Real-life Emotion Detection from Speech in Human-Robot Interaction: Experiments across Diverse Corpora with Child and Adult Voices

Marie Tahon, Agnes Delaborde, Laurence Devillers

Real-life Emotion Detection from Speech in Human-Robot Interaction: Experiments across Diverse Corpora with Child and Adult Voices

Marie Tahon, Agnes Delaborde, Laurence Devillers

Department of Human-Communication, LIMSI-CNRS, France
Department of Computer Sciences, University Paris 11
{marie.tahon, agnes.delaborde, laurence.devil}@limsi.fr

Abstract

We focus in this paper on the detection of the emotions in the voice of a speaker in a Human-Robot Interaction context. This work is part of the ROMEO project, which aims to design a robot for both elderly people and children. Our system offers several modules based on a multi-level processing of the audio cues. The affective markers produced by these different modules will allow to pilot the emotional behaviour of the robot. Since the models are built with recording data and the system will test real-life data, we need to estimate our emotion detection system performances in cross-corpus situations. Cross-validation experiments on a three class detection show that derivatives and energy features may be removed from our feature set for this specific task. Cross-corpora experiments on anger-positive-neutral data suggest that detection performances may be better with two different models: one for child voices, one for adult voices.

Index Terms: emotion detection, human-robot interaction, cross-corpus, realistic corpus

1. Introduction

Our study focus is Human-Robot Interaction. We are involved in the French ROMEO project[1], which aims to design a 1m40 high social humanoid robot which will be able to assist elderly and disabled persons at home in everyday life activities, but also which will be able to play games with children (for example with the grand-children of the user). So as to interact as naturally as possible with the user, the robot will be endowed with an emotional system. It will allow it to be able to adapt its behaviour according to its own emotional state, and will be sensitive to the user's emotional state as well. The robot will evolve in real-life conditions and then face a rich multimodal contextual environment which needs to be processed. In the community, robotic emotional systems are endowed with a processing of visual, tactile or audio inputs. Non-verbal cues are useful at different levels: in maintaining the communication (backchannel), in the comprehension of the message (signs of agreement/disagreement) but also in an interpersonal affective dimension: in the short term (positive/negative emotion) or longer term (affective dispositions). We can refer to works related to WE-3RV by Miwa et al. [1] dealing on visual and audio inputs (low or loud sounds), studies on iCat focused on the facial expressions of the participant [2] or Kismet's color, face and motion detection [3]. Nonetheless, emotions carried by speech are seldom used in Human-Robot Interaction (HRI).

We focus on designing an audio detection system for a HRI interface, which will process the audio cues at different levels. The system is based on several modules, each focusing on different specific affective measures: emotion label, activation, valence, speaking rate, presence of affect bursts, etc. These different affective markers will be used for driving the behaviour of the robot.

In the framework of the final application, the HRI system will be supplied with both adult and child voices. In order to train our models, we need data related to our final applicative context, featuring children and adults. Our system is supposed to detect emotions on an audio file recorded in unknown acoustic conditions. In order to estimate the robustness of our emotion detection system to different recording conditions, we have done several cross-corpora experiments. Another method to improve robustness to different recording conditions would be to adapt our models during the interaction, but this aspect will not be treated in this paper. In the community, there are few available realistic HRI corpora, the best known being the AIBO corpus, which is a collection of 51 children interacting with Sony's pet robot Aibo in a specific context [4]. In our context, we try to use this corpus for building models for children and we also collected new data corresponding more precisely to our applicative context. The focus of the paper is to study the performances of the anger-positive-neutral model we can build on the databases at our disposal. A next step will be to build all the models to give predictions which can be used in the HRI emotion detection system (active/passive, positive/negative, anger against the remaining emotions, etc.).

The normal approach in emotion detection from speech is to subdivide one corpus in two sets: one for training the model, and the other for testing. When using only one corpus, most variables are constant: microphone, room acoustics, sampling frequency, speaker group, annotations, etc. In the case of realistic corpora, variables vary much more than in acted or prototypical corpora [5]. Then, doing cross-corpora using realistic corpora is challenging. One of the first studies on cross-corpora classification of realistic emotions [6] shows that significant improvement over random guess is observed in a few cases for valence classification. Normalisation in the context of cross-corpus evaluation is an important challenge, as well as canceling the recording conditions effects. We have tested the normalisation to speaker; as [5] shows it was the better. In the context of the ROMEO project, we wonder if one single model for both children and adults is enough for emotion detection, or if we need to have two distinct models. A first approach is to test performances in cross-corpus classification: training on the adults and testing on children, or the contrary. A second method that will not be studied here, would be to test leave-one-speaker-out on both corpora

Section 2 deals with the final emotion detection system we intend to build for the HRI. Section 3 presents the new corpora we collected in the final application context. Section 4 summarises first experiments on performances of the three corpora: NAO-HR, IDV-HR, and AIBO, then experiments on cross-corpora classification.

[1] http://www.projetromeo.com

2. Multi-level Processing of Emotional Audio Cues in Human-Robot Interaction

We presented in a previous study on the emotional and interactional markers [7] a modeling of the emotional social interaction between a Human and a Robot. We argue that a multi-level use of audio non-verbal cues contributes to an efficient piloting of the decisions of the robot. Low level cues can be computed from the speech signal [8]: duration of speaker turns, F0, energy, and other acoustic coefficients such as MFCC, etc. Multi-level markers can be derived from these cues and provide a system with emotional information such as positive/negative emotion, activation/non activation behaviour, emotion labels (Joy, Sadness, Fear, Anger), speech delivery, rhythm and duration. On a higher level of analysis, these data can be processed so as to get cues about the emotional and interactional tendencies of the speaker: we can obtain emotional and interactional markers such as ill-at-ease, talkative, shy, or dominant. This multi-level processing is presented in Figure 1. A speaker identification system would also bring sociological metadata such as the age bracket of the speaker, the sex, and to be able to recognise a specific user and then keep an automatic track of his or her emotional and interactional profile.

The detection of the expressed emotions is then organised in two levels: the first level is the single speaker turn, when the emotion is immediately processed, and the other level requires the use of a history gathering the markers of the emotions expressed by the speaker, after several speaker turns and also the use of a history of the reactions of the robot. This emotional and interactional profile will be a basis for the selection of the most desirable behaviour of the robot towards the user, depending on the context of interaction. In this study, we look into the performances of short-term emotion detection.

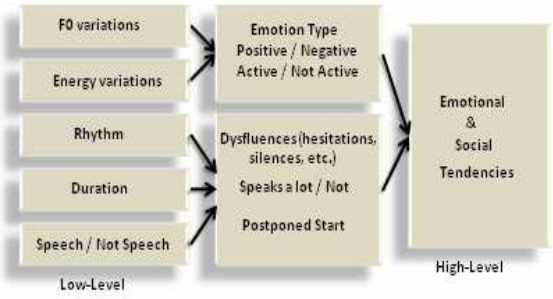

Figure 1: *Multi-level detection of the emotional and interactional cues from speech (from [7])*

3. Data Collection

A first data collection, NAO-HR [7], features children playing games with the humanoid robot Nao, and a second study concerns visually-impaired elderly people in the context of an interaction with Nao as a robotic domestic assistant (IDV-HR). In both experiments, different communicational strategies of the robot were applied so as to induce different emotional reactions in the participants.

3.1. Recording Protocol

In NAO-HR, each child carries a high quality lapel-microphone (AKG PT40 Pro Flexx) which records the totality of the session.The sampling frequency is 16kHz. The video recording of the children is used for potential verifications, and as a showcase for our studies. Each participant of the IDV-HR data collection is offered to sit comfortably in front of NAO, which is sitting down on a coffee table. The participant is recorded with the same lapel-microphone. The sampling frequency is 44kHz. A camera is placed behind the robot and films the upper part of the body of the speaker for further studies.

3.2. Wizard of Oz

3.2.1. Scenarii with Children

We designed Wizard-of-Oz systems which allow us to gather spontaneous emotional data, in task-related contexts. In NAO-HR, pairs of two children aged between eight and thirteen played games with the robot. In a first game, both the children and the robot play a question-answer game, arbitrated by a human game master (experimenter). The second and third game present the robot as capable of recognizing emotions and songs, and each child has to act emotions or hum songs in such a way as to be recognized by the robot. The communicational strategies applied by the robot in the course of these games are divided into positive (desirable) strategies and negatives (undesirable) ones: depending on the moment of the experiment, the robot encourages or congratulates the children, as well as triggers competition or presents some technical failures (repetitions, crashes). An experimenter controls the robot from and adjacent room and selects the behaviours which are to be played.

3.2.2. Scenarii with Visually Impaired People

The corpus IDV-HR features elderly people interacting with the robot. The speaker is asked to play three sessions of five scenarios in which he pictures himself in a situation of waking up in the morning. The robot would come to him to chat about either his health, or the program of the day, etc. The utterances of the robot are spoken through a Speech-to-Text module, and are based on pre-established and fixed sentences. Each of these five scenarios is devoted to a different affective state, which the speaker is asked by the robot to act: well-being, minor illness, depressed, medical distress, happy. Each series of five scenarios differ from the other, by the social attitude of the robot (positive: friendly, empathetic, encouraging, or negative: directive, doubtful, machine-like).The robot is remotely controlled by an experimenter who selects the different social attitudes and the utterances which match the content of the speaker's speech the best.

3.3. Annotation Protocol

On each speaker's track, we define segment boundaries. A segment is emotionally homogenous, i.e. the emotion is considered as being the same and of a constant intensity along the segment [9]. Two expert annotators perceptively annotated (as far as possible they did not take into account the semantic content) the segment along this annotation scheme:

- Three **affective state labels** describe each emotional segment. The resulting annotation of a segment can then represent complex emotions (such as both positive and negative expressions). The affective state labels are grouped together in five macro-classes: positive, anger, sadness, fear, neutral.
- **Valence**: On the whole, does the speaker feel a positive or a negative sensation? *positive; negative; either positive or negative; positive and negative; valence non decidable*
- **Activation**: the strength of the expressed emotion. Scale of -2 to 2, from *very weak* to *very intense*

The annotation scheme also includes interactional dimensions for other studies relative to the HRI.

3.4. Description of the Corpora

The NAO-HR corpus is made up of 1287 emotional segments, for a total amount of 31'07". Ten children were recorded (five males and five females from eight to thirteen). 22 speakers were recorded in the framework of IDV-HR (11 males and 11 females for a median age of 59). So far, 8 speaker sessions were emotionally labelled, for an amount of 2198 emotional segments (1h 20'). Table I presents the inter-speaker agreement scores for the annotations of the macro-classes for both corpora: the Kappa values are computed first on all the macro-classes. However, the NAO-HR corpus offers only few instances of fear and sadness. Moreover, so as to allow the comparison between our data and the corpus AIBO, we need to restrict our study subcorpus to the macro-classes anger, positive and neutral (as explained in 4.1). We present the Kappa values for these three annotation macro-classes. In the experiments described in this paper, we only keep consensual instances.

Table I. *Annotation agreement scores for the two corpora*

Corpus	# Coders	Kappa values	Macro-classes
NAO-HR	2	0.4	positive, anger, fear, sadness, neutral
NAO-HR	2	0.7	positive, anger, neutral
IDV-HR	2	0.7	positive, anger, fear, sadness, neutral
IDV-HR	2	0.8	positive, anger, neutral

4. Cross-corpus Experiments

4.1. Performance comparison for several Human-Robot Interaction audio corpora

All the emotion detection features in this study are based on the OpenEar Interspeech 2009 Challenge [10]. As our final emotion detection system is going to be an embedded system, we would like first to reduce the number of acoustic features for the specific task studied in this paper. We have chosen the AIBO corpus to be able to compare the results obtained on IDV-HR and NAO-HR with a well-known reference. We have separated the two schools of the AIBO corpus in order to have the same recording conditions. AIBO corpus contains Anger, Motherese, Empathy and Neutral. In order to compare AIBO, IDV-HR and NAO-HR together, we are going to use the macro-classes: anger, positive and neutral. Table II summarises the number of instances used in all the experiments. As our sets of instances are nearly balanced, we will report only the Unweighted Average Recall percentage (average percentage of correctly detected instances per class).

Table II. *The number of instances used in the experiments*

#instances	Neutral	Positive	Anger
IDV-HR	282	267	282
NAO-HR	51	51	27
AIBO (Mont)	372	372	324
AIBO (Ohm)	501	501	447

Our first experiment was to reduce the number of acoustic features for the specific task: neutral, positive and anger recognition. Table III shows the cross-validation (10 folds) performances for the four corpora. The confidence ratio is estimated on the basis of 384 features without any normalisation. Optimisation of parameters and classification are done with libSVM tool [11]. We have tested three different features sets: the first one contains the 384 features of the Interspeech 2009 challenge, the second contains only the 192 basic features (no derivatives) and the third one contains the basic features after having removed energy features (180 features). We have noticed in previous studies that energy features are important for activation detection, but removing them does not make the results fall down on our specific task. We have also tested the performance with normalisation to speaker called NS (z-norm: for each acoustic features, median and standard deviation are computed for one speaker only) compared to no normalisation (NO). As we can see in Table III, performances are quite the same for the three sets of features. As we would like to have the smallest number of features in our final ROMEO application, we will use only basic features (without derivatives) without energy features.

Table III. *Cross-validation performances (UAR, %) for different sets of features and normalisation (NO: no normalisation, NS: normalisation to speaker)*

#features	384	192	180	180	Confi dence
Normalisation	NO	NO	NO	NS	
AIBO-Mont	62.83	61.22	61.40	59.84	2.90
AIBO-Ohm	37.18	37.47	37.19	63.20	2.49
NAO-HR	52.65	61.65	52.29	57.44	8.62
IDV-HR	40.71	39.09	39.61	41.95	3.34

As our NAO-HR corpus is a small corpus, the results are probably biased: high number of features for small number of instances. Normalisation to speaker (NS) allows us to improve significantly the performances on AIBO-Ohm on cross-validation test. Normalisation to speaker also improves performances on NAO-HR but the gain remains in the confidence interval.

4.2. Experiments on cross corpora emotion detection

The final emotion detection system will be able to recognise emotions expressed by adults and children. In order to improve this detection, we would like to know if we can mix both NAO-HR and IDV-HR corpora, or if we need to have two different training sets. The following experiments are made with the 180 features set and libSVM tool for optimisation of parameters and classification. All features have been normalized using normalisation to speaker. We have made different cross-corpus tests between NAO-HR, IDV-HR and AIBO-Mont and AIBO-Ohm. The results reported in Table IV are the UAR performances (and confidence interval) of cross-corpus tests. For example: training on AIBO-Mont, testing on AIBO-Ohm performs 50.20% UAR. A first result is that cross-corpus between AIBO-Mont and AIBO-Ohm is high performing: both corpora are in German, they have the same task, annotation protocol is similar and speakers belong to the same age group. Training on AIBO (both schools) and testing on NAO-HR is better than testing on IDV-HR. NAO-HR and IDV-HR have nearly the same annotation protocol, are both in French, but speakers do not belong to the same age group. We can notice similar trends when training on NAO-HR: tests on AIBO (34.68% and 38.75%) have better performances than tests on IDV-HR (29.02%). Generally speaking, every cross-corpus experiments with IDV-HR are below the random guess (33%). Our conclusion is that we can not mix together IDV-HR and NAO-HR, but it seems feasible to mix together AIBO and NAO-HR.

Table IV. *Performances in cross-corpus classification, column for test, line for train*

	AIBO-Mont	AIBO-Ohm	NAO-HR	IDV-HR
AIBO-Mont		50.20 (2.37)	40.25 (8.47)	33.24 (3.20)
AIBO-Ohm	53.18 (2.99)		47.06 (8.61)	31.29 (3.15)
NAO-HR	38.75 (2.92)	34.68 (2.45)		29.02 (3.08)
IDV-HR	30.30 (2.77)	30.29 (2.36)	25.05 (7.48)	

As we see in section 4.1, IDV-HR presents relatively poor performances on cross-validation. Our hypothesis is that this corpus corresponds to a very specific HRI situation, with visually impaired people whose age varies from 28 to 80; expressed emotions are quite shaded in IDV-HR, contrary to NAO-HR. Therefore, we will probably need a specific model for this particular corpus. Some instances in NAO-HR and IDV-HR are very short (less than 1s). On such durations, the pitch, spectrum and voiced part estimation is not absolutely reliable. It can introduce bias in emotion detection and explain the differences of UAR performances between AIBO (minimum duration is 1s) and NAO-HR and IDV-HR.

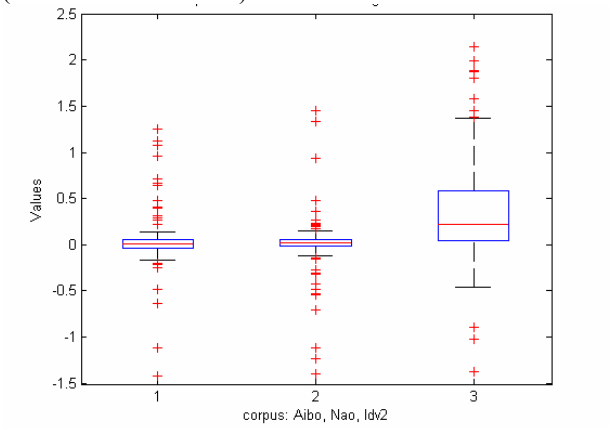

Figure 2. *Distribution of distances between anger instances and all the instances*

Looking at the confusion matrices, we can notice that the anger class is the one which is the worst recognised. This is probably due to the fact that the tasks are different from one corpus to another, and then, anger has a different meaning across corpora. One method to show this point is to compute the distance between anger instances and all the instances [12]. Figure 2 shows the distribution of the distances. We can see that anger has quite the same repartition for AIBO and NAO-HR corpora and is very different from IDV-HR anger. This could explain the relatively poor performances in cross-corpus combining IDV-HR with AIBO or NAO-HR.

5. Conclusion

We focus in this paper on an anger-positive-neutral detection module of an audio HRI emotion detection system, based on a multi-level processing of the audio cues. Due to the real-life condition of the test, we carried out different experiments on emotion detection on cross-corpora.

Our first experiments show that derivatives and energy may be removed of the feature set for our specific anger-positive-neutral detection task. As we have a smaller set of features, the real-time system will probably be faster. The second important point is that normalisation to speaker seems to improve cross-validation performances, which confirms previous studies [5]. Our last result is that we see that it is possible to mix two audio realistic corpora recorded with children in HRI, but it seems more complex to mix a corpus with adult speakers and a corpus with children speakers. Therefore, we will suppose that two different models would lead to better performances.

Our results on real-life corpora must be validated on larger amount of data. In order to build the final embodied models, we will study the influence of instances duration, other normalisations and last but not least, the real-time audio segmentation. Further studies will need to be carried out to develop the other modules which will supply our HRI system with data: emotion detection, activation detection, speaking rate, affect bursts detection, etc. We will notably study the influence of instances duration on emotion detection.

6. Acknowledgements

This work is financed by national funds FUI6 under the French ROMEO project labeled by CAP DIGITAL competitive centre (Paris Region).

7. References

[1] Miwa, H. Itoh, K. Ito, D. Takanobu, H. Takanishi, A., "Introduction of the Need Model for Humanoid Robots to Generate Active Behavior". In proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, vol.2, pp.1400-1406, 2003.

[2] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A. and McOwan, P., "Affect recognition for interactive companions: challenges and design in real world scenarios". Multimodal User Interfaces, vol.3, pp.89-98, 2010.

[3] Breazeal, C., "Robot in Society: Friend or Appliance?". In proc. of Autonomous Agents Workshop on Emotion-Based Agent Architectures, Seattle, WA, 1999.

[4] Batliner, A.; Hacker, C.; Steidl, S.; N¨oth, 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 proc. of the 4th International Conference of Language Resources and Evaluation, pp. 171–174, 2004.

[5] Schuller, B.; Vlasenko, B.; Eyben, F.; Wöllmer, M.; Stühlsatz, A.; Wendemuth, A. & Rigoll, G., "Cross-corpus acoustic emotion recognition: variances and strategies". In proc. of IEEE Transaction on Effective Computing, vol. 1, n°1, 2010.

[6] Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S., "Cross-corpus classification of realistic emotions – some pilot experiments". In proc. of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, pp. 77-82, 2010.

[7] Delaborde, A., Devillers, L., "Use of Nonverbal Speech Cues in Social Interaction between Human and Robot: Emotional and Interactional markers". In proc. of the Third International Workshop on Affective Interaction in Natural Environments, ACM Multimedia, Firenze, Italy, 2010.

[8] Devillers, L., Vidrascu, L. and Lamel, L., "Challenges in real-life emotion annotation and machine learning based detection". Journal of Neural Networks, vol. 18 pp. 407-422, 2005.

[9] Devillers, L. and Martin, J.-C., "Coding Emotional Events in Audiovisual Corpora". In proc. of the International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, 2008.

[10] Schuller, S., Steidl, S., Batliner, A. "The INTERSPEECH 2009 emotion challenge". In proc. of the 10th Annual Conference of the International Speech Communication Association (Interspeech), Brighton, U.K., 2009.

[11] Chang, C.-C. and Lin, C.-J., "LIBSVM : a library for support vector machines". ACM Transactions on Intelligent Systems and Technology, vol. 2, n°3, pp. 27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] Tahon, M., Devillers, L., "Acoustic measures characterizing anger across corpora collected in artificial or natural context". In proc. of the 5th International Conference Speech Prosody, Chicago, 2010.