Influence of expressive speech on ASR performances: application to elderly assistance in smart home

Frédéric Aman, Véronique Aubergé, and Michel Vacher

[1] Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2] CNRS, LIG, F-38000 Grenoble, France
http://www.liglab.fr/
{Frederic.Aman@imag.fr,Veronique.Auberge@imag.fr,Michel.Vacher@imag.fr}

Abstract. Smart homes are discussed as a win-win solution for maintaining the Elderly at home as a better alternative to care homes for dependent elderly people. Such Smart homes are characterized by rich domestic commands devoted to elderly safety and comfort. The vocal command has been identified as an efficient, well accepted, interaction way, it can be directly addressed to the "habitat", or through a robotic interface. In daily use, the challenges of vocal commands recognition are the noisy environment but moreover the reformulation and the expressive change of the strictly authorized commands. This paper focuses (1) to show, on the base of elicited corpus, that expressive speech, in particular distress speech, strongly affects generic state of the art ASR systems (20 to 30%) (2) how interesting improvement thanks to ASR adaptation can regulate (15%) this degradation. We conclude on the necessary adaptation of ASR system to expressive speech when they are designed for person's assistance.

Keywords: Expressive speech, distress call, Ambient Assisted Living, Home Automation

1 Introduction

Smart homes aim at anticipating and responding to the special needs of elderly people living alone at their own home by assisting them in their daily life. A large variety of sensors are used in this framework [1] but one of the best interfaces seems to be the speech interface, that makes possible interaction using natural language and that is well adapted to people with reduced mobility. The purpose of this study is to develop a system able to detect distress calls uttered by elderly people. In a previous study, we reported that elderly voices resulted in degraded performances for generic Automatic Speech Recognition (ASR) systems [2]. The content of test sentences was related to distress but these sentences were read and uttered in a neutral manner without any particular emotion; in this article we are using "neutral" when utterances are not highly perturbated by intentions, attitudes or emotions.

In a real situation and in the framework of home automation, vocal commands can be expressive for 2 reasons:

1. if the user integrates a representation of the home as a communicative entity [3], thus prosody and morphosyntax can implement intentions, attitudes and social affects;
2. the command is motivated by an emotional context (panic, happiness to speak with a relative, etc.).

In a distress situation, a cry for help (i.e., "I fell", "Help me") may be charged in emotion, prosodic modifications influencing voice quality [4][5] so that emotions could be perceived by the interlocutor. Many studies are related to emotion recognition from voice (distress, fear, stress, etc.) [6][7][8][9], in [10], the authors used energy, pitch, jitter and shimmer. Scherer et al. [11] compare effects of different emotions on prosodic parameters with regard to neutral voice.

An immediate and more reachable issue consists in making ASR more robust in presence of emotion instead of recognizing the distress state but few studies were done in this domain excepted the study of Vlasenko et al. [12]. They observed a performance degradation with expressive speech when ASR were trained with neutral voice.

In former studies [13], the different ASR performances between acted expressive voice and spontaneous expressive voice from a corpus recorded using a Wizard of Oz technique was brought to light; in this study, results were compared to those of neutral voice. The aim of our study is to highlight the lack of robustness of ASR systems with expressive voice. For in-depth study, we recorded the *Voix Détresse* corpus which is made of distress sentences uttered in expressive and neutral manner in an elicitation protocol. A reduction of performance decrease was observed thanks to the use of an adapted acoustic model. Expressive voice characterisation is presented at the end.

2 Method

Ideally, this work should have been held on natural dataset. But in the spontaneous corpus of distress, situations (calls to emergency centers, etc.) are: (1) not addressed to a system (human isolation), but to a person devoted to assistance (voice human contact); (2) after a phone call, that is not in the so serious situation that the elderly is not anymore able to give a call. It is why, in this primary study, we decided to be closer to the natural situations but to control the relevant cues that will be the real situations for which there is yet no available data. Thus we proposed a devoted elicitation protocol. The Voix Détresse French corpus was recorded in our laboratory in the following elicitation protocol:

- a list of 20 prototypical utterances was selected from the *AD80* corpus [2], that was validated to be representative of distress commands spontaneously produced by subjects in real distress situations at home;
- to elicitate expressive speech in the specific context of these distress situations, for each utterance is associated a devoted photography showing an elderly in this given distress home situation; the subject had to produce the utterance (as many times as the subject decided) trying to stand in that individual's shoes;
- before to enter the elicitation procedure, in order to get a control sub-corpus, each speaker had to read the 20 distress sentences in a neutral manner, that is without any elicitation support.

Composition	Younger group	Elderly group
All speakers	12F, 8M, 23-60 years old	5F, 67-85 years old
Neutral voice	400 sentences, 5min 41s (total)	100 sentences, 1min 38s (total)
Expressive voice	782 sentences, 12min 44s (total)	199 sentences, 4min 15s (total)

Table 1. *Voix Détresse* corpus.

Young voice	*Training*	*Test*	Elderly voice	*Training*	*Test*
Neutral	200	200	Neutral	50	50
Expressive	393	389	Expressive	99	100

Table 2. Number of sentences of the *Training* and *Testing* partitions of the corpus.

The expected emotions for these distress elicitations were mainly negative emotions like fear, anger, sadness. The elderly speakers had some difficulties to enter the elicitation protocol, so that only 5 speakers, all females, had sufficient performances of "quite-natural" distress expressions to be selected for the corpus. It has to be noted that the females are the main users of the Smart homes, because the elderly population over 80 is mainly females, and moreover some French investigations show that the females are more interested by such Smart homes solutions. To complete these low number of elderly subjects, together with collecting a non elderly control sub-set, 20 other subjects, from 23 to 60 years old, were recorded in the same elicitation protocol. *Voix Détresse* is described in Table 1. The corpus is made of 1481 sentences, its length is 24min 19s.

One half of the corpus was reserved for model training and the remaining part for testing. Since 25 speakers were recorded and regarding each speaker, approximately 10 neutral sentences were reserved for adapting and 10 for testing, and about 20 expressive sentences for training and 20 for testing (see Table 2).

For our study, ASR was operated using *Sphinx3* [14] based on a context dependant triphone acoustic model. Our HMM model was trained with *BREF120* corpus [15] which is made of 100 hours of texts read by 120 speakers (65 females, 55 males), this generic model was called *BREF120*.

For the decoding, a trigram language model (LM) with a 10K lexicon was used. It results from the interpolation of a generic LM (with a 10% weight) and a specialized LM (with a 90% weight). The generic LM was estimated on about 1000M of words from Gigaword, a collection of French newspapers [3], this model is unigram and made of 11018 words. The specialized LM was estimated from the distress and call for help sentences of the *AD80* corpus, it is made of 88 unigrams, 193 bigrams et 223 trigrams. This combination has been shown as leading to the best WER for domain specific application [16]. 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.

[3] http://catalog.ldc.upenn.edu/LDC2006T17

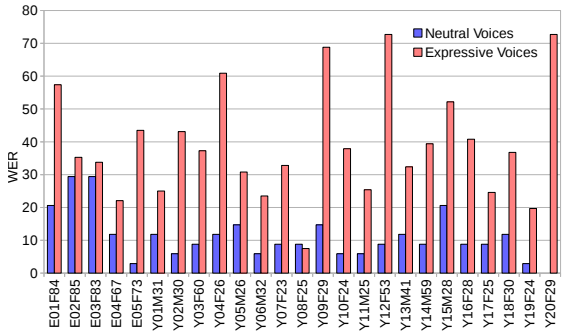


Fig. 1. *WER from Sphinx3 decoding (BREF120 acoustic model) for neutral and expressive voices*

Voice	Younger group	Elderly group
Neutral	9.27	18.82
Expressive	39.22	38.42

Table 3. WER (%) from Sphinx3 decoding (BREF120 acoustic model) for the two groups.

3 Experiment

3.1 Decoding with BREF120 acoustic model

Decoding was operated on the *Test* part of the *Voix Détresse* corpus. WER results are presented in Figure 1 for each speaker and as the average for each group in Table 3. A severe performance degradation can be observed: absolute WER difference is 29.95% for younger speakers and 19.6% for elderly speakers.

A Welch's t test was operated in order to determine if WER differences between neutral and expressive voices are significant. Test results indicate a significant difference between the corresponding WER (t=-7.2026. df=21.365. p<0.05 (p=3.834e-07) for younger speakers and t=-2.5165. df=7.874. p<0.05 (p=0.03645) for elderly speakers).

As shown in Figure 1, we can observe a WER increase for expressive voice with regard to neutral voice for all speakers, excepted for *Y08F25*, and these variations are more or less important according to the speaker. One explanation of these differences could be due to the difficulty of the participants, who are not professionnal, to "play the game" of distress during the records.

Overall, persons more comfortable to play distress situations are persons with a higher WER for distress sentences. In real situations, it is feared that very expressive sentences will be badly recognized by ASR, while it is essential to act in this situation.

3.2 Decoding with an adapted acoustic model

In order to establish the possibility of improvement, a MLLR adaptation of *BREF120* model to each speaker was operated: (1) *BREF120_N* by using the 10 neutral *Training*

Group	BREF120	BREF120_N	BREF120_E	BREF120_N+E
Young neutral	9.27	7.80	11.03	7.21
Young expressive	39.22	28.86	22.45	20.23
Elderly neutral	18.82	17.06	16.48	15.88
Elderly expressive	38.42	35.48	31.64	30.50
Average Neutral	11.18	9.65	12.12	8.94
Average Expressive	39.06	30.18	24.28	22.28

Table 4. WER (%) for the different voices using generic and adapted models.

	BREF120	BREF120_N	BREF120_E	BREF120_N+E
BREF120	-	p=0.8298	p=0.9529	p=0.5977
BREF120_N	-	-	p=0.5174	p=0.9786
BREF120_E	-	-	-	p=0.2923
BREF120_N+E	-	-	-	-
WER (%)	11.18	9.65	12.12	8.94

Table 5. *Neutral voice* group: p-value of Tukey HSD test and WER.

sentences uttered by the speaker, (2) *BREF120_E* by using the 20 expressive sentences, and (3) *BREF120_N+E* by using neutral and expressive sentences. For each speaker, 3 different models adapted to his voice were obtained. Decoding was operated as in section 3.1 on the 739 neutral and expressive sentences ot the *Test* part of the corpus using the 3 adapted acoustic models. Results are given in Table 4 with regard to the standard *BREF120* model.

ANOVA was operated on the 2 groups *neutral* and *expressive voices* (samples are Gaussian and homogeneity of group variances is verified). There is no significant difference inside the samples of the *neutral voice* group ($F(3;96)=1.293$; $p=0.281$) but a significant difference (p-value < 0.05) inside the samples of the *expressive voice* group ($F(3;96)=7.828$; $p=9.96e-05$). Results of the Tukey HSD test are given in Table 5 for *neutral voices* and in Table 6 for *expressive voices*. For neutral voices, no significant difference is observed between the WER obtained from the different acoustic models, $p>0.05$ for each acoustic model pair. By contrast for expressive voices, a significant WER difference exists between the generic *BREF120* model and the models adapted to expressive voice *BREF120_E*, with $p<0.05$ ($p=0.0011$), and an absolute difference between average WER equals to 14.78%. Conclusion is the same for *BREF120* and *BREF120_N+E*, with $p<0.05$ ($p=0.0002$), and an absolute difference of 16.78%. By contrast, there is no significant difference between the other models.

Therefore, with regard to expressive sentences, speaker adaptation from neutral sentences is not sufficient to improve significantly the WER. Models adapted to expressive voice may be used, but adaptation can be done from mixing neutral and expressive sentences because there is no significant difference between models *BREF120_E* and *BREF120_N+E*. In the same way, neutral sentences can be decoded using adapted model *BREF120_N+E* without important performance decrease.

	BREF120	BREF120_N	BREF120_E	BREF120_N+E
BREF120	-	p=0.0976	**p=0.0011**	**p=0.0002**
BREF120_N	-	-	p=0.4120	p=0.1683
BREF120_E	-	-	-	p=0.9526
BREF120_N+E	-	-	-	-
WER (%)	39.06	30.18	24.28	22.28

Table 6. *Expressive voice* group: p-value of Tukey HSD test and WER.

Parameters	Neutral voice	Expressive voice	Welch's t test (significant difference if p<0.05)
Flow (Phonemes/s)	13.58	11.71	p<0.05 (p=0.0019)
F0 (Hz)	162.74	260.43	p<0.05 (p=1.16e-06)
Jitter (%)	3.07	2.30	p<0.05 (p=0.00015)
Shimmer (%)	13.63	9.07	p<0.05 (p=1.41e-07)
HNR (dB)	12.44	14.76	p<0.05 (p=0.00077)

Table 7. Average prosodic parameters as a function of kind of voice, and Welch's t test p-values.

4 Expressive voice characterisation

Average values of prosodic parameters, flow, F0, *jitter*, *shimmer* et Harmonic to Noise Ratio (HNR), were measured and compared between neutral and elicited expressive voices. Data were the 739 neutral and expressive sentences of the *Testing* part of *Voix Détresse*.

We observed for expressive voice, with regard to neutral voice and in average, a decrease of the flow, an increase of F0 and HNR, and an increase of *jitter* and *shimmer*. Averages concerning all speakers for each parameter are given in Table 7 with significance tests of the difference between neutral and expressive voices.

In a first step, we considered for our analysis that the distress corpus (i.e., the expressive sentences) of *Voix Détresse* was homogeneous with regard to expressiveness of acted emotion (eventually with variable expressive intensity), and varied possibly only with the speaker (difference of interpretation or idiosyncratic physiologic variation). However, by more precise hearing of the considered utterances, sentence by sentence, we can suppose that the nature of the distress varies according the situation suggested by the picture submitted for elicitation, and that this nature could be homogeneous by situation (i.e., by sentence) without dominant variability of the subject. For example, the sentence "Je ne me sens pas bien" (I don't fell very well) was associated to a picture showing a very demoralized man taking his arm, supported by an other person, this person was very close and had already started to help the man in distress situation. By contrast, the sentence "A moi" (Help me) is associated to a picture showing an isolated drowning person: it is a question of a request for emergency assistance by an isolated person in a critical condition. The two situations are two opposite situations of our corpus: in the first case a helper took care of the person but in the second case a person in a critical situation was calling but there was no identified helper.

That is what we wanted to verify by evaluating prosodic parameters not by speaker but by context, The 489 expressive sentences were spread into 10 contexts. *Jitter*, *shim-*

Distress sentences	Jitter (%)	Shimmer (%)	F0 (Hz)	Intensity (dB)
A moi !	1.028	5.962	317	70.27
Oh la la !	1.367	6.383	224	65.72
J'ai un malaise !	1.411	6.778	232	63.53
Aidez-moi !	1.423	6.801	279	66.46
Du secours s'il vous plaît !	1.550	6.791	276	64.80
Me laissez pas tout seul !	1.550	7.131	276	64.27
Qu'est-ce qu'il m'arrive !	1.654	7.470	239	62.21
Au secours !	1.679	5.414	294	69.79
Je ne peux plus bouger !	1.708	8.181	282	64.72
Je ne me sens pas bien !	1.718	8.372	220	58.08

Table 8. Prosodic parameters for some different French distress sentences.

mer, F0 and flow were measured, results are given in Table 8. It appears that 2 extreme situations "A moi" (Help me) and "Je ne me sens pas bien" (I don't feel very well) are characterized, throughout all speakers, by extreme values of F0, *jitter* and *shimmer*; this fact could confirm our hypothesis on the variability of the nature of the distress. According to the 2D model of Russel [17] of cognitive psychology, emotions can be represented along 2 axis: valence (positive/negative) and *arousal* (active/passive).

Thus, the context "A moi" could correspond to a very "active" distress, by hearing they are perceived as very stressed. The context "Je ne me sens pas bien" could be more "passive" and is perceived as the less stressed. Measures corroborate these observations: we can observe in Table 8 that the context "A moi" has the lowest *jitter* and the highest F0 with regard to the 9 other contexts, and the context "Je ne me sens pas bien" has the highest *jitter* and the lowest F0.

We can see that some analog expressions are characterized by the same parameter values, as "Du secours s'il vous plaît" and "Ne me laissez pas tout seul" with *jitter*=1.550% and F0=276Hz for these 2 contexts.

5 Conclusion

After adaptation, WERs of expressive voices remain higher than WERs of neutral voices with the generic model *BREF120*, the difference is 13.1%. Moreover, the record of spontaneous expressive voices is very difficult and the *Voix Détresse* corpus is made of acted expressive voice. In all likelihood, performances may be lower in a real case.

Acoustic characteristics of distress are very subtle. It is clear that in the harsh acoustic conditions of a smart home, we have to be extremely vigilant on the physiologic, functional and communicative context of utterances to recognize: a high amount of data must be recorded in order to obtain a robust system. Moreover these data must be finely characterized as regard to their expressiveness. Those conditions are necessary but insufficient ones for the use of speech technologies in real conditions for the assistance of elderly persons.

Acknowledgements This study was supported by the French funding agencies ANR and CNSA through the project CIRDO - Industrial Research (ANR-2010-TECS-012). The authors would like to thank the persons who agreed to participate in recordings.

References

1. Peetoom, K., Lexis, M., Joore, M., Dirksen, C., De Witte, L.: Literature review on monitoring technologies and their outcomes in independently living elderly people. Disability and Rehabilitation: Assistive Technology (2014) 1–24
2. Aman, F., Vacher, M., Rossato, S., Portet, F.: Analysing the Performance of Automatic Speech Recognition for Ageing Voice: Does it Correlate with Dependency Level? In: 4th Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France (August 2013) 9–15
3. Clarcke, A.C.: 2001 : A Space Odyssey. New American Library (1968)
4. Audibert, N.: Prosodie de la parole expressive: dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés. PhD thesis, INPG, Ecole Doctorale "Ingénierie pour la Santé, la Cognition et l'Environnement" (2008)
5. Vlasenko, B., Prylipko, D., Philippou-Hübner, D., Wendemuth, A.: Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In: Proceedings of Interspeech 2011. (2011) 1577–1580
6. Vidrascu, L.: Analyse et détection des émotions verbales dans les interactions orales. PhD thesis, Université Paris Sud-Paris XI, Discipline : Informatique (2007)
7. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication **53**(9-10) (November 2011) 1062–1087
8. Vaudable, C.: Analyse et reconnaissance des émotions lors de conversations de centres d'appels. PhD thesis, Université Paris Sud-Paris XI (2012)
9. Chastagnol, C.: Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes. PhD thesis, Université Paris Sud-Paris XI (2013)
10. Soury, M., Devillers, L.: Stress detection from audio on multiple window analysis size in a public speaking task. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. (Sept 2013) 529–533
11. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication **40**(1–2) (2003) 227 – 256
12. Vlasenko, B., Prylipko, D., Wendemuth, A.: Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. In: Proc. of the KI 2012. (2012) 103–107
13. Aman, F., Auberge, V., Vacher, M.: How affects can perturbe the automatic speech recognition of domotic interactions. In: Workshop on Affective Social Speech Signals, Grenoble, France (August 2013) 1–5
14. Seymore, K., Stanley, C., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R., Thayer, E.: The 1997 CMU Sphinx-3 English broadcast news transcription system. DARPA Broadcast News Transcription and Understanding Workshop (1998)
15. Lamel, L., Gauvain, J., Eskenazi, M.: BREF, a large vocabulary spoken corpus for french. In: Proceedings of EUROSPEECH 91. Volume 2., Geneva, Switzerland (1991) 505–508
16. Lecouteux, B., Vacher, M., Portet, F.: 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) 2273–2276
17. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6) (1980) 1161–1178