Cross-Corpus Experiments on Laughter and Emotion Detection in HRI with Elderly People

Marie Tahon[1(✉)], Mohamed A. Sehili[1], and Laurence Devillers[1,2]

[1] Human-Machine Communication Department, LIMSI-CNRS, 91403 Orsay, France
marie.tahon@limsi.fr
[2] University Paris-Sorbonne IV, 28 Rue Serpente, 75006 Paris, France

Abstract. Social Signal Processing such as laughter or emotion detection is a very important issue, particularly in the field of human-robot interaction (HRI). At the moment, very few studies exist on elderly-people's voices and social markers in real-life HRI situations. This paper presents a cross-corpus study with two realistic corpora featuring elderly people (ROMEO2 and ARMEN) and two corpora collected in laboratory conditions with young adults (JEMO and OFFICE). The goal of this experiment is to assess how good data from one given corpus can be used as a training set for another corpus, with a specific focus on elderly people voices. First, clear differences between elderly people real-life data and young adults laboratory data are shown on acoustic feature distributions (such as F_0 standard deviation or local jitter). Second, cross-corpus emotion recognition experiments show that elderly people real-life corpora are much more complex than laboratory corpora. Surprisingly, modeling emotions with an elderly people corpus do not generalize to another elderly people corpus collected in the same acoustic conditions but with different speakers. Our last result is that laboratory laughter is quite homogeneous across corpora but this is not the case for elderly people real-life laughter.

Keywords: Laughter recognition · Emotion recognition · Human-Robot Interaction · Elderly people · Cross-corpus protocol

1 Introduction

Assistive social robots must be able to decode verbal and non-verbal expressions of the user. The success of a social robot also relies on its ability to rightly interpret the inputs and properly react to them. In such a context, social signal processing designs high level cues which describe conversations, user profiles and engagement [1] during Human-Robot interactions. For example, social and interactional markers extracted from speech signal can be used to build up a user profile [2].

The authors are working under the French project ROMEO2[1] which aims at building a 140 cm high humanoid social robot. The robot is designed to be

[1] http://projetromeo.com

A. Tapus et al. (Eds.): ICSR 2015, LNAI 9388, pp. 633–642, 2015.
DOI: 10.1007/978-3-319-25554-5_63

a friendly assistant robot for non-autonomous people such as elderly people. It will be able to adapt its behavior but also to build user profiles. This project faces two main issues: social cues must be 1) robust to realistic and unseen data (spontaneous speech, noisy environments, uncontrolled acoustics), 2) adapted to non-autonomous users, especially elderly-people. The present study focuses on the decoding of two social markers of elderly-people interacting with a robot using speech input: affective states [3] and laughter [4].

The two main drawbacks of the standard corpora used in the community are the very small size of audio corpora and data variability in terms of task, speaker, age and audio environment which compromises the significance of results and improvements [5]. As a consequence, there is a critical need for data collection with end-users (with different types of speakers, ages) and real tasks for emotion recognition systems since realistic emotions could not be found in acted databases [6]. So far, very few HRI databases have been collected with diverse kind of participants: children (AIBO [7] and NAO-HR [8]), young adults (SEMAINE [9]) or visually-impaired people (IDV-HR [10]). At the present time, very few real-life emotional speech databases were recorded with elderly people: ARMEN [11] and ROMEO2 [12]. Speaker identification has been shown to be easier on elderly people than on young adults [13] because voice quality is very different between these two age groups (creaky voice, low loudness, voice pathology, etc.).

Because social markers extraction must be robust to unseen data, the present study features cross-corpus experiments which also ensures speaker independent conditions. It consists of using one corpus for modeling emotion and laughter and another one as test set. A third corpus is eventually used for development purposes. By this way, recognition rates are lower but more realistic than with cross-validation experiments. Schuller et al. [14] performed binary valence recognition with cross-corpus experiment on seven corpora. Average recalls are slightly over the random guess, from 50% to 55% with young adults. A previous experiment on children and adults voices [10] has shown a possible merging between children voices corpora, however it seems more complex to merge adult speakers and children speakers. A lot of interesting work on laughter detection in HRI has been reported in the ILHAIRE project[2]. But, as far as the authors know, none of them has been done in cross-corpus. Recently, a cross-corpus experiment on laughter was carried on three spontaneous HRI corpora [15]. The goal of the presented cross-corpus experiment is to assess how data from one given corpus can generalize to another corpus, variability being expressed under the project ROMEO2, in terms of age and acoustic conditions. Two groups are tested: one is composed of young adults recorded in laboratory conditions (OFFICE and JEMO [16] corpora), the other one is an elderly people's recorded in real-life conditions (ARMEN [11] and ROMEO2 [12] corpora).

Section 2 summarizes the acoustic features used for emotions and laughter modeling. The four French HRI databases are described in section 3.

[2] www.ilhaire.eu/project

Methodology and results are presented in section 4. The conclusion is drawn in the last section.

2 Acoustic Cues

In this work, many acoustic features are used to model laughter and emotions in voice. These features globally carry three kinds of information: spectral information, temporal shape information, and voice quality. Such acoustic features are supposed to carry most of emotional information [17], [18]. Several studies [19], [20], [21] found that fundamental frequency, instance duration energy and formants are also relevant for clear and well-identified laughters.

Spectral and temporal shape information is extracted using Yaafe[3] and contains perceptual features, ZCR (Zero Crossing Rate) and 24 Specific Loudness Energy bands. A total of 10 statistical coefficients (SetFunc) are calculated for each vector attribute. Prosodic and phonetic information is extracted with Praat[4]. Pitch-related features include mean, standard deviation, maximum and minimum of pitch (extracted in semitones). Intra (respectively inter) pitch is the pitch difference within a voice region (respectively across consecutive voice regions) and glissando. Formant-related features are: mean and standard deviation of the three first formants, mean and standard deviation of the formant differences F_2-F_1 and F_3-F_2. Micro-prosody features are: jitter, shimmer, HNR and proportion of voiced parts in the segment. More details on these acoustic features can be found in [22]. The extraction step yields a 301-dimension vector per audio segment as summarized in table 1.

Table 1. Acoustic feature set: 301 features. SetFunc is a set of 10 functionals: mean, std, slope and high-level statistics. std stands for standard deviation.

LLD	functionals	Nb func.
ZCR	SetFunc	10
Roll Off 95%	SetFunc	10
Spectral Slope	SetFunc	10
Spectral Flatness	SetFunc	10
Specific Loudness 1-24	SetFunc	24×10
Pitch	mean, max, min, std, intra, inter, glissando	7
Formants	mean, std F_1, F_2, F_3	6
	mean, std $F_2 - F_1$, $F_3 - F_2$	4
Micro-prosody	local jitter, local shimmer, HNR, punvoiced	4

3 Databases

The four databases used in the following cross-corpus experiments, are presented in this section. Two of them, ARMEN and ROMEO2, were collected with elderly-people during HRI (60 speakers of more than 60 years old). The other two,

[3] http://yaafe.sourceforge.net/
[4] http://www.fon.hum.uva.nl/praat/

JEMO and OFFICE (66 speakers of less than 60 years old), were collected during emotion games. The four corpora are in French and there is no lexical constraints. All corpora were manually segmented and annotations were performed by two expert annotators. Only consensual emotional segments are used in this work.

3.1 ROMEO2 Corpus

The ROMEO2 corpus [12] was collected in a French EHPAD (public accommodation for non-autonomous old people). 27 participants (3 men and 24 women) were recorded. A Wizard-of-Oz scheme controls the robot so that its behavior adapts seamlessly and quickly to most situations. Each interaction was split into different scenarios: greetings, reminder events (take medicine), social interaction (call a relative) and cognitive simulation (song recognition game). This corpus is very rich in terms of elderly-people speech. The study of interactions with elderly people also suppose to deal with hearing difficulties. The consensual data constitute 98 min of emotional instances.

3.2 ARMEN Corpus

The ARMEN corpus was collected in a French EHPAD within the ANR Tescan ARMEN. 77 patients from medical centers (elderly and impaired people), of which 48 men and 29 women between 18 and 90 years old participated in this data collection. The consensual data constitute about 70 minutes of the corpus. The collected data are used to explore approaches which aim at resolving the performance generalization problem of emotion detection systems run on different data [11]. In the present paper, the authors use a subset of ARMEN that contains elderly speakers only (36 speakers over 60 year old).

3.3 OFFICE Corpus

The OFFICE corpus was collected with two scenarios (jokes and emotion game) written in order to spark emotional speech and laughter. 7 speakers from 18 to 52 were recorded at LIMSI with a high-quality microphone during an interaction with the robot Nao [15]. In the "joke" scenario, the robot tells jokes in order to provoke a user's laughter. In the "emotion game" scenario, the user is asked to act emotions (anger, sadness, happiness or neutral state) so as to be recognized by the robot. The collected data contain emotional speech and affect bursts (laughter) but also noise, cough and blow (breathing or blowing). Each record was then segmented and transcribed, the number of segments per emotional class and affect bursts is summarized in table 2.

3.4 JEMO Corpus

The JEMO corpus was recorded in laboratory conditions to obtain emotions in the context of a game within the ANR Affective Avatar project. The goal of

the game was to make the machine recognize an emotion (anger, joy, sadness or neutral state) without providing any context [16]. The lexical content was totally unconstrained, and the speaker tried and modulate freely their emotional expressions so as to be recognized by the system. As a result, the participants produced very expressive emotions in order to be as close as possible to the entries expected by the system. The corpus contains thus prototypical emotions produced in a "game" scenario. The total duration of the corpus is 41 minutes and it includes 59 participants (30 men and 29 women aged from 16 to 48 y. o.)

3.5 Characteristics of the Databases

The databases described previously mainly contain, besides laughter, the four Ekman's emotions: neutral state, anger, positive state, sadness. Since the ROMEO2 corpus has a very small number of anger instances, only positive, neutral states and sadness will be modeled in the present study. In the presented corpora, laughter can suppose either positive feelings (joy, amusement, etc.) or negative states (such as contempt [23], sadness or embarrassment). The number of consensual instances for each emotional class used in this work is shown in table 2.

Table 2. Content description for each data corpus. POS: positive, NEG: negative, NEU: neutral, SPE: total speech, LAU: laughter (non-speech).

Corpus	# Subjects	Age	Duration	# Segments				
				POS	NEG	NEU	SPE	LAU
ARMEN	36	60-90	68 min	308	64	1162	1534	253
ROMEO2	24	75-99	98 min	673	404	1306	2583	205
OFFICE	7	18-50	10 min	107	134	62	303	123
JEMO	59	16-48	29 min	201	307	341	849	73

ARMEN and ROMEO2 are elderly people real-life databases collected with similar acoustic environments (same EHPAD) with similar protocols, but different speakers. One is collected with a humanoid robot (ROMEO2), the other with a virtual agent (ARMEN). JEMO and OFFICE were collected in the same laboratory conditions but with different speakers and protocols.

4 Cross-Corpus Experiments with Elderly and Young People Voices

The goal of this experiment is to assess how data from one given corpus can generalize to another corpus. The inter-corpus variability that interests us here, is expressed in terms of age and acoustic conditions. Two groups are tested: one is composed of young adults, the other of elderly people. Four acoustic conditions are tested which correspond to the four corpora.

4.1 Comparison of Acoustic Features Between Elderly and Young Adults People

Elderly people speech contains tremor, pitch breaks, a lot of hesitations and fillers. Speakers' voice quality is also different from that of young adults. Figure 1 shows pitch standard deviation and local jitter distributions across corpora. While local jitter distributions are almost the same for the four corpora, F_0 standard deviation reaches significantly higher values in elderly people real-life voices than in laboratory young voices. This result shows that looking for relevant acoustic features which are good for distinguishing young and elderly people voices, is a real challenge. In the present study, age and acoustic conditions are mixed together because available corpora are not big enough to analyze all conditions separately. A previous study showed that speaker recognition was easier for elderly than for young speakers [13]. Our hypothesis is that acoustic features change more with age group condition than with acoustic environment, but further investigations are needed.

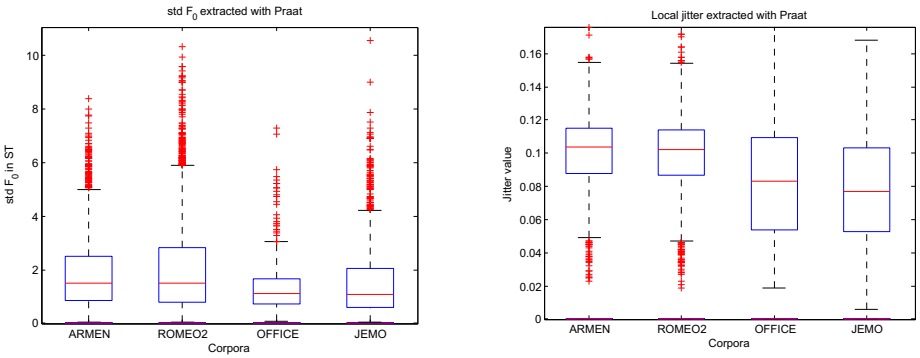

Fig. 1. Feature distributions across the four corpora: std F_0 (left), local jitter (right).

4.2 Methodology: Cross-Corpus Experiments

Emotion and laughter cross-corpus experiments are realized following the same protocol. ROMEO2 and JEMO corpora have been equally divided into three subsets: one for training (C1), one for development purposes (C2) and a last one for testing (C3). The three subsets are randomly composed so that they have the same number of segments for given class. Thus, by using JEMO or OFFICE (young subjects), ARMEN or ROMEO2 (elderly people) as train corpora and ROMEO2 or JEMO as test corpora, we actually want to check how age divergence and acoustic conditions variability affect the recognition performance. Good rates are expected when train and test data are from similar age groups, whereas lower rates are expected when train and test data belong to different age groups.

The cross-corpus protocol ensures speaker independent conditions, expect when training and testing on the same corpus (baseline). The subjects are not equally represented in each subset.

Automatic classification is performed with SVM (Support Vector Machines) using libsvm[5]. Classification was run with a linear or RBF (Radial Basis Function) kernel with parameter optimization on development subsets. Results are given in terms of UAR (Unweighted Average Recall). The confidence interval depends on the number of the tested segments N and the obtained performance UAR (equation 1).

$$Confidence = UAR \pm 1.96\sqrt{\frac{UAR \times (1 - UAR)}{N}} \tag{1}$$

4.3 Cross-Corpus Results

The results of the cross-corpus experiments are reported in table 3. Experiments conducted with the same corpus for both training and testing (baseline condition) are reported in bold, they serve for comparison with cross-corpus experiments results.

Table 3. Cross-corpus UAR ± confidence results for emotion and laughter recognition, baseline in bold. # is the number of tested instances (a third of the initial corpus).

Train	Test			
	NEU/NEG/POS		SPE/LAU	
	ROMEO2-C3 (#793)	JEMO-C3 (#282)	ROMEO2-C3 (#862)	JEMO-C3 (#307)
ARMEN	39.2 ± 3.4	40.6 ± 5.7	67.0 ± 3.1	69.1 ± 5.2
OFFICE	44.7 ± 3.5	44.2 ± 5.8	59.2 ± 3.3	81.6 ± 4.3
ROMEO2-C1	**46.3 ± 3.5**	42.0 ± 5.8	**87.2 ± 2.2**	71.3 ± 5.1
JEMO-C1	40.7 ± 3.4	**61.2 ± 5.7**	68.3 ± 3.1	**82.3 ± 4.3**

Emotion Recognition Results. In the context of emotion recognition, the baseline performances obtained with both ROMEO2 and JEMO corpora, are the highest. Using data from the same corpus for training and testing not only yields the best performance but also seems to lead to a fairly more balanced recall between the three classes of emotion. For example, with OFFICE for training and JEMO-C3 for testing the minimum recall is reached by the neutral class at 8.9% (probably because there is very few neutral instances); with JEMO-C1 for training, the minimum recall is reached by the negative class at 57.8%. The recognition rates are lower while testing on ROMEO2 than testing on JEMO. This is due to the fact that JEMO is prototypical while ROMEO2 is real-life.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

The recognition rate obtained with models trained on ARMEN and tested with ROMEO2-C3 was expected to be similar to the one obtained with models trained on ROMEO2-C1. This is actually not the case (UAR=39.2% with ARMEN for training and UAR=46.3% with ROMEO2-C1), thus denying our hypothesis.

Based on these results, the use of other elderly people corpus for training emotions does not help improving the performances when testing on elderly people. However, when testing on JEMO-C3, all training corpora, give similar results. Elderly people real-life corpora are much more complex than laboratory corpora, and they are significantly different one from another (between ARMEN and ROMEO2).

Laughter Recognition Results. Similar results are obtained on cross-corpus laughter recognition. The recognition of ROMEO2 (respectively JEMO) laughter is better if the model is trained with similar data (with ROMEO2-C1 sub-corpus (respectively with JEMO-C1)). However, in cross-corpus conditions, building a model with elderly people is not necessary when testing on elderly people: the best performance is obtained with JEMO-C1, then comes OFFICE and last is ARMEN.
The use of elderly people voices for training the models degrades the recognition rates (with ARMEN and ROMEO2-C1). Training a laughter model with the corpus OFFICE leads to a performance similar to the baseline. One of the main conclusions of these experiments on laughter is that JEMO and OFFICE laughters are acoustically homogeneous, however, they differ from ARMEN's and ROMEO2's. Despite the small size of the OFFICE corpus and the absence of very aged subjects, it performs better than ARMEN, be that against ROMEO2 or JEMO.

It seems that laughter is significantly different on one hand between prototypical corpora and real-life corpora, and on the other hand between two different real-life corpora. Laboratory laughter is quite homogeneous across corpora (between OFFICE and JEMO) but this is not the case for elderly people's real-life laughter.

5 Conclusion

The study gives some pilot results with elderly-people voices during interaction with a robot. Two social markers which are very useful in HRI, are detected: laughs and emotions. The automatic recognition of these two markers is presented in cross-corpus conditions. Four corpora are used in the experiments: two of them were collected with young adults (JEMO and OFFICE) and the other two with elderly people (ROMEO2 and ARMEN) during HRI. Our goal was to assess how data from one given corpus can generalize to another corpus, variability being expressed in terms of age and acoustic conditions.

Our first main result is that a comparison of acoustic features (such as F_0 standard deviation or local jitter) distributions across corpora, show clear differences between age and acoustic environments groups. This result confirms

the fact that speaker recognition best performs on elder adults [13]. The second result obtained with cross-corpus experiments on emotion recognition is that elderly people real-life corpora are much more complex than laboratory corpora and they are significantly different one from another (ARMEN and ROMEO2). Surprisingly, modeling emotions with an elderly people corpus do not generalize to another elderly people corpus collected in the same acoustic conditions (here same EPHADs) but with different speakers. Our last result is that laboratory laughter is quite homogeneous across corpora (JEMO and OFFICE) but this is not the case for elderly people real-life laughter.

The complexity of elderly people real-life corpora may be due to age group and emotional behavior. This study shows that modeling emotions with an elderly people corpus do not generalize to another elderly people corpus even if the training and testing corpora are collected within the same acoustic environments and with similar scenarios. Further experiments are needed to investigate the advantage of merging elderly and young people real-life corpora or building separate models. The authors use available corpora, therefore further experiments with new HRI corpora are needed to dissociate the effect of age group on acoustic features independently from the acoustic environment.

Acknowledgments. This work was financed by the French project BPI ROMEO2. The authors thank the association APPROCHE and the EHPADs for their help in corpus collection.

References

1. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signals, their function, and automatic analysis: a survey. In: Conference on Multimodal Interfaces (ACM), Chania, Greece, pp. 61–68 (2008)
2. Delaborde, A., Devillers, L.: Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers. In: International Workshop on Affective Interaction in Natural Environements (AFFINE), Firenze, Italy (2010)
3. Breazeal, C.: Emotion and sociable humanoid robots. Human Computer Studies **59**, 119–155 (2003)
4. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. ACM Transactions on Interactive Intelligent Systems (TiiS) **2**(1), Article No. 4 (2012). Special Issue on Affective Interaction in Natural Environments
5. 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), 1062–1087 (2011). Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing
6. Batliner, A., Steidl, S., Nöth, E.: Laryngealizations and emotions: how many babushkas? In: Proc. Internat. Workshop on Paralinguistic Speech - Between Models and Data (ParaLing' 07), Saarbrucken, Germany, pp. 17–22 (2007)

7. Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., Wong, M.: You stupid tin box - children interacting with the aibo robot: a cross-linguistic emotional speech corpus. In: LREC, Lisbon, Portugal, pp. 171–174 (2004)
8. Delaborde, A., Tahon, M., Barras, C., Devillers, L.: Affective links in a child-robot interaction. In: LREC, Valletta, Malta (2010)
9. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schröder, M.: The semaine database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. IEEE Transactions on Affective Computing **3**(1), 5–17 (2012)
10. Tahon, M., Delaborde, A., Devillers, L.: Real-life emotion detection from speech in human-robot interaction: experiments across diverse corpora with child and adult voices. In: Interspeech, Firenze, Italia (2011)
11. Chastagnol, C., Clavel, C., Courgeon, M., Devillers, L.: Designing an emotion detection system for a socially-intelligent human-robot interaction. In: Towards a Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice. Springer (2013)
12. Sehili, M.A., Yang, F., Leynaert, V., Devillers, L.: A corpus of social interaction between nao and elderly people. In: International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Satellite of LREC (2014)
13. Tahon, M., Delaborde, A., Barras, C., Devillers, L.: A corpus for identification of speakers and their emotions. In: LREC, Valletta, Malta (2010)
14. Schuller, B., Zhang, Z., Weninger, F., Rigoll, G.: Selecting training data for cross-corpus speech emotion recognition: prototypicality vs. generalization. In: AVIOS Speech Processing, Tel-Aviv, Israel (2011)
15. Tahon, M., Devillers, L.: Laughter detection for on-line human-robot interaction. In: Interdisciplinary Workshop on Laughter and Non-verbal Vocalisations in Speech, Enschede, Netherlands (2015)
16. Brendel, M., Zaccarelli, R., Devillers, L.: Building a system for emotions detection from speech to control an affective avatar. In: LREC, Valletta, Malta (2010)
17. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: Ressources, features and methods. Speech Communication **48**(9), 1162–1181 (2006)
18. Schuller, B., Batliner, A.: Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. John Wiley & Sons (2013)
19. Bachorowski, J.-A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. Journal of the Acoustical Society of America **110**(3), 1581–1597 (2001)
20. Campbell, N.: Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation. In: International Conference on Spoken Language Processing, Jeju Island, Korea (2004)
21. Szameitat, D.P., Darwin, C.J., Szameitat, A.J., Wildgruber, D., Alter, K.: Formant characteristics of human laughter. Journal of Voice **25**(1), 32–38 (2011)
22. Devillers, L., Tahon, M., Sehili, M., Delaborde, A.: Inference of human beings' emotional states from speech in human-robot interactions. International Journal of Social Robotics, Special Issue on Developmental Social Robotics (in press, 2015)
23. Schröder, M.: Experimental study of affect bursts. Speech Communication **40**(1–2), 99–116 (2003). Special Session on Speech and Emotion