Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences

Frédéric Aman, Michel Vacher, Solange Rossato, François Portet

Speech Recognition of Aged Voice in the AAL Context: Detection of Distress Sentences

Frédéric Aman, Michel Vacher, Solange Rossato and François Portet
Laboratoire d'Informatique de Grenoble, GETALP Team
UMR CNRS/UJF/INPG/UPMF 5217
Grenoble, France
{frederic.aman, michel.vacher, solange.rossato, francois.portet}@imag.fr

Abstract—**By 2050, about a third of the French population will be over 65. In the context of technologies development aiming at helping aged people to live independently at home, the CIRDO project aims at implementing an ASR system into a social inclusion product designed for elderly people in order to detect distress situations. Speech recognition systems present higher word error rate when speech is uttered by elderly speakers compared to when non-aged voice is considered. Two specialized corpora in French, AD80 and ERES38, were recorded in this framework by aged people, they were used first to study the possibility of adaptation of standard ASR to aged voice. Then we looked at whether the variability of the WER between speakers could be correlated with the level of dependence. Then, we assessed the performance of distress sentence detection by a filter and we demonstrated a significant drop in performance for those with the lowest degree of autonomy.**

Keywords—Speech recognition; AAL; Dependence; Elderly; Keyword detection

I. Introduction

Life expectancy has increased in all countries of the European Union in the last decade. In the beginning of 2013, France has 17.5% of its citizens of at least 65 years old. Furthermore, 9% of the people in France are at least 75 years old. The increase of life expectancy and the ageing of baby boomers are the main factors of ageing [1]. The number of dependent elderly people will increase in the coming years, with 50% more of dependent people by 2040 according to a projection conducted by the INSEE French institute [2]. At the end of 2010, 12 million people in France received the Personalized Allocation of Autonomy (APA), where 61% of them were at home and 39% were in specialized institutions [1]. A survey shows that 80% of people above 65 years old would prefer to stay living at home if they lose autonomy [3]. The notion of dependency is based on the alteration of physical, sensory and cognitive functions, on the restriction of the activities of daily living, and on the need for help or assistance. Therefore, ageing can cause functional limitations that – if not compensated by technical assistance or environmental management – lead to activities restriction. Then, the person needs the assistance of someone for regular elementary activities (nursing home services, domestic help, etc.) [4].

Some technological solutions based on robotics, automation, cognitive science, and computer networks have been developed to compensate physical or mental decline, and to provide assistance if necessary through surveillance by detecting distress and accidental falls. Some systems also facilitate social contacts, helping caregivers as well as reassuring relatives. However, elderly are often confused by complex interfaces in technological systems. Therefore, the usual interfaces (remote controls, mice or keyboards) must be complemented by more accessible and natural interfaces such as a system of Automatic Speech Recognition (ASR).

In this context, the CIRDO project[1] promotes autonomy and support for elderly people by caregivers through the social inclusion product e-lio[2] and thanks to the integration of innovative services (automatic speech recognition, analysis of situations or scenes in an uncontrolled environment) to promote independence and support for caregivers, patients with chronic diseases or Alzheimer's disease or related. In addition, this project will allow the validation of generic technologies, ergonomic and a psychological evaluation on the use of services but also critical investigation of the knowledge gained by professionals in human services.

The use of elderly voice can be an issue for performance of speech recognition system. Indeed, ageing voice is characterized by some specific features such as imprecise production of consonants, tremors and slower articulation [5]. From the anatomical point of view, some studies have shown age-related degeneration with atrophy of vocal cords, calcification of laryngeal cartilages, and changes in muscles of larynx [6], [7]. Some authors [8], [9] have reported that classical ASR systems exhibit poor performances with elderly voice because most acoustic models of ASR systems are acquired from non-aged voice samples. Privat et al. [10] studied the usability of a system of dictation in French for elderly vs. young people in various situations of vocal production, showing that certain classes of phonemes such as the front and rounded vowels were badly recognized. These few studies were relevant for their comparison between ageing voice vs. non-ageing voice on ASR performance, but their fields were quite far from our topic of automation commands recognition, and no study was done in French language.

[1] http://liris.cnrs.fr/cirdo

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

Another issue for our work was the non-existence of a speech corpus in French containing distress utterances and automation commands.

This paper presents the ASR ability to detect sentences of distress in the case of ageing voice. In Section II, we present some projects and studies related to distress detection in the Ambient Assisted Living (AAL) context with a special attention to audio analysis used in this framework. The audio processing system CirdoX is presented in Section III.A and the proposed method of detection of emergency calls in the home is described in Section III.B. The specialized corpora that we recorded and the ASR system are described in Section IV, then we show the necessity to adapt acoustic models to the aged voice in Section V and present findings of the adaptation in the same section. In Section VI, we introduce the problem of the variability of performances of the detection of emergency call for the elderly people case before discussing the results of distress sentences filtering for the different levels of elderly dependence. Section VII is related to the conclusion of this study.

II. Related works

In the Ambient Assisted Living context, a *Health Smart Home* is a habitation equipped with a set of sensors, actuators and automated devices to provide ambient intelligence for daily living task support, early detection of distress situations, remote monitoring and promotion of safety and well-being [11]. A large number of research projects have contributed to the field, among them House_n [12], Casas [13], ISpace [14], Aging in Place [15], Ger'Home [16] or Soprano [17]. The main trends of these projects are related to distress detection, health status monitoring and cognitive stimulation; a great variety of sensors are used like wearable video cameras, embedded sensors, medical sensors, switches and infrared detectors.

Microphones and audio technologies are less taken into consideration. Regarding security in the home, speech technologies can help a person in danger to call for help from anywhere in a multi-room home without having to use a touch interface that can be out of reach. For instance, the Sweet-Home project [18] takes advantage of a smart home equipped with home automation through sensors, microphones and actuators driven by an intelligent controller allowing in context interaction and detection of distress situations. The Reside and DesdHIS projects studied the detection and classification of sounds in order to send an alarm in presence of cry or glass breaking [19]. Other projects considered the fall detection using a wearable microphone which is often fused with other modalities like motion thanks to the use of an accelerometer [20], [21]. Doukas et al. proposed a patient awareness system to detect body fall and vocal stress in speech expression through analysis of motion data and acoustic, but in this case the person is constrained to wear sensors continuously [21]. Another approach proposed a dialog system in order to replace traditional emergency systems and fit the life style of the elderly [22]. However, the vocabulary of the prototype was limited (yes or no) and the system was not tested with aged users. In [23], a communicative avatar was designed to interact with a person in a smart office. This study takes advantage of beam forming for speech enhancement and for a geometric area of recording but was not tested in a multi-room realistic home. Another popular approach is the use of companion robots like in the CompanionAble project [24]. The companion robot is able to move nearby the person and then the user is close to the avatar. The system is able to recognize commands uttered in Dutch, no specific application to distress sentence recognition were reported by the authors. This system is limited by the capacity of the robot to move close to the person.

To the best of our knowledge, the main focus of these projects is the detection of the distress of the person but doesn't include the detection of distress sentences and calls uttered by aged voices. Indeed, in the CIRDO project, the aim is to enable detection of distress situation and call for help at home. In this perspective the project addresses the important issues of aged voice recognition and distress sentence detection.

III. Distress sentence detection

We developed an application for the recognition of distress situations by analyzing the audio signal to detect sound and voice. This work is planned to be coupled with an analysis of video scenes developed by the LIRIS laboratory in Lyon.

A. Audio processing in line: CirdoX

Fig. 1 shows in the first place the processing chain and processing of the audio signal, and then its interaction with video analysis and the data fusion system in charge of interaction with the social inclusion device e-lio. The method used in the last step for detection of emergency calls is presented in Section B.

CirdoX is made of two parallel processes communicating by IPC (Inter-Process Communication) signals. This application is designed to be modular, with independent modules so that each module can be chosen from among several plug-ins corresponding to different techniques.

1) First process

The audio stream is captured by a microphone and recorded by the Acquisition module at the input of "**Process 1**". This module can work through the plug-ins using the PortAudio library (audio card), Kinect or a National Instrument card. Then, the Detection module detects the occurrence of a signal (sound or voice). The detection of the occurrence of an audio event is based on the change of energy level of the three highest frequency coefficients of the Discrete Wavelet Transform (DTW) in a floating window frame (last 2048 samples without overlapping). Each time the energy on a channel goes beyond a self-adaptive threshold, an audio event is detected until the energy decrease below this level for at least an imposed duration [19]. Then, a wave file is created by the Wave Creation module.

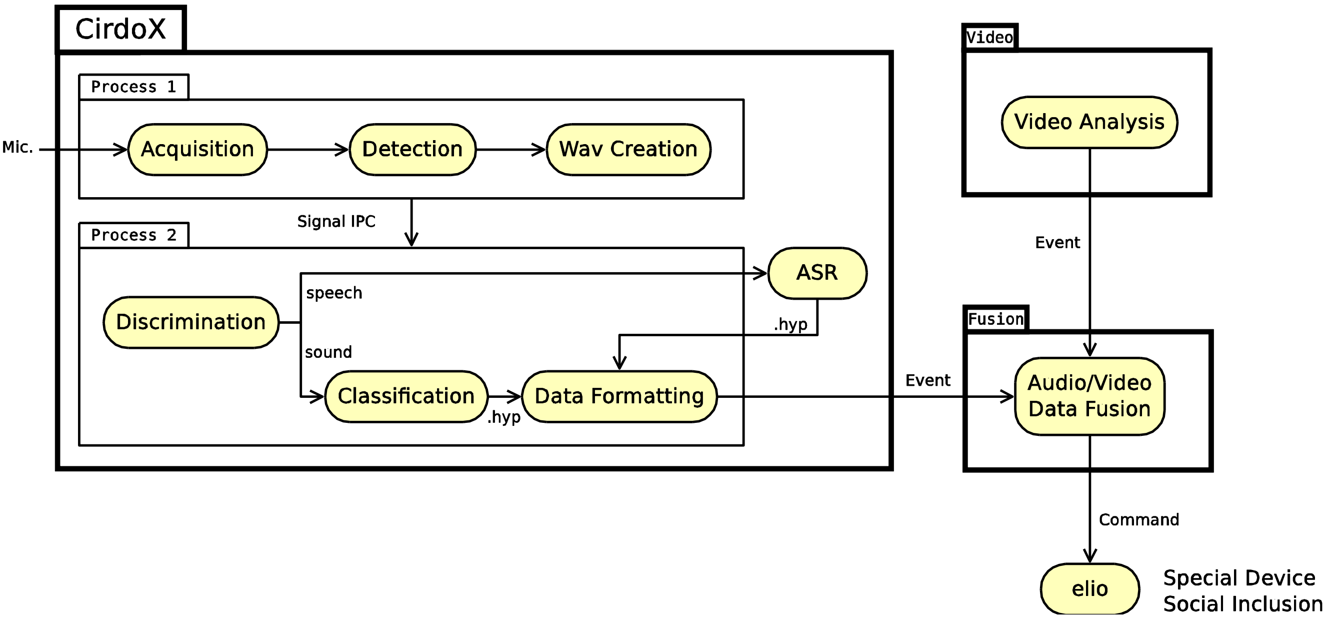


Fig. 1. Organisation of CirdoX and its insertion into the global system.

2) Second process

This output is then analyzed by "**Process 2**". The Discrimination module determines whether the signal is sound or speech. This module can be executed through the plug-ins using a Gaussian Mixture Model (GMM) or decision tree. In the case of sound, the Classification module defines to which sound class the signal belongs. It can work through GMM, decision tree and also Hidden Markov Model (HMM). In the case of speech, the signal is sent to the ASR module in order to output a sentence. Both hypotheses from Classification and ASR module are received by the Data Formatting module that sends a socket containing all the relevant data to the Fusion module of the global system.

3) Fusion system

Finally, in addition to audio information presented before, the Fusion module also receives data from the recognition of distress by video analysis in order to make a fusion between video and audio data. This one determines whether or not a distress situation is occurring and sends a socket containing its decision to the social inclusion product, e-lio (see Fig. 2). Then, e-lio executes an appropriate action, such as calling a relative, a physician or a carer.

Fig. 2. The e-lio solution by Technosens.

B. Detection of emergency calls in the home

As a preliminary approach, we used a filter to detect the distress sentences into the ASR hypotheses by using Levenshtein distance [25]. This 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 (insertion, deletion, substitution) required to change one word into the other. In our case, we use it to determine whether a sentence recognized by the ASR is close to a distress sentence by calculating the difference between the reference sequence which is the correct transcription of such sentence, and the hypothesis which is the recognized sentence.

Based on the GRePS laboratory's work [26] that interviewed elderly people in nursing homes to identify and describe what situations of distress they could have experienced, we created a list of sentences that can occur during a distress situation.

The filter calculates the Levenshtein distance between the ASR output hypothesis and all the distress sentences of the list. The sentence from the list with the best Levenshtein distance is selected according to a threshold. The lower the distance (score from 0 to infinity), the better the matching will be. To not be biased by orthography, the distance is calculated on a phonemic level. This approach takes into account some recognition errors such as word endings errors or slight variations. Moreover, in many cases, a miss-decoded word is phonetically close to the correct one due to close pronunciation.

IV. Corpora and ASR system

Very few corpora are related to ageing voices in French. These different corpora are stem of projects related to the study of French language like the "Corpus de Français Parlé

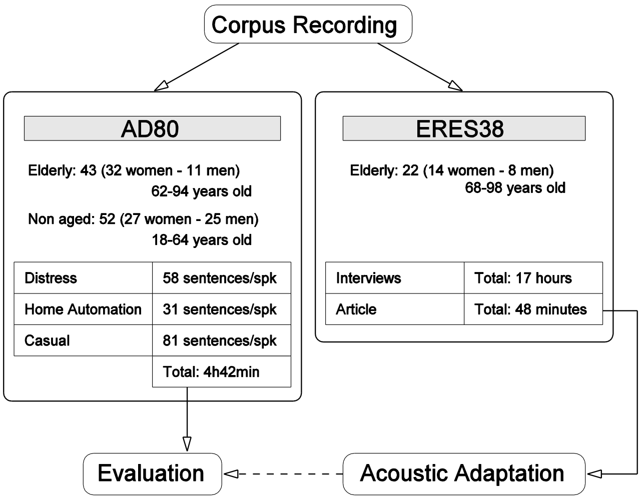
Parisien des années 2000"[3]. This corpus is made of recordings of inhabitants of different districts of Paris in order to study the influence of French spoken language over France and the French speaking world. The "Projet Phonologie du Français Contemporain"[4] is a database of records according to the region or the country. The records of 38 ageing persons (above 70 years old) are included, each record is made of a word list, a small text and two interviews.

Other available sources come from videos of testimonies of Shoah survivals and recorded in the framework of "Mémorial de la Shoah"[5] which collects testimonies and organizes conferences. These videos are not annotated. This corpus is then a collection of interviews and spontaneous speech.

Therefore, an important issue for our work was the non-existence of speech corpus containing distress calls and home automation commands. Thus, we recorded the French AD80 and ERES38 corpora (see Fig. 3). To enable comparison of system performance between elderly and non-elderly, the corpus contains recordings made by these two groups of people.

A. AD80 corpus

After a bibliographical study in collaboration with the GRePS laboratory and in the prolongation of older studies [27], we defined a list of home automation orders and of distress calls that the person could utter during a distress situation to request for assistance. This list was completed with casual sentences. At the beginning of the CIRDO project, a survey was conducted by the GRePS laboratory [26] in order to determine what sort of sentences were truly uttered by aged people in distress case and during a fall. The collected sentences were added to the list. Ten samples of each sentence category (distress, home automation order, casual) are given in TABLE I.



Distress	58 sentences/spk
Home Automation	31 sentences/spk
Casual	81 sentences/spk
	Total: 4h42min

Interviews	Total: 17 hours
Article	Total: 48 minutes

Fig. 3. The recorded corpora AD80 and ERES38.

This aged and non-aged corpus is called the AD80 corpus (Anodin-Détresse: *anodin* means colloquial and *détresse* means distress).

The non-aged part of the corpus was previously recorded in our laboratory in 2004 and was complemented in 2013 with sentences based on [26]. The collection of the aged part of the AD80 corpus was performed sporadically from 2009 to 2012 in collaboration with a rehabilitation centre, volunteers and a nursing home. Targeted speakers were persons aged of more than 60 years old, able to read and with no mental disorder or pathology altering the voice. The recording was done with a single microphone positioned at about 30 cm from the speaker's mouth. Most speakers were sat, but some were in a wheelchair or laying in a bed. The recording was done using a computer and a homemade software to prompt sentences to be read by the speaker and to record the utterances using voice activity detection. Given the targeted application (in-home voice commands and distress calls) the participants were requested to read a list of short distress/home automation and casual sentences such as *Aidez-moi* (Help me) or *Il fait beau* (It's sunny).

Indeed, the AD80 corpus was acquired from 95 speakers (36 men and 59 women) who were asked to read short sentences as mentioned before. The speakers are divided into two groups: the elderly group composed of 43 speakers (11 men and 32 women), 62 to 94 years old, with 2796 distress and home automation sentences for a duration of 1 hour 5 minutes, and 3006 casual sentences for a duration of 1 hour 6 minutes, and the non-elderly group composed of 52 speakers, 18 to 64 years old, with 3903 distress and home automation sentences for a duration of 1 hour 18 minutes, and 3897 casual sentences for a duration of 1 hour 12 minutes.

Speakers around 60-69 years old were placed in the groups according to their level of autonomy loss, their physical ageing, and whether or not they were in a nursing house. For the 43 speakers of the aged AD80 corpus, a GIR score was obtained after clinicians filled the AGGIR grid (French national test). This score is computed from the performances of the person to classify it in one of the six groups: GIR 1 (total dependence) to GIR 6 (total autonomy).

Finally, the AD80 corpus is made up of 13,602 annotated sentences, with 4 hours and 42 minutes of recording.

B. The training ERES38 corpus

Another aged corpus, the ERES38 corpus (Entretiens RESidences 38: *Entretiens* means interviews) was acquired for model adaptation purpose. Contrary to AD80, this is a collection of spontaneous speech and the annotation task was an important effort. This corpus was recorded in 2011 in the living place of the person. During the interviews, we requested each speaker to read a text but they were also asked to talk freely about their life. The text was an article about gardening created by the experimenters in order to target phoneme issues reported in [10], [28].

[3] http://ed268.univ-paris3.fr/syled/ressources/Corpus-Parole-Paris-PIII
[4] http://www.projet-pfc.net
[5] http://www.memorialdelashoah.org

TABLE I. EXAMPLES OF SENTENCES OF THE AD80 CORPUS.

Sample	Distress Sentence	Home Automation Order	Casual Sentence
1	Qu'est-ce qui m'arrive !	Appelle quelqu'un e-lio !	Bonjour madame !
2	Oh là !	e-lio appelle quelqu'un !	Ça va très bien.
3	Oh là ! Je saigne ! Je me suis blessé !	e-lio appelle les secours !	Ce livre est intéressant.
4	Aïe ! J'ai mal !	e-lio appelle ma fille !	Il fait soleil.
5	Je peux pas me relever !	e-lio appelle un docteur !	J'ai ouvert la porte.
6	Aidez-moi !	e-lio appelle le SAMU !	Je dois prendre mon médicament !
7	Au secours !	e-lio appelle les pompiers !	J'allume la lumière !
8	Je me sens mal !	e-lio appelle une ambulance !	Je me suis endormi tout de suite !
9	Je suis tombé !	e-lio tu peux téléphoner au SAMU ?	Le café est brûlant !
10	Du secours s'il vous plaît !	e-lio tu peux appeler une ambulance ?	Où sont mes lunettes ?

The ERES38 corpus was acquired from 22 elderly people (14 women and 8 men) between 68 and 98 years old. The corpus included 48 minutes of read speeches and 17 hours of interviews. The speakers lived in specialized institutes, such as nursing homes and were cognitively intact without severe disabilities.

C. ASR

The ASR toolkit chosen in our study was Sphinx3 [29]. This decoder used a context-dependent acoustic model with 3-state left-to-right HMM. The acoustic vectors are composed of 13 MFCC coefficients, the delta and the delta delta of each coefficient. This HMM-based context-dependent acoustic model was trained on the BREF120 corpus [30] which is composed of about 100 hours of annotated speech from 120 non-elderly French speakers (different from AD80). We called it the **generic acoustic model**.

A general language model (LM) was estimated from French newswire collected in the *Gigaword* corpus. It was 1-gram with 11,018 words. Moreover, to reduce the linguistic variability, a 3-gram domain language model was learnt from the sentences used during the corpus collection described in Section A, with 88 1-gram, 193 2-gram and 223 3-gram models. Finally, the language model was a 3-gram-type which results from the combination of the general language model (with a 10% weight) and the domain one (with 90% weight). This combination has been shown as leading to the best WER for domain specific application [31]. 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 correctly recognize the utterances.

V. ASR PERFORMANCES ISSUED WITH AGED VOICES

The experiments carried out in automatic speech recognition have shown performance degradation for "atypical" population such as children or elderly people [32], [9], [33] and have shown the interests of an adaptation to the target populations [33], [34]. A study of Gorham-Rowan and Laures-Gore [35] highlights the effects of ageing on the speech utterance and the consequences on the speech recognition. Speech recognition adapted to the voice of elderly people is still an under-explored area. The relevant languages are mainly English [9] and Asian languages such as Japanese [8]. A very interesting study [9] used some recordings of speeches delivered in the Supreme Court of the United States over a decade. These recordings are particularly interesting because they were used to study the evolution of recognition performance on the same person depending on his age over 7-8 years. A limitation of this study is that it relates to people with good diction and with experience in public speaking. These studies show that the performance of recognition systems decreases steadily with age, and that a special adaptation to each speaker can get closer to the scores obtained from the youngest speakers without adaptation. The implicit consequence is that the recognition system is adapted to a single speaker. To make the system adapted to the person, Renouard et al. [36] proposed to use the recognized words to adapt online the recognition models. Proposed in the context of home assistance, this research does not appear to have been pursued. It can be emphasized that no study has considered French aged voice in smart home condition. Moreover, most studies considered the chronological age as global explanatory factor while many other effects can also be responsible for ASR performance degradation as raised by [37]. There is thus no certainty that age can predict the reliability of a voice-based control system. That is why our study includes an evaluation from the dependence perspective.

A. General acoustic models are unadapted to aged voice

When performing ASR using the **generic acoustic model** on the distress/home automation sentences of the AD80 corpus, we obtained an average WER of 9% for the non-elderly group, and an average WER of 43.5% for the elderly group, reflecting a significant decrease in performance for aged speech. Fig. 4 represents the WER according to the age of each speaker.

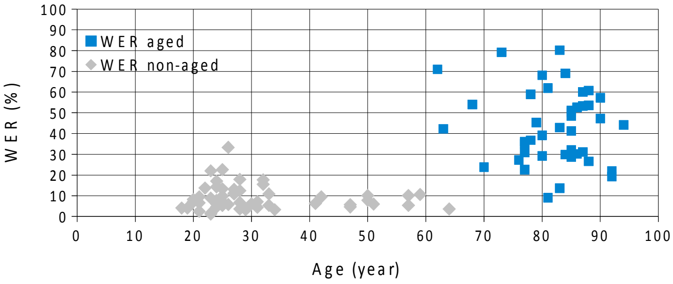


Fig. 4. WER as a function of age for aged and non-aged groups.

Although the corpus contains few speakers between 35 and 60 years old, we can observe that performance decreases steadily with age as shown on the previous studies [8], [9], [37], and that the variability of performance is very significant from the age of 60 years old. For instance, some 83 years old speakers have their WER ranging from 13.6% to 80.2%. Standard deviation is 6% for the non-elderly group and 17.3% for the elderly group. In other words, the WER is far less predictable in the elderly group than in the non-elderly group. Consequently, we have to deal with the fact that speech recognition with such a system can work very well with some of the elderly speakers, and very badly with others.

B. *Influence of dependence on ASR performances*

The Maximum Likelihood Linear Regression (MLLR) was used to adapt globally the **generic acoustic model** to the voice of elderly people [38] with the records of ERES38 and we obtained an **elderly acoustic model**. With the global MLLR adaptation using ERES38, the average WER was 14.52%. Compared to the 43.47% WER without adaptation, the absolute difference was -28.95%.

Among the elderly group and with this acoustic adaptation, we observed a great disparity between the WER for older people in the same age category, this confirms the results of Section A. Therefore, we studied other criteria and focused on the elderly dependency.

TABLE II shows the repartition of the elderly group of the AD80 corpus as a function of the GIR grid, it can be observed that none of them are in total dependence (GIR1) and 35% are in total autonomy (GIR6). Due to the small number of speakers in GIR2, GIR3 and GIR5, we merged GIR2 with GIR3 and GIR4 with GIR5 in order to obtain more balanced classes.

TABLE II. COMPOSITION OF THE AGED PART OF THE AD80 CORPUS ACCORDING TO THE GIR GRID.

AD80	GIR1	GIR2	GIR3	GIR4	GIR5	GIR6	Total
Members	0	4	2	21	1	15	43
		GIR2-3		GIR4-5		GIR6	
		6		22		15	

The ASR was operated using **elderly acoustic model** on the distress/home automation sentences of the AD80 corpus. Fig. 5 shows the variation of the WER as a function of the age and for the 3 considered GIR groups.

There is an important dispersion of the results inside each group, the WER averages for GIR2-3, GIR4-5 and GIR6 are respectively 25.2%, 13.2% and 12.2%, and the WER standard deviations are respectively 16.8%, 8.4% and 7.6%. The dispersion is the highest for the most dependent people (GIR2-3), for 50% of this group the WER is equal or upper to 20% (see TABLE III), so that only half of the group obtains average performance.

TABLE III. WER UPPER THAN 20% FOR EACH GIR GROUP.

AD80	GIR2-3	GIR4-5	GIR6
Members	50%	18%	26%

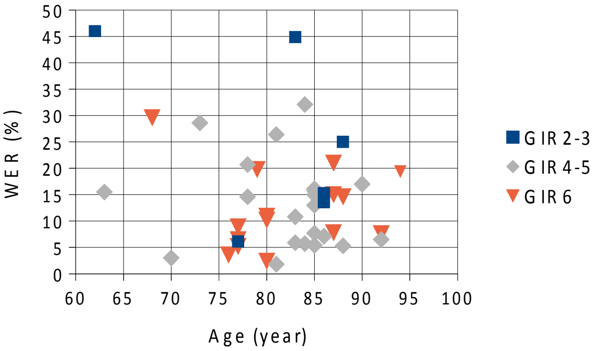


Fig. 5. WER as a function of age and dependency levels.

This suggests that the grid is too general to give a reliable indication. Therefore in addition, we conducted a preliminary study on the correlation between WER and each of the 17 parameters of the AGGIR grid. It seems that the parameters characteristic of motor control of the upper limbs and continence may be more correlated with WER. Indeed, these parameters could be representative of an advanced general physical degradation that also affects the voice control and thus diminishes the performance of ASR system

VI. DISTRESS DETECTION RESULTS

Many sentences are recognized by the ASR, but only two categories of them are relevant to the system and must be taken into account: home automation controls and calls related to a distress situation. In addition, users do not want the sentences they utter when they call their relatives are treated by the system [11]. It is then necessary to determine whether the hypothesis output by the ASR between is part of one of the two categories of interest. Therefore, we use a distance between the test sentence and typical phrases characteristics of these two categories, which allows to exclude phrases of everyday conversation.

A list was generated to categorize the recognized sentences, this list was made of home automation orders and distress calls of the AD80 corpus as described in Section IV-A. A detected sentence occurs when the Levenshtein distance (normalized by the number of phonemes) between the ASR output hypothesis and a sentence from the list is under the threshold. If a distress sentence is detected, it is considered as true positive (TP), otherwise a casual sentence which is under the threshold is considered as a false positive (FP). A casual sentence above threshold is true negative (TN), and a distress sentence above threshold is false negative (FN).

In order to assess the Levenshtein distance filter, we realized a decoding with the elderly speakers from the AD80 corpus, including 2796 distress sentences and 3006 casual sentences, for a total duration of 2 hours 12 minutes. The casual utterances were used as disrupters, with some sentences far from the distress ones: for instance *Les patates sont cuites* (Potatoes are cooked), or closer, for instance *Le médecin a appelé}* (The doctor called).

The average WER with the adapted acoustic model was 14.5% for the distress sentences, and was much higher for the casual sentences, 87.5%, due to the adapted-to-distress language model.

Then, we drew a ROC curve (Fig. 6) representing the True Positive Rate (TPR) as a function of the False Positive Rate (FPR) by varying the threshold on the Levenshtein distance, for the 43 elderly speakers.

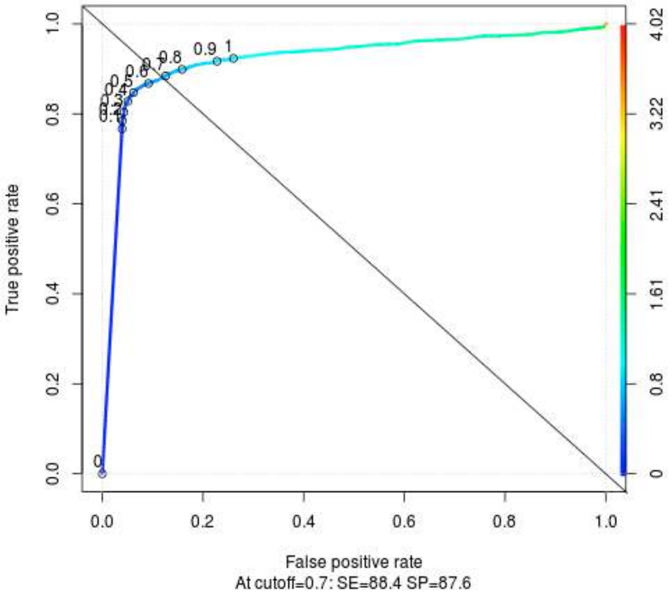


Fig. 6. ROC curve representing TPR as a function of FPR.

At the point of equal error (cutoff = 0.7), sensibility and specificity were equal to 88.4% and 87.6%. We obtained the positive and negative test showed in TABLE IV with threshold = 0.7.

TABLE IV. POSITIVE AND NEGATIVE TEST.

Threshold = 0.7	Distress	Casual
d ≤ threshold	TP = 2472	FP = 374
d > threshold	FN = 324	TN = 2632

Also, we found a recall, precision and F-measure equal to 88.4%, 86.9% and 87.2%. This is quite similar with another study conducted by Lecouteux et al. [39], where they detected home automation orders among casual utterances with a multisource ASR in a smart home environment. On their baseline system (best SNR channel), they found WER, recall, precision, and F-measure respectively equal to 18.3%, 88.0%, 90.5% and 89.2%.

In TABLE V, we evaluated the distress sentences detection for the different levels of elderly dependence GIR2-3, GIR4-5 and GIR6 by using recall, precision and F-measure.

TABLE V. DISTRESS SENTENCES DETECTION IN FUNCTION OF DEPENDENCE.

	Recall	Precision	F-measure
GIR 6	89.8%	88.2%	89.0%
GIR 4-5	89.5%	87.2%	88.3%
GIR 2-3	78.9%	80.4%	79.7%

Once more, we observed that the system performance is quite depending on the level of the elderly dependence. Due to the lower F-measure found on GIR2-3 comparing to GIR4-5 and GIR6 (-9%), we suggest that such a system could be provided to elderly and usable by them only if their autonomy score ranges from GIR4 to GIR6.

VII. CONCLUSION

In this paper, we showed that our ASR system presents higher WER for elderly voice than for young voice, with an average WER equal to 43.5% for the aged group without acoustical adaptation vs. 9% for the non-aged group, for elderly the WER is 14.5% after adaptation using the ERES38 corpus. We observed that the WER is related to the level of dependence of elderly people, with WER = 25.2% for GIR 2-3, WER = 13.2% for GIR 4-5 and WER = 12.2% for GIR 6. We presented the CirdoX software, used to filter and detect the distress sentences. We evaluated this filter based on Levenshtein distance and showed that the recall and precision reach 88.4% and 86.9%. Moreover, we showed that such a system cannot detect distress sentences well enough in the case of elderly in GIR 2-3 because of the high WER in this group.

In a future work, the whole system CirdoX including sound classification, speech recognition and fusion with video scene recognition is going to be evaluated in realistic condition in a real smart home [40]. Professional actors would play some scenarios, including for example falls to the ground because of the foot grasped in the carpet, sudden weakness, etc. We will assess how video can improve distress detection.

ACKNOWLEDGMENT

This study was funded by the National Agency for Research under the project CIRDO - Industrial Research (ANR-2010-TECS-012). The authors would like to thank to Mrs Vézignol and Bonnefond-Jimenez, Mr Debrus, Mrs Aman, Bron, Lalande and Martins of the medical institutions SSR "Les Cadières" and EHPAD "Château de Labahou" for their help in corpus recording.

Special thanks to R. Dugheanu, J. Le Grand and Y. Sasa for their active contribution, and to various elderly and caregivers who agreed to participate in the recordings.

REFERENCES

[1] V. Bellamy and C. Beaumel, "Bilan démographique 2012. la population croît, mais plus modérément", INSEE Première, vol. 1429, pp. 1-4, 2013.

[2] M. Duée and C. Rebillard, "La dépendance des personnes âgées : une projection en 2040", Données sociales - La société française, pp. 613-619, 2006.

[3] CSA, "Les français et la dépendance", http://www.csa.eu/fr/s26/nos-sondages-publies.aspx, 2003, accessed: 12/03/2013.

[4] C. Tlili, "Perspectives démographique et financières de la dépendance", Rapport du groupe de travail sur la prise en charge de la dépendance, 2011.

[5] W. Ryan and K. Burk, "Perceptual and acoustic correlates in the speech of males", Journal of Communication Disorders, vol. 7, pp. 181-192, 1974.

[6] N. Takeda, G. Thomas, and C. Ludlow, "Aging effects on motor units in the human thyroarytenoid muscle", Laryngoscope, vol. 110, pp. 1018-1025, 2000.

[7] P. Mueller, R. Sweeney, and L. Baribeau, "Acoustic and morphologic study of the senescent voice", Ear, Nose, and Throat Journal, vol. 63, pp. 71-75, 1984.

[8] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Acoustic models of the elderly for large-vocabulary continuous speech recognition", Electronics and Communications in Japan, Part 2, vol. 87, pp. 49-57, 2004.

[9] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices", in 9th International Conference on Speech Science and Speech Technology (InterSpeech 2008), Brisbane, Australia, 2008, pp. 2550-2553.

[10] R. Privat, N. Vigouroux, and P. Truillet, "Etude de l'effet du vieillissement sur les productions langagières et sur les performances en reconnaissance automatique de la parole", Revue Parole, vol. 31-32, pp. 281-318, 2004.

[11] 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, pp. 127-144, 2013.

[12] S. S. Intille, "Designing a home of the future", IEEE Pervasive Computing, vol. 1, no. 2, pp. 76-82, 2002.

[13] C. Chen and D. Cook, "Behavior-based home energy prediction", in International Conference on Intelligent Environments, 2012, pp. 1-7.

[14] A. Holmes, H. Duman, and A. Pounds-Cornish, "The iDorm: Gateway to heterogeneous networking environments", in Proceedings of international ITEA workshop on virtual home environments. Paderborn, Germany: ITEA Press, 2002, pp. 20-37.

[15] M. Skubic, G. Alexander, M. Popescu, M. Rantz, and J. Keller, "A smart home application to eldercare: Current status and lessons learned", Technology and Health Care, vol. 17, no. 3, pp. 183-201, 2009.

[16] N. Zouba, F. Bremond, M. Thonnat, A. Anfosso, E. Pascual, P. Mallea, V. Mailland, and O. Guerin, "A computer system to monitor older adults at home: Preliminary results", Gerontechnology Journal, vol. 8, no. 3, pp. 129-139, July 2009.

[17] P. Wolf, A. Schmidt, and M. Klein, "Soprano - an extensible, open aal platform for elderly people based on semantical contracts", in 3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'08), 18th European Conference on Artificial Intelligence (ECAI 08), Patras, Greece, 2008.

[18] M. Vacher, B. Lecouteux, and F. Portet, "Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment", in EUSIPCO (European Signal Processing Conference), Bucarest, Romania, August 27-31 2012, pp. 1663-1667.

[19] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat, "Information extraction from sound for medical telemonitoring", Information Technology in Biomedicine, IEEE Transactions, vol. 10, pp. 264-274, April 2006.

[20] M. Popescu, Y. Li, M. Skubic, and M. Rantz, "An acoustic fall detector system that uses sound height information to reduce the false alarm rate", in Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008, 20-25 Aug. 2008, pp. 4628-4631.

[21] C. Doukas and I. Maglogiannis, "Enabling human status awareness in assistive environments based on advanced sound and motion data classification", in Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments. New York, NY, USA: ACM, 2008, pp. 1:1-1:8.

[22] M. Hamill, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems", Journal of NeuroEngineering and Rehabilitation, vol. 6, 2009.

[23] G. L. Filho and T. J. Moir, "From science fiction to science fact: a smarthouse interface using speech technology and a photo-realistic avatar", Int. J. Comput. Appl. Technol., vol. 39, no. 1/2/3, pp. 32-39, Aug. 2010.

[24] P. Milhorat, D. Istrate, J. Boudy, and G. Chollet, "Hands-free speechsound interactions at home", in EUSIPCO (European Signal Processing Conference), Bucarest, Romania, August 27-31 2012, pp. 1678-1682.

[25] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals", Soviet Physics-Doklady, vol. 10, pp. 707-710, 1966.

[26] M.-E. Bobillier-Chaumon, B. Cuvillier, S. Bouakaz, and M. Vacher, "Démarche de développement de technologies ambiantes pour le maintien à domicile des personnes dépendantes : vers une triangulation des méthodes et des approches", in Actes du 1er Congrès Européen de Stimulation Cognitive, Dijon, France, 23-25 May 2012, pp. 121-122.

[27] M. Vacher, J. Serignat, S. Chaillol, D. Istrate, and V. Popescu, "Speech and sound use in a remote monitoring system for health care", vol. 4188/2006, pp. 711-718, 2006.

[28] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Contribution à l'étude de la variabilité de la voix des personnes âgées en reconnaissance automatique de la parole (assessment of the acoustic models performance in the ageing voice case for ASR system adaptation) [in french]", in Actes de la conférence JEP-TALN-RECITAL 2012, vol. 1: JEP, Grenoble, France, June 2012, pp. 707-714.

[29] K. Seymore, C. Stanley, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer, "The 1997 CMU Sphinx-3 English broadcast news transcription system", DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[30] L. Lamel, J. Gauvain, and M. EskEnazi, "BREF, a large vocabulary spoken corpus for french", in Proceedings of EUROSPEECH 91, vol. 2, Geneva, Switzerland, 1991, pp. 505-508.

[31] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions", in Interspeech 2011, Florence, Italy, 2011, p. 4.

[32] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly", in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 1996, pp. 349-352.

[33] M. Gerosa, Giuliani, and F. D., Brugnara, "Towards age-independent acoustic modeling", Speech Communication, vol. 51(6), pp. 499-509, 2009.

[34] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson, "Recognition of elderly speech and voice-driven document retrieval", in IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '99, vol. 1, 1999, pp. 145-148.

[35] M. Gorham-Rowan and J. Laures-Gore, "Acoustic-perceptual correlates of voice quality in elderly men and women", Journal of Communication Disorders, vol. 39, pp. 171-184, 2006.

[36] S. Renouard, M. Charbit, and G. Chollet, Independent Living for Persons with Disabilities, 2003, ch. Vocal interface with a speech memory for dependent people, pp. 15-21.

[37] T. Pellegrini, I. Trancoso, A. Hämäläinen, A. Calado, M. S. Dias, and D. Braga, "Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese", in Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings, 2012, pp. 139-147.

[38] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Analysing the performance of automatic speech recognition for ageing voice: Does it correlate with dependency level?" in Speech and Language Processing for Assistive Technologies, Satellite workshop of Interspeech2013, Grenoble, France, Aug. 21-22 2013, pp. 1-7.

[39] B. Lecouteux, M. Vacher, and F. Portet, "Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions", in 12th International Conference on Speech Science and Speech Technology (InterSpeech 2011), Florence, Italy, Aug. 28-31 2011, pp. 2273-2276.

[40] M. Gallissot, J. Caelen, F. Jambon, and B. Meillon, "Une plate-forme usage pour l'intégration de l'informatique ambiante dans l'habitat. L'appartement Domus", Technique et Science Informatiques (TSI), vol. 32(5), pp. 547-574, 2013.