How affects can perturbe the automatic speech recognition of domotic interactions

Frédéric Aman, Véronique Auberge, Michel Vacher

How affects can perturbe the automatic speech recognition of domotic interactions

Frédéric Aman, Véronique Aubergé and Michel Vacher

LIG, UMR5217 UJF/CNRS/Grenoble-INP/UMPF, 38041 Grenoble, France

{frederic.aman, veronique.auberge, michel.vacher}@imag.fr

Abstract

In Smart Home, the vocal home automation orders, for comfort purposes, or assistive devoted, have been pointed as the more relevant interaction for ambient assisted living. Even if the orders are very strictly formulated, when they are daily used (directed to the smart home, or to a robot mediator), they become often pronounced with various affects. In this paper we have evaluated how some state of the art ASR systems shut down with expressive orders, acted or spontaneous, and how the ASR training with neutral and/or acted and/or spontaneous expressive commands corpus can greatly modify the ASR performances.

Index Terms: ASR, Smart Home, Vocal Commands, Expressive Speech

1. Introduction

Different studies have been shown that the more convenient and acceptable modality for Smart Home users to control the ambient intelligence is the vocal commands. It is mainly relevant for weakened people, like aged people, how can have some motricity and/or visual difficulties [5]. In such kind of usages, that are nearer to assistive technologies than to comfort (it can be devoted to help aged people to keep life at home), it is usually asked to users to remember quite strictly the commands, that is the exact morpho-syntactico-lexicology of each command. In particular some distress commands are necessarily included to the home automation controls. That is that one main problem of ASR systems, in such context, is supposed to be resistive to the speech signal degradations because of the environment: long distance capture, mixing with home noises, etc. But the commands are supposed to be pronounced quite neutral. However, it can be challenged that in real use, these commands will be expressive in two ways: (1) if the user builds a representation of the intelligent apartment as a communicative entity (HAL paradigm) –it can even be embodied by a mediator robot– that is that the command prosody can implement intentions, attitudes and other interactional socio-affective values, (2) if the command is motivated by pulled emotions. It is thus clear that more the command is related to something important for the user, more the speech signal is expressive. However, if many studies are related on emotions/social affects automatic recognition [16], some evaluations on the ASR performances specifically to expressive vs. non expressive natural speech could not be found in the field on automatic spontaneous speech recognition. This paper is devoted to measure how some selected state of the art ASR systems can deal with expressive and non expressive spoken commands, since a daily system must sufficient on non expressive commands, that is for the majority of commands, but must be particularly efficient for expressive commands, of course especially for distress commands.

2. About domotic vocal commands

2.1. Smart Home and Ambient Assisted Living

The evolution of communication technologies has facilitated the emergence of new ways of designing an habitat through the concept of Smart Home. The Smart Home is a residence equipped with computer technology that anticipates and meets the needs of the occupants trying to optimally manage their comfort and safety by pressing the home and implementing connections with the outside world [18]. One of the biggest challenges in Ambient Assisted Living (AAL) is to develop smart homes that anticipate and respond to the needs of their inhabitants, this is especially important when they have disabilities. A special case is that of Smart Homes equipped with specialized devices for the detection of distress and for remote monitoring of the health [19]. The aim of the project GER'HOME [20] is to design, test and certify technical solutions supporting services to help maintain home elderly. While respecting the criteria of satisfying needs, respect for ethical and legal constraints, economic interest, the proposed systems must provide assistance in three main areas:

1. health (monitoring of health and loss autonomy),
2. safety (prevention and detection of distress),
3. and assistance with home automation (compensation of handicaps by providing easy access to household appliances).

There is a fourth aspect, the communication with relatives and outside which is essential for people who are isolated at home. More generally, Smart Homes tend to be equipped with interfaces more complex and much more difficult to control by the user. Those who benefit most from these new technologies are people losing their autonomy, people with motor disabilities or weakened by various diseases [1]. Many proposed systems are primarily concerned with monitoring, they often use sensors placed on the person and which are reserved for cases of severe disability [2] [3], cameras [4] which currently pose problems of acceptability [5] and ethics [6]. The sensors, door contacts and meters (water, electricity, heating) are widely used. Audio consideration of information is almost absent from all projects except Companionable [7], HERMES [8] and CHIL [9].

Potential users, however, are less able to use complex interfaces because of their disability or lack of familiarity with new technologies. It therefore becomes essential to provide assistance to facilitate everyday life and access to all the so-called "home automation" through Smart Home systems. The usual touch interfaces should be supplemented by more accessible interfaces, seeking neither sight nor movement, thanks to a system reactive for speech. Such systems also find their utility when, even momentarily, the person can hardly move. The sound analysis involves two distinct areas, the automatic speech recognition (voice command and dialogue) and identification of sounds. Many challenges must be

overcome before these technologies can be used on a living [10], which explains the small number of projects based on sound analysis. Regarding the particular speech, recognition is difficult because of the strong distance between the speaker and the microphone, resulting in a greater influence of reverberation, noise and environmental sounds and cone sensitivity microphone. Moreover, the results obtained with recognition systems are highly degraded in relation to children and the elderly [11]. With age, the floor is prone to earthquakes, imprecise articulation of consonants and slower articulation [12].

2.2. Cirdo context

In this context, the CIRDO project (French ANR) promotes autonomy and support for elderly people by caregivers through the social inclusion product e-lio 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. One major aim of this project is to detect distress situation of the person living alone at his home through audio and video analysis. In this framework, our objective in the project is to integrate an ASR system that will include detection of distress situations and voice commands. This is a big challenge because this includes the automatic speech recognition of emotional aged voice.

2.3. Expressivity in vocal commands

The ASR for natural speech is not a really solved problem, even if the role, that is the sub-speech, is quite known [17]. If the addressed speech is vocal commands, the main treated problems are phonologico-syntactic (reformulations, fillers, etc). But the prosodic perturbations of speech signals are not really addressed as a problem for ASR performances. On another side, many studies have been devoted to automatic emotion recognition. However, it is surely easier for a user to control a strict lexico-syntactic formulation of the command, than to control the prosodic expressivity, since some emotions are not voluntary expressed. The pragmatic problem in daily applications of home automation vocal commands, is thus how the speech recognition can be efficient both on neutral and expressive vocal commands, and even really robust to recognize the command for very expressive distress commands.

3. Methodology

In speech recognition, the acoustic models are often trained on large amount of recordings of newspaper readings or journalist speeches. In real situation, speech can be very distant from the read or journalistic speech, particularly when it is spontaneous expressive speech.

In this paper, we study the impact of emotions on ASR systems on the acoustic level. We decoded some expressive speeches in order to answer the following questions: when speech data contain emotions, does the ASR system well resist in term of accuracy by comparing with speech without emotion? Does the choice of different modalities of data used for testing and for model adaptation has an influence on the decoding accuracy?

We compared three modalities of speech:
- spontaneous expressive speech: the speaker was talking with real emotions,
- acted expressive speech: the speaker was talking by acting the emotions,
- modal speech: the speaker was talking without any emotion.

We tested the decoding on these 3 modalities by using different ASR systems, and with several acoustical models adapted on these different modalities.

The metrics of the systems analyzed in our paper are the correct-word rate (CER) and the correct-phoneme rate (CPR).

3.1. The corpora

The corpus used in this study is made of spontaneous and acted expressive speech, and of modal speech without emotion. Data of 6 speakers from the E-Wiz corpus [13] recorded in 2003 by the GIPSA laboratory were retained for the expressive speech, and we recorded, in 2012, 7 other speakers for the modal speech data.

Table 1: *Number of speakers by data category.*

Data category	Gender and number
Acted and spontaneous expressive speech	3 females 3 males
Modal speech	4 females 3 males
Total	13

3.1.1. Acted and spontaneous expressive speech corpus

The stimuli used in this study for expressive speech were extracted from the French expressive corpus E-Wiz [13]. This corpus was recorded using the Wizard of Oz technique, in which the subject is convinced to be interacting with a complex person-machine interface while the apparent behavior of the application is remote-controlled by the experimenter. Subjects were recruited with the pretext of participating in the last pre-commercialization tests of a novel voice-recognition-based language-learning application, presented as acting directly on subjects brain plasticity to enable a fast and easy learning of foreign vowels pronunciation. Most of the tasks consisted in perceptual discrimination of pairs of synthetic vowels, visually presented in the acoustic triangle. The interactions of the subjects with the system were restrained to a command language composed of the French monosyllabic color names "brique" (brick), "jaune" (yellow), "rouge" (red), "sable" (sand) and "vert" (green) and the command "page suivante" (next page), enabling the collection of at least 20 utterances of each stimulus per subject, balanced across the successive phases of the scenario. The performances attributed to the 17 subjects participating in the experiment were manipulated according to a predefined scenario. Subjects perception skills were first presented as among the better observed so far, prior to getting worse and worse. In the last step of the scenario, modified audio stimuli were presented to subjects to induce random choice of answers, while pretending that the learning software might have damaged their perceptual abilities. This scenario enabled the induction of both positive and negative emotions. The affects expressed were annotated by the subjects themselves from the video recording, as a first step before perceptual validation. An

adapted protocol was set up for the 7 subjects who were also actors: those subjects were requested immediately after the Wizard of Oz task to express on the same utterances the affects they reported to have felt during the experiment, as well as the most frequently studied emotions (sadness, anger, fear, disgust, surprise and joy), using their acting methods. The experimenters insisted that the actors should express the affects felt in the experiment the same way they had been expressing them before. The actors recruited were practicing improvisation theater and/or street acting.

The E-Wiz corpus has been used by [14], where the authors worked on the graded structure in vocal expression of emotion. They retained the productions of 6 speakers (3 males, 3 females), all actors, and they classified the expressions into 3 categories, which were referred to as ”spontaneous”, ”acted non-prototypical” and ”acted prototypical” expressions. Three broad classes of emotion, for which there were a sufficient number of utterances, were selected: anger, fear, and joy. In a pretest, they retained 193 productions of the 6 speakers, and they evaluated these productions by 15 French speaking naive listeners, who were allowed to listen to the stimuli as many times as they wanted and have to select for each one either an emotion class among the 3 proposed (anger, fear, or joy) or an ”other emotion” label. Among stimuli identified with accuracy above chance, the authors retained 12 stimuli in each emotion for each category, for a total of 108 stimuli.

Table 2: *Number of utterances by type of stimuli for acted category.*

Utterance	Anger	Fear	Joy	Total
brique	3	1	2	6
jaune	6	5	8	19
rouge	4	3	5	12
sable	4	6	0	10
vert	4	4	3	8
page suivante	6	5	6	17
Total	24	24	24	72

Table 3: *Number of utterances by type of stimuli for authentic category.*

Utterance	Anger	Fear	Joy	Total
brique	1	3	0	4
jaune	1	1	1	3
rouge	0	2	1	3
sable	1	0	0	1
vert	1	5	2	8
page suivante	8	1	8	17
Total	12	12	12	36

In our study, we kept the same 108 productions but we merged the categories ”acted non-prototypical” and ”acted prototypical” into one category that we referred to as ”acted”. The repartition of the emotions into the categories spontaneous expressive speech and acted expressive speech are given in Tables 2 and 3. These two categories can be considered as subcategories of the larger category expressive speech.

3.1.2. Modal speech corpus

To compare expressive speech with speech without expression, we collected data to construct the third category that we called ”modal speech”. We recorded 7 French speaking subjects (3 males, 4 females) for whom we asked to read the utterances with the least expressive manner possible, by using a recording application displaying the stimuli on a monitor. The stimuli were the same as for the expressive speech with the 5 monosyllabic color names ”brique”, ”jaune”, ”rouge”, ”sable”, ”vert” and the command ”page suivante”. Each speaker read almost 10 occurrences of each type of stimuli, for a total of 415 utterances as shown in Table 4.

Table 4: *Number of utterances by type of stimuli for modal category.*

Utterance	Modal
brique	70
jaune	70
rouge	70
sable	69
vert	69
page suivante	67
Total	415

3.2. The selected ASR systems

We compared the ASR performance on two decoders: the ASR toolkit Sphinx3 [15] and the Google speech web service.

The Google speech web service is an on-line service providing by the Google’s servers. It can be considered as a black box because we have no knowledge of its internal workings. It is very easy to use, but in the other hand we have access only to few basic parameters. The models are very generic, trained on a large amount of Google’s data. It is accessible via a POST request at the address:

www.google.com/speech-api/v1/recognize.

On the contrary, Sphinx3, an open source system, is used locally off-line. We can manage the configuration parameters of the ASR and can train our own models specific to the task.

3.2.1. Google speech server

Google provides a web service of speech recognition for web applications. An example of use of this web service is Google Chrome. Indeed, Google added in Google Chrome 11 the support for the HTML speech input API which allows web developers to implement speech recognition into web pages. The API calls internally this web service for processing speech to text. The web service requires a recorded audio file that is passed to the Google server via POST request, and the server responds with a JSON object containing the n-best hypothesis and their confidence scores.

3.2.2. Sphinx

With Sphinx3, we 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 which is

composed of about 100 hours of annotated speech from 120 French speakers. We called it the generic acoustic model.

3.2.3. Acoustic Model Adaptation

The most common method to overcome the ASR limitation is to apply speaker adaptation. Speaker adaptation consists in generating a new acoustic model from a generic one and some new annotated speech in limited quantity. One of the most popular technique is to apply the Maximum Likelihood Linear Regression (MLLR) which is particularly adapted when a limited amount of data per class is available. MLLR is an adaptation technique that uses small amounts of data to train a linear transform which warps the Gaussian means so as to maximize the likelihood of the data. The principle is that acoustically close classes are grouped and transformed together.

4. Experiments

4.1. Google decoding

We launched the decoding with Google speech web service on the three categories spontaneous expressive speech, acted expressive speech and modal speech. We compared the resulting 1-best hypothesis with the reference by using the NIST SCLITE toolkit for computing the CER (Table 5).

Table 5: *Correct-word rate, number of words in reference, number of correct words in hypothesis, substitutions, deletions, insertions and homophones vert/verre, by category modal speech, acted expressive speech and spontaneous expressive speech.*

Cat.	CER	Word	Corr.	Sub.	Del.	Ins.	H.
modal	65.5%	482	316	163	3	13	46
acted	39.3%	89	35	51	3	4	4
spont.	30.2%	53	16	33	4	3	6

The correct-word rates are equal to 65.5% for modal speech, 39.3% for acted expressive speech and 30.2% for spontaneous expressive speech. For expressive speech (acted+spontaneous), the CER is equal to 35.9%. We can observe an absolute difference of 29.6% between the modal speech and the expressive speech. That demonstrate the influence of the emotions, degrading the performance of decoding. We can also observe that acted expressive speech is quite better than spontaneous speech, with an absolute difference of 9.1%.

Because of mono-word utterances, sometimes Google decoded the word ”vert” (means green) by its homophone ”verre” (means glass). The percentages of word resulting in homophones vert/verre was 9.5% for modal, 4.5% for acted and 11.3% for spontaneous. If we consider than hypothesis utterances are correct if phonetically identical, the CER are respectively for modal, acted and spontaneous equal to 75.1%, 43.8% and 41.5%.

4.2. Sphinx decoding

We separated each speech category into two parts, half for acoustic adaptation, and the other half for testing. We adapted the BREF120 generic acoustic model with the data of the different speech categories.

We obtained 4 new acoustic models:
- BREF120 adapted to modal speech
- BREF120 adapted to acted expressive speech
- BREF120 adapted to spontaneous expressive speech
- BREF120 adapted to expressive speech (acted+spontaneous).

We realized with Sphinx3 a phonemic decoding by testing the different speech categories with the BREF120 acoustic model and the different adapted acoustic models. The phonemic references for comparing with hypothesis are obtained by forced alignment with Sphinx3. The results are shown in Table 6.

Table 6: *Correct-phoneme rate (CPR) for the phonemic decoding of modal speech, acted expressive speech, spontaneous expressive speech and expressive speech for the acoustic models: BREF120, adapted to modal speech, adapted to acted expressive speech, adapted to spontaneous expressive speech, adapted to expressive speech.*

AM/Test	modal	acted	spont.	emotions
BREF120	69.83%	58.37%	46.21%	53.82%
modal	81.90%	66.36%	57.94%	63.24%
acted	77.64%	66.35%	60.00%	63.99%
spont.	79.48%	65.24%	64.80%	65.07%
emotions	80.30%	67.46%	62.40%	65.57%

We can observe that without adaptation (BREF120 generic acoustic model), the test with modal data gives the best CPR (69.83%), the worst CPR is given by spontaneous expressive speech (46.21%), and for the acted expressive speech, the CPR is 58.37%. The ranking of the speech categories is the same than with the Google decoding. The absolute difference of CPR between modal speech and expressive speech is 16.01%.

In all the cases, the acoustic adaptation improves the results.

For the decoding of spontaneous speech expressive speech, the using of the acoustic model adapted with spontaneous expressive speech produces the best result (CPR=64.80%) compared to the other adapted models. The recording of modal speech by reading text is obviously easier than recording spontaneous expressive speech, but the adaptation with modal speech gives the worst result (CPR=57.94%). The adaptation with acted expressive speech produces CPR=60.00%, less than the result with spontaneous expressive speech, showing that the properties of acted expressive speech are different than the properties of spontaneous expressive speech.

If the models are not adapted with the same category of data that the category of data used for testing, the decoding with the model adapted to expressive speech permits to increase the results for all the speech categories.

For the decoding of acted expressive speech, the best model is the one adapted with expressive speech (CPR=67.46%).

When testing with modal speech, the less efficient model is the model adapted to acted expressive speech (CPR=77.64%).

Finally, when decoding with expressive speech, the best model is the model adapted with expressive speech (CPR=65.57%) and the worst is the model adapted with modal speech (CPR=63.24%).

5. Conclusion

In this study, we addressed the pragmatic problem of how some ASR systems, efficient for neutral commands – that are

not perturbed by morpho-syntactic reformulations, neither fillers – can deal in real-life home automation commands, when the utterances are mixed within expressive and non expressive commands. This problem cannot be related with the automatic recognition of emotions, since the lexical recognition of commands is a sufficient issue, even for distress commands (the distress information is largely contained by the lexico-morpho-syntactic recognition). We could observe, but mixing all the combinations of expressive/non expressive corpus for the systems training, and the systems running, that there are many variations of ASR performances depending on how are trained the system, whatever the ASR tested system. That means that for real applications in home automation commands, especially for weakened users, this problem must absolutely be taken into account, and the ASR systems must be specifically adapted. The question is opened to understand if the ASR could be helped by the emotion automatic recognition processing (that are not directly required).

6. Acknowledgement

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 the volunteer subjects who accepted to perform this experimentation and give their time to do so correctly.

7. References

[1] Z. Callejas and R. Lopez-Cozar, "Designing smart home interface for the elderly," Accessibility and Computing, vol. 95, pp. 10–16, 2009.

[2] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Trster, "Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection." in European conference on Wireless Sensor Networks, ser. Lecture Notes in Computer Science, R. Verdone, Ed., vol. 4913. Springer, 2008, pp. 17–33.

[3] Y. Charlon, W. Bourennane, and E. Campo, "Mise en œuvre d'une plateforme de suivi de l'actimétrie associée à un système d'identification," in Symposium Mobilité et Santé (SMS 2011), 2011.

[4] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," ACM Computer Survey, vol. 43, pp. 1–43, 2011.

[5] 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. [Online]. Available: http://dx.doi.org/10.1007/s00779-011-0470-5

[6] E. M. Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," Pervasive Computing, vol. 2, pp. 158–175, 2004.

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

[8] A. G. Jianmin Jiang and S. Zhang, "HERMES: A FP7 funded project towards computer-aided memory management via intelligent computations," in 3rd Symposium of Ubiquitous Computing and Ambient Intelligence, 2009, pp. 249–253.

[9] O. Brdiczka, M. Langet, J. Maisonnasse, and J. Crowley, "Detecting human behavior models from multimodal observation in a smart home," IEEE Transactions on Automation Science and Engineering, vol. 6, no. 4, pp. 588 –597, oct. 2009.

[10] 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.

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

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

[13] V. Aubergé, N. Audibert, and Rilliard, "E-wiz: A trapper protocol for hunting the expressive speech corpora in lab," in LREC 2004, Lisbon, Portugal, 2004, pp. 179–182.

[14] P. Laukka and N. Audibert, "Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"?" in Proceedings of the International Workshop on Paralinguistic Speech, ParaLing'07, Saarbrücken, Allemagne, Aug. 2007, pp. p. 1–4.

[15] 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.

[16] B. Schuller, A. Batliner, S. Steidl and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge", Speech Communication, Volume 53, Issues 9–10, pp. 1062-1087, 2011.

[17] R. Dufour., Y. Estève and P. Deléglise, "Investigation of spontaneous speech characterization applied to Speaker role recognition", Interspeech, 2011.

[18] F. Aldrich, "Smart homes: Past, present and future". In R. Harper, ed., Inside the Smart Home, pp. 17-39. Springer London, 2003

[19] M. Chan, D. Estèves, C. Escriba and E. Campo, "A review of smart homes- present state and future challenges," in Computer Methods and Programs in Biomedicine, 91(1):55-81, 2008.

[20] 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," in Gerontechnology Journal, 8(3):129-139, 2009.