled assistive technology," in *SLPAT 2013*, 2013, pp. 67–72.

[11] J. F. Gemmeke, B. Ons, N. Tessema, H. Van Hamme, J. Van De Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: an overview of the ALADIN project," in *Interspeech 2013*, 2013, pp. 2039–2043.

[12] P. Milhorat, D. Istrate, J. Boudy, and G. Chollet, "Hands-free speech-sound interactions at home," in *EUSIPCO 2012*, Aug. 2012, pp. 1678 –1682.

[13] M. Matassoni, R. F. Astudillo, A. Katsamanis, and M. Ravanelli, "The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Interspeech 2014*, Sep. 2014, pp. 1613–1617.

[14] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, "Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation," *ACM Transactions on Accessible Computing, Special Issue on Speech and Language Processing for AT (Part 3)*, vol. 7, no. 2, (in press), 36 pages.

[15] M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili, and P. Chahuara, "Evaluation of a Real-Time Voice Order Recognition System from Multiple Audio Channels in a Home," in *Interspeech 2013*, Aug. 2013, pp. 2062–2064.

[16] M. Vacher, B. Lecouteux, and F. Portet, "Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment," in *Interspeech 2014*, Sep. 2014, pp. 1008–1012.

[17] M. Bobillier Chaumon, F. Cros, B. Cuvillier, C. Hem, and E. Codreanu, "Concevoir une technologie pervasive pour le maintien à

domicile des personnes âgées : la détection de chutes dans les activités quotidiennes," in *Activités Humaines, Technologies et bien-être, Congrès EPIQUE (Psychologie Ergonomique)*, Belgique - Bruxelles, July 2013, pp. 189–199.

[18] S. Bouakaz, M. Vacher, M.-E. Bobillier-Chaumon, F. Aman, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, E. Desserée, P. Traineau, J.-P. Vimon, and T. Chevalier, "CIRDO: Smart companion for helping elderly to live at home for longer," *Innovation and Research in BioMedical engineering (IRBM)*, vol. 35, no. 2, pp. 101–108, Mar. 2014.

[19] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences," in *The 7th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013*, Cluj-Napoca, Romania, Oct. 2013, pp. 177–184.

[20] M. Vacher, D. Istrate, and J. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," in *Proc. 12th European Signal Processing Conference*, S. LTD, Ed., Vienna, Austria, sep. 2004, pp. 1171–1174.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[22] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The Sweet-Home speech and multimodal corpus for home automation interaction," in *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 4499–4506.

[23] F. Aman, "Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile," Ph.D. dissertation, Université de Grenoble, Ecole doctorale MSTII, 2014.

[24] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011.

[25] L. Zouari and G. Chollet, "Efficient gaussian mixture for speech recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, 2006, pp. 294–297.

[26] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.

[27] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions," in *Proc. InterSpeech*, 2011, pp. 2273–2276.