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

To cite this version:
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. 6th Workshop on Speech and Language Processing for Assistive Technologies, SIG-SLPAT, Sep 2015, Dresden, Germany. pp.1-7. hal-01179930

HAL Id: hal-01179930
https://hal.archives-ouvertes.fr/hal-01179930
Submitted on 23 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Recognition of Distress Calls in Distant Speech Setting: a Preliminary Experiment in a Smart Home

Michel Vacher\textsuperscript{1}, Benjamin Lecouteux\textsuperscript{2}, Frédéric Aman\textsuperscript{1}, Solange Rossato\textsuperscript{2}, François Portet\textsuperscript{2}

\textsuperscript{1}CNRS, LIG, F-38000 Grenoble, France
\textsuperscript{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 \cite{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 \cite{3,4,11,5}. 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 \cite{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), which 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 \cite{7}. It was demonstrated that VUI is useful for system integrating speech commands \cite{8}.

The use of speech technologies in home environment requires to address particular challenges due to this specific environment \cite{9}. There is a rising number of smart home projects considering speech processing in their design. They are related to wheelchair command \cite{10}, vocal command for people with dysarthria \cite{11,8}, companion robot \cite{12}, vocal control of appliances and devices \cite{13}. Due to the experimental constraints, few systems were validated with real users in realistic situation condition like in the SWEET-HOME project \cite{14} during which a dedicated voice based home automation system was able to drive a smart home thanks to vocal commands with typical people \cite{15} and with elderly and visually impaired people \cite{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\textsuperscript{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 \cite{17} in which a representative set of seniors were included.

\textsuperscript{1}http://www.technosens.fr/
is set on a table in the living room in front 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:

\[ p(x|j) = \sum_{j=m}^{M} c_{jmi} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(x; \mu_{jmi}, \Sigma_{i}) \]

\[ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jmi} \]

\[ w_{jmi} = \frac{\exp \{ w_{jmi} \} \gamma_{jmi}}{\exp \{ \gamma_{jmi} \} + \sum_{v=1}^{V} \exp \{ w_{jvi} \} \gamma_{jvi}} \]

where \( x \) denotes the feature vector, \( j \in \{1...J\} \) is the HMM state, \( i \) is the Gaussian index, \( \mathbf{v}_{jmi} \) is the substate weight. Each state \( j \) is associated to a vector \( \mathbf{v}_{jmi} \in \mathbb{R}^S \) (\( S \) is the phonetic subspace dimension) which derives the means, \( \mu_{jmi} \) and mixture weights, \( w_{jmi} \) and it has a shared number of Gaussians, \( I \). The phonetic subspace \( \mathbf{M}_i \), weight projections \( w_i \) and covariance matrices \( \Sigma_i \) i.e. the globally shared parameters \( \Phi_i = \{ \mathbf{M}_i, w_i, \Sigma_i \} \) 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 aim 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 GMM-UBM).

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.

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

(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:

\[ \text{CER} = \frac{\text{Number of missed calls}}{\text{Number of calls}} \]  

(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 detect a false alarm rate in this framework.
develop the language model used by our system. This corpus was
The sentences of the AD80 corpus [19] served as basis to de-
3.1.2. Voice commands and distress calls
rehearsals before starting the recordings so that the participant
very tiring for the participants and it was necessary to include
an experiment was 2h 30min per person. This experiment was
ations. Figure 3 shows a young participant wearing the simulator
reduced his vision and hearing to simulate aged physical condi-
falls by the video analysis system. If the participant's age was
called "true-false" added to challenge the automatic detection of
narios of fall, one blocked hip scenario and two other scenarios
participant was introduced to the context of the research and was
me? Oh shit, shit!
Blast! What's happening to
after the fall were also identified "...
Son's reaction once on the floor. Phrases uttered during and
fall context, the movements during the fall as well as the per-
lay 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 anno-
tion process. In the same way, speeches mixed with sound produced by the fall were ignored. At the end, each speaker ut-
tered 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 differ-
ent ("oh mein on est bien là tiens - oh I am in a sticky situation"),
In practice, participants cut some sentences (i.e., inserted a de-
y 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 esti-
mated 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-

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 Cir-
DOX 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 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 wore 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 mein 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).
3.1.3. 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 recruit-
ment 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.
Table 1: Composition of the audio corpus Cirdo-set

<table>
<thead>
<tr>
<th>Spk.</th>
<th>Age</th>
<th>Sex</th>
<th>Nb. of interjections or short sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>30</td>
<td>M</td>
<td>22</td>
</tr>
<tr>
<td>S02</td>
<td>30</td>
<td>M</td>
<td>14</td>
</tr>
<tr>
<td>S03</td>
<td>34</td>
<td>F</td>
<td>16</td>
</tr>
<tr>
<td>S04</td>
<td>34</td>
<td>F</td>
<td>15</td>
</tr>
<tr>
<td>S05</td>
<td>29</td>
<td>M</td>
<td>24</td>
</tr>
<tr>
<td>S06</td>
<td>64</td>
<td>F</td>
<td>23</td>
</tr>
<tr>
<td>S07</td>
<td>61</td>
<td>M</td>
<td>23</td>
</tr>
<tr>
<td>S08</td>
<td>44</td>
<td>M</td>
<td>25</td>
</tr>
<tr>
<td>S09</td>
<td>16</td>
<td>M</td>
<td>32</td>
</tr>
<tr>
<td>S10</td>
<td>16</td>
<td>M</td>
<td>19</td>
</tr>
<tr>
<td>S11</td>
<td>52</td>
<td>M</td>
<td>12</td>
</tr>
<tr>
<td>S12</td>
<td>28</td>
<td>M</td>
<td>15</td>
</tr>
<tr>
<td>S13</td>
<td>66</td>
<td>M</td>
<td>24</td>
</tr>
<tr>
<td>S14</td>
<td>52</td>
<td>F</td>
<td>23</td>
</tr>
<tr>
<td>S15</td>
<td>23</td>
<td>M</td>
<td>20</td>
</tr>
<tr>
<td>S16</td>
<td>40</td>
<td>F</td>
<td>29</td>
</tr>
<tr>
<td>S17</td>
<td>40</td>
<td>F</td>
<td>24</td>
</tr>
<tr>
<td>S18</td>
<td>25</td>
<td>F</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>40.76</td>
<td>413</td>
<td>341</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Distress Sentence</th>
<th>Home Automation Command</th>
<th>Casual Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oh là ⋆</td>
<td>e-lio tu peux appeler une ambulance</td>
<td>Je me suis endormi tout de suite</td>
</tr>
<tr>
<td>Aie ⋆</td>
<td>e-lio appelle les secours</td>
<td>Je dois prendre mon médicament</td>
</tr>
<tr>
<td>Distress Sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 CIDRO project 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


