


Inference of Human Beings' Emotional States from Speech in Human–Robot Interactions

Laurence Devillers^{1,2}  · Marie Tahon¹ · Mohamed A. Sehili¹ · Agnes Delaborde¹

Accepted: 24 March 2015

© Springer Science+Business Media Dordrecht 2015

Abstract The challenge of this study is twofold: recognizing emotions from audio signals in *naturalistic Human–Robot Interaction* (HRI) environment, and using a *cross-dataset recognition* for robustness evaluation. The originality of this work lies in the use of six emotional models in parallel, generated using two training corpora and three acoustic feature sets. The models are obtained from two databases collected in different tasks, and a third independent real-life HRI corpus (collected within the ROMEO project—<http://www.projetromeo.com/>) is used for test. As primary results, for the task of four-emotion recognition, and by combining the probabilistic outputs of six different systems in a very simplistic way, we obtained better results compared to the best baseline system. Moreover, to investigate the potential of fusing many systems' outputs using a “perfect” fusion method, we calculate the oracle performance (oracle considers a correct prediction if at least one of the systems outputs a correct prediction). The obtained oracle score is 73 % while the auto-coherence score on the same corpus (i.e. performance obtained by using the same data for training and for testing) is about 57 %. We experiment a reliability estimation protocol that makes use of outputs from many systems. Such reliability measurement of an emotion recognition sys-

tem's decision could help to construct a relevant emotional and interactional user profile which could be used to drive the expressive behavior of the robot.

Keywords Human–robot interaction · Emotion recognition · Prediction reliability · Real-life data

1 Introduction

Social robots should be endowed with three major emotional and communicational skills: the ability to efficiently decode the verbal and non-verbal expressions of the user, the ability to interpret and react by taking this information into account, and the ability to properly encode its own emotional responses. The success of a social robot is highly dependent on the performances of these three competences.

This paper focuses on the decoding of emotional behavior of human subjects interacting with a robot using audio input, as well as on the reliability of emotion prediction. The present work is part of the Human-Robot Interaction (HRI) project ROMEO2, which notably aims at providing a system that has the ability to socially interact with elderly users, by building a dynamic representation of their emotional and interactional tendencies, and selecting the most adapted and relevant social behavior as a response.

Accurate emotion recognition is important in assistive social robotics. For instance, a proper understanding of situations where an elderly person gets irritated or depressed will help to trigger the most suitable behavioral response from the robot. Nevertheless, emotion recognition is currently still a challenge, and its reliability needs to be taken into account to provide trustworthy human-robot interaction systems.

✉ Laurence Devillers
devil@limsi.fr

Marie Tahon
mtahon@limsi.fr

Mohamed A. Sehili
sehili@limsi.fr

Agnes Delaborde
agnes.delaborde@limsi.fr

¹ LIMSI-CNRS, Orsay, France

² Université Paris-Sorbonne IV, Paris, France

Over many decades, many works about the detection of the affective behavior in human voice have been realized in psychology [34], phonetics and linguistics. This issue has also been addressed for about fifteen years in computer modeling [16,23]. There is an increasing number of studies in this field, though to this day there are no applied standards for data annotation, nor for audio features computation. The most prevailing efforts concern the evaluation of different recognition systems. This has become possible thanks to challenges carried out using corpora (standard “acted” and realistic databases) made available for researchers in order to compare the performance of their systems within the Interspeech conference from 2009 to 2013 [35,38–40].

The corpora used in these experiments are mainly acted and the results are not very useful to real-life application. Most current approaches are based on automatic machine learning techniques to acquire models for prototypical emotions (based on expressions annotated on a representative corpus) in order to detect similar patterns in new test expressions. These techniques are fairly good when using the same material for training and testing but are less accurate on realistic and new emotional expressions.

Speech signal contains numerous features about the speaker: their sociological and physical (age, health) and emotional states. Acted data vary a lot across speakers and contexts, but realistic data vary much more. Actually, from one person to another, voice characteristics are very different. They also vary for the same person from one situation to another according to the context. Thus, there is a mounting evidence that there is no small set for basic and prototypical expressive emotions in a real-life context but rather an huge number of complex and mixed emotions. If there is no ground truth in the comprehension of emotional expressions, we can imagine that there are several ways to model emotions in different contexts depending on different corpora and different sets of acoustic features. The experiments presented in this paper consist in drawing a parallel between several models that have learned emotional expressions from different realistic corpora and acoustic features sets.

Our research mainly addresses the processing and interpretation of non-verbal audio data, in order to build, over the long run, a high-level interpretation of the user’s emotional and interactional tendencies. These emotional expressions, automatically detected by a robot during interactions, allow it to dynamically interpret the profile of the user in order to update the expressive answering strategies and actions according to the user’s behavior [7,14,46]. In order to fulfil this goal, it is essential to collect and analyze the emotional behaviors of potential end-users of these technologies in realistic interaction contexts [16,18].

As we present in the next section of this document, the reliability of HRI systems is one of the current issues in the community, but it is often assessed on the performance of

the whole system, rather than on each separate module composing the system. Based on this observation, we propose a way to measure the reliability of affective behavior detection using a set of models obtained from multi-corpus and multi-feature sets approaches. In a realistic application, the user would tend to trust the robot if its reactions are reliable, even if some of the user’s emotions are not taken into account by the robot. Therefore, false positive emotion detections must be minimized. The use of measures of reliability (e.g. prediction from the context, measuring the confidence of models’ outputs) are very important so that the user continue to trust the robot.

The rest of this article is organized as follows. Section 2 presents some of the related work. Section 3 describes the three realistic corpora used in our experiments. Two of them were collected with potential end-users (elderly and disabled people aging from 16 to 90 years old) in human-robot interaction, in order to test the generalization power of our models. In Sect. 4, we give a detailed description of the different sets of acoustic descriptors upon which our study relies. Section 5 reports our experimental results for auto-coherence, cross-validation and cross-corpus evaluations. It also presents a potential method to combine the predicted outputs so as to improve the overall accuracy of the system and to measure the reliability of its decisions. A discussion of the obtained results and future perspectives will be presented in Sect. 6.

2 Related Works

2.1 Reliability of Emotion Detection Systems

A decade ago, emotion recognition models were trained using acted artificial data, such as the Danish Emotional Speech (DES) [21] or the Berlin Emotional Speech-Database (EMO-DB) [8]. These data were generally collected with few actors using the same lexical content. Nowadays, the trend is to train models on realistic data since realistic emotions could not be found in acted databases [4]. The two main drawbacks of the standard corpora used in the community are the very small size of audio content and the data variability in terms of task, speaker, age and audio recording environment which compromise the significance of results and improvements [37]. As a consequence, there is a critical need for data collection with end-users (with different types of speakers, ages, etc.) and real tasks for emotion recognition systems.

Realistic data are more appropriate to build an affective state recognition system for robots evolving in realistic environments. However, in real-life contexts, neutral speech is predominant and emotions are quite sparse, shaded, complex and often mixed. Hence, there is a critical need to build

context-independent macro-classes and mixtures of emotions [3, 16]. Some prototypical databases are also collected in the community; they contain consensual emotions defined as “utterances that are consistently recognized by a set of human evaluators” [30]. Because they contain prototypical emotions, their collection with large panels of speakers and different acoustic environments is simplified. That is why prototypical and realistic corpora are useful to investigate cross-corpus variabilities.

In order to implement reliable emotional models in systems such as robots, new data must be collected during interactions with these systems. So far, very few HRI databases have been collected. Among them, the AIBO database [2] was collected during a child-robot interaction with the Sony AIBO dog. The SEMAINE database [29] was recorded during emotionally rich interaction with an automatic agent. The Herme database [24] was collected in the Science Gallery of Dublin, during conversations between visitors and a robot. Emotional databases are mainly audio databases, but recent databases are collected with different modalities such as video, tactile and physiological inputs. The collection of rich affective data aims at building automatic recognition systems applied to HRI. For example, the system developed by [1] uses voice and facial modalities to recognize acted emotions.

Experiments on a single corpus do not allow for an estimation of the reliability of emotion detection systems in real-life conditions. In fact, as one can learn from the literature, many studies are carried out on one single corpus and results cannot be generalized to other corpora. Cross-corpus experiments consist in using one or many corpora as training data and another *different* set of corpora as test data. Following this kind of evaluation protocols, recognition rates are low but more realistic [41]. Two protocols are predominant in cross-corpus experiments (with N corpora). The leave-one-corpus-out protocol (the “unit” protocol) consists in training on $N - 1$ corpus and testing on the remaining one [22, 36]. The model can be trained with few corpora and improved with an unsupervised adaptation [49]. It allows to merge different data: fixed or variable linguistic information, realistic or acted data, children or adults. In Marchi et al. [28], the authors conclude that the accumulation of similar speech data in the training set improves the recognition performances. The other protocol (the “vote” protocol) consists in training using one corpus and testing on the $N - 1$ remaining corpora. Then, the recognition rate is the majority vote of each test [42, 45, 46]. For binary valence recognition, the unweighted average recall (UA) rates on seven corpora—acted (DES, EMO-DB), induced (eNTERFACE) and realistic (VAM, SAL and SUSAS)—are from 50 to 55 %. Such experiments are very interesting, but the results are slightly over the random guess. The cross-corpus protocol used in this article is based on a fusion of the probabilistic outputs of many

systems. Comparing our results with other cross-corpus works would indeed be relevant, but our work differs in that we work on four different classes, which are not the same in the works cited in this section.

Cross-corpus experiments give the experimenters an estimation of how accurate is the emotion detection system in real contexts. This should be the first element to take into account in order to improve the credibility of the robots. The next element consists in integrating reliability measures in the systems. Such measures could help the robots at the decision-taking level which may be conducted very restrictively because of the importance of potential consequences and with respect to the current task of the robot. As far as we know, including reliability measures has been studied from a theoretical point of view [11], and not yet from an engineering point of view.

2.2 Reliability of the HRI System

The emotional and interactional behaviors and tendencies of the user can be represented at different levels [7, 14]: from a low-level representation (extraction of low-level cues, such as acoustic cues, and a subsequent interpretation in terms of “voiced” or “unvoiced” signal), to a high-level interpretation (personality or intentions inferred from acoustic cues for example), crossing a mid-level interpretation (emotional state predicted from acoustic cues).

For example, research teams from Stanford University and MIT [26] carried out an exploratory process on the cooperation between a human and a robot through the expression of robotic backchannels. The system interprets audio and video data in order to localize the user and understand voice commands (speech recognition). From this interpretation, the robot is able to provide backchannels at appropriate time points. Other studies have been carried out by Castellano et al. [9], in a gaming and educative context between children and the iCat robot. In this context, only the state of the game is interpreted so as to infer the child’s emotional state, and then select a matching robotic behavior. Ochs et al. [31] present an empathic rational dialog agent which deduces the type and the intensity of the emotions potentially felt by the user, from an understanding of the communicative acts expressed. High-level and even bio-inspired schemes currently constitute one of the main challenges in HRI. Emotirob [19] deals with a multimodal detection (speech, prosody, vision) which impacts the internal cognitive state of the robot, and its behavior selection. Buendia and Devillers [7] shows an interest in extracting and interpreting several social cues (identity, speech, emotions) so as to maintain a long-term relationship with a robot able to deal with high level social attitudes such as understanding a lie or showing compassion.

The level of interpretation and the dimensions processed are highly dependent on the interactional applicative

end-context. Nonetheless, the higher the complexity of the interpretation, the higher is the risk of misinterpreting the social cues expressed by the user. The issue of reliability in HRI has often been addressed in the community, especially at the interaction level. For example, the reliability of a non-social robot's automation affects the user's trust and their desire to let the system be autonomous [13, 15]. Hegel et al. [25] investigate the notion of reliability of the social signals transmitted by the robot, through its physical appearance and the related expectancies it can raise in the user. If the cues sent by the robot match the expectancies of the user, its reliability will increase in the eyes of the user. This use of "reliability" refers mainly to the overall trust of the user towards the system. Yagoda and Gillian [48] consider the reliability as a parameter in their proposed HRI trust scale.

An analysis of reliability is a keystone for assessing the performance of a HRI system. Splitting a system into minimal units, and affecting qualitative and objective efficiency measures to compute the best cost to fulfill the task has been proposed in the PARADISE theoretical evaluation framework [47]. Similarly, the evaluation can rely upon a benchmark of metrics used to assess low level units of behavior such as head position or gestures [12], or tasks divided into subtasks [44] so as to provide a comparative analysis of interactional behaviors on the whole.

The success of a social and/or assistive robot is highly dependent on the system's performances (efficiency of the models used), but also on the correctness of subjective information used as input (e.g. emotion, personality). Hence, there is a pressing need to take into account the reliability of each module of the system in order to adapt the final social response of the robot. Although this aspect is deemed primordial by the community, the evaluation of each step of the automated process in social HRI is complex and rarely fully implemented. As an illustration, reinforcement learning techniques provide a trail for the evaluation of the success of the robot's social action: for example, Kreier et al. [27] propose a platform for a robotic bartender which selects the best action to perform according to a set of rewards/penalties triggered by the evaluation of its performance of the expected task. However, the reliability of the detected inputs is not taken into account. As of now, no social and assistive robotic systems are guided along a track of reliability evaluation which weighs in the final decision for the robot's behavior, from the early signal detection, the interpretation in terms of user model or context representation, to the final selection of behavior.

3 Databases

The experiments were carried out using three different corpora collected in realistic conditions within many national

research projects.¹ The three used corpora were segmented and doubly annotated according to the protocol and annotation scheme described in [17]. Cohen's kappa (κ) score (or Cronbach's α for continuous data) is generally used to measure the concordance degree between the annotations of two or more human annotators. It can also be used to measure the intra-annotator agreement. Consensual data is then used to train models. The quality and reliability of the annotations is a vital issue. It is checked thanks to these inter-annotator concordance tests and to perceptual test carried out with other observers, generally on a randomly selected part of the corpus. The effort on the annotation is often underestimated, it is nevertheless essential. It is actually frequent to see fairly low kappa scores (<0.5) when it comes to emotion annotation on realistic data (not played by actors). These scores show the difficulty of the annotation task and the diversity of the present emotions. In everyday life, emotions are rarely discrete and primary, they are often mixed up, subtle and follow a dynamic process.

Nonetheless in this present work, restraining the number of classes is compulsory for classification. We thus merged the emotional states into their respective macro-class Anger, Joy, Sadness and Neutral. We do not work on Ekman's two other primary emotional states [20], disgust and surprise, since our databases do not present enough samples for these. The following list presents the merging we performed:

- Anger: nervousness, irritation, impatience, cold anger and fury;
- Joy: positive annotations of amusement, relief, satisfaction as well as frank joy with laughter;
- Sadness: different types of sadness expressed through slow, hesitating voices with less energy, or even depressive voices;
- Neutral: non-expressive state.

The Table 1 sums up the corpora's distributions in terms of subjects, durations, segment numbers and emotion tags.

3.1 JEMO

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 [6]. The lexical content was totally unconstrained, and the speaker tried and modulate freely his or her emotional expressions so as to be recognized by the system. As a result, the participant

¹ Many efforts are made to release at least one of these corpora to the community, which will require specific data formatting and obtaining the agreement of all the participants.

Table 1 Content description for each data corpus

Corpus	Subjects(M/W) Age	Dur. (min)	# Seg.	Anger	Joy	Sad.	Neut.
JEMO	59 (30/29) 16–48 y. o.	41	1249	291	310	307	341
ARMEN	77 (48/29) 18–90 y. o.	70	1807	403	506	318	580
IDV-HR	22 (11/11) 28–80 y. o.	82	2063	512	508	495	551

produced very expressive emotions in order to be as close as possible to the entries expected by the system. The corpus was recorded in December 2010 in the LIMSI laboratory. The total duration of the corpus is 41 min and it includes 59 participants (30 men and 29 women).

3.2 ARMEN

In order to obtain spontaneous and as close as possible to real-life data, a data collection system which simulates a natural interaction was developed. It implements an expressive virtual agent embedded on a robot. It was used within the ANR Tescan ARMEN project [10] to collect 2 emotional corpora with the participation of about 80 patients from medical centers (elderly and impaired people) in the region of Montpellier (France) in 2010 and 2011. Two people conducted the data collection process: an interviewer and an operator, who did not talk to the subjects but only set up the recording environment and operated the virtual character (avatar) without the knowledge of the subject. The experiment would take place as follows: first the interviewer would greet the subject and explain him/her the purpose of the experiment (collecting data for a future assistive robot) while the operator was attaching the lapel microphone onto the subject's clothes, adjusting the camera's height, checking that the audio and video were working, and starting the recordings. Then the subject would interact with the virtual character in the frame of different scenarios, detailed below (each scenario being explained beforehand by the interviewer). At the end, the interviewer would ask the subject to rate the quality of the interaction and the virtual character on a list of adjectives, using a 5 levels Likert scale to indicate if he/she agreed or not with the proposed adjectives. The scenarios were outlined by physicians and functional therapists from the reeducation center, written by the authors and validated by the physicians. They are designed to satisfy several constraints: matching the test cases for the functionalities of the robot and being close to the final user experience, eliciting emotions to collect useful training data, being easy to relate to for the subjects, offering variability but staying in a limited dialog to ensure robustness. There were four scenarios. In the first one "introduction" scenario), designed to put the subjects at ease and

build a beginning of proximity, the virtual character and the subject would introduce themselves. Then in a short training phase the VC would ask the subject to try to pronounce the phrase "My voice conveys emotions" with different emotional intonations (anger, happiness, sadness). Then pictures of staff and others patients of the center known by the subject, taken at various positive occasions (birthdays, games, activities...), were presented and the virtual character would ask the subject to name the people, if they were friends, explain the situations and if they had good memories of it. 77 persons, of which 48 men and 29 women between 18 and 90 years old participated in this data collection (about half of the participants were over 60 years old). The consensual data constitutes about 70 min of the corpus. The collected data was used to explore approaches which aim at resolving the performance generalization problem of emotion detection systems run on different data [10].

3.3 IDV-HR

The French IDV-HR corpus [46] (HR for Human–Robot) was recorded in a model apartment for disabled people, in the *Institut de la Vision*² in October 2010 during the ROMEO project. 22 subjects (11 men and 11 women) took part in this data collection for a total duration of 4 h, 7 min and 43 s after segmentation. 82 min of consensual data are used in our experiments (a large amount of the original corpus is labeled as neutral).

During the recording, Nao asks the subject to picture himself or herself in five different scenarios which could possibly occur when waking up in the morning: (1) feeling well, (2) a bit sick, (3) health emergency, (4) in a bad mood, (5) exhilarated by something pleasant to come. During the series of scenarios, the robot presents one distinctive social behavior (among directive, doubtful, encouraging, kind, machine-like, empathic); its utterances are coded according to a French grammatical representation allowing to finely represent its current social behavior. The subject engages in a discussion with the robot according to the latter's questions and suggestions. The robot's dialogue is controlled by a Wizard-

² Vision Institute, 11 rue Moreau, 75012 Paris.

Table 2 IDV-HR extract: Nao (in doubtful mode) talks to the subject (a 47 year-old female) during the “something pleasant to come” scenario

Nao	How are you this morning?
Subject	<i>I am really fine, I think I'll go running a bit, do some exercise. Then I think I'll go shopping. That's great, this day is going to be awesome</i>
Nao	It seems like you're feeling OK, but I'm not so sure. Maybe you know what you plan to do today?
Subject	<i>Yes. I'll go running. In the park. Then I'll go shopping</i>
Nao	Yes, that may be nice. And perhaps you plan to go somewhere else afterwards?
Subject	<i>Afterwards I'll go to the theater with my lover</i>
Nao	I think it may be nice, but I'm not really sure. Is it nice?
Subject	<i>But- sure it is! Yes!</i>

of-Oz experimenter, who selects the most appropriate answer among a finite set, every answer being linguistically coded according to the current social behavior. We present a translated extract of the corpus in Table 2.

The subject plays two or three times the set of five scenarios, and for each series the Wizard-of-Oz selects a new robotic social behavior. Subjects are asked about their perception of the robot's social and communicative skills through self-report questionnaires.

We decide to use this real-life database as the reference test in our cross-corpus experiment.

4 Audio Features Sets

Acoustic descriptors used for emotion detection [34] are borrowed from the fields of phonetics, speech recognition and music recognition and have been used to measure many phonation and articulation aspects. These descriptors include acoustic parameters in the frequency domain (e.g. fundamental frequency F_0), in the amplitude domain (e.g. energy), in the time domain (e.g. rhythm) and in the spectral domain (e.g. spectral envelop or energy per spectral bands). Many studies [34,35] show the interest of combining many of these parameters.

In this section, we present three sets of acoustic descriptors obtained using different libraries which include parameters from different domains. They are used to carry out experiments on three corpora (Sect. 5). Acoustic descriptors are in great majority calculated thanks to the open-source libraries Open-Smile, Yaafe, Aubio and Praat. Parameters known within the community as important for emotion recognition are used in these sets. There is an overlap between the three sets for certain parameters, however, they each have their own characteristics. The three tables below (4–6) show the content

Table 3 Repartition of the three sets in the eight big families of descriptors (EBBark: Energy per Bark Band, VQ: Voice Quality)

Feature family	Set 1 (384)	Set 2 (334)	Set 3 (293)
EBBark	0	240	136
Cepstrum	288	0	78
Spectrum	0	60	54
F_0	48	15	7
ZCR	24	10	0
Energy	24	0	6
VQ	0	9	4
Formants	0	0	8

of each set, classified by family. We have selected eight big families of descriptors including prosody (fundamental frequency, energy and voice quality), spectrum (descriptors of envelope energy per spectral bands), cepstrum (MFCC, especially used in speech recognition), zero-crossing rate (ZCR) and formants. A mean and variance normalization to corpus has been performed on all acoustic features. A comparative summary of the acoustic families is given in Table 3.

4.1 Set 1 (S1-384-IS09)

This set of 384 parameters (Table 4) is widely used within the international community and allows performance comparison. The 384 features are extracted using the Open-Smile tool and were used for the 2009's Interspeech challenge on emotion detection [38]. It contains many low-level descriptors (LLD) such as the fundamental frequency, the voicing likelihood, ZCR, RMS energy (raw energy per perceptual filter). This set also includes the derivatives (Δ) of all low-level descriptors. Furthermore, for all low-level descriptors, 12 statistical functions (functionals) are applied: maximum, minimum, range (maximum–minimum), maxima and minima positions, mean, first and second coefficients of a linear regression (one of them corresponds to the slope mentioned above), the correlation coefficient Q , standard deviation (std), kurtosis (estimation of the flattening of a probability distribution), skewness (estimation of the asymmetry of a probability distribution around its mean value).

4.2 Set 2 (S2-334-LIMSI)

This set of 334 parameters uses two open-source libraries (Yaafe and Aubio) and is mainly based on perceptual energy (or loudness, 204 parameters out of 334) extracted on 24 Bark bands. It also contains many voice-quality descriptors based on the calculation of voice jitter and shimmer (Table 5). 10 statistical functions are applied to the acoustic features: maximum, minimum, mean, median, standard deviation, slope, centroid, spread, kurtosis and skewness.

Table 4 First set of acoustic features (S1-384-IS09)

LLD	Functionals	Voiced
ZCR	12 func.	12
Δ ZCR	12 func.	12
ZCR	–	24
Voice quality		0
Voicing likelihood	12 func.	12
Δ Voicing likelihood	12 func.	12
F_0 (Hz)	12 func.	12
ΔF_0 (Hz)	12 func.	12
Pitch		48
RMS energy	12 func.	12
Δ RMS energy	12 func.	12
Energy		24
Spectrum		0
Energy per spectral bands		0
MFCC 1-12	12 func.	144
Δ MFCC 1-12	12 func.	144
Cepstrum		288
Formants		0

12 functionals: min, max, range, max and min positions, mean, 1st and 2nd linear regression coefficients, correlation (Q), std, kurtosis, skewness

4.3 Set 3 (S3-293-LIMSI)

We propose a set of hybrid (perceptual and cepstral) coefficients of 293 acoustic descriptors (Table 6) obtained with praat and also some scripts on matlab. This set includes a certain number of descriptors from the scientific community, be they in the speech and emotions field or in the field of music. A few descriptors are redundant, which allows for a comparative study of the most robust among them. The descriptors are generally a combination of low-level descriptors and standard statistical functions applied to the LLD. Given that the descriptors have different behavior according to the type of signal, we propose to extract them by three, easy to identify, signals: voiced signals, non-voiced signals and the whole audio signal. The descriptors are extracted for each sub-part of the signal and then the average value is calculated for all sub-parts of the same type. For instance, the mean of minimum values of the fundamental frequency obtained for each voiced region of the signal. In order to reduce F_0 estimation errors, only voiced regions of above 4 ms are taken into account. For each voiced region, the mean, standard deviation, minimum and maximum of the F_0 are calculated. To which we add the glissando $G = \frac{F_{max} - F_{min}}{T_{max} - T_{min}}$ (for a frequency in semitones).

Table 5 Second set of acoustic features(S2-334-LIMSI)

LLD	Functionals	Voiced	All
ZCR	10 func.		10
Jitter (absolute)	Mean, std	2	
Jitter (relative)	Mean, std	2	
Shimmer	Mean, std	2	
Shimmer (dB)	Mean, std	2	
Punvoiced			1
Voice quality		8	1
F_0	Mean, max, min, median, slope	12	
Voiced frames			1
Number of voice break			1
Degree of voice break			1
Pitch		12	3
Energy		0	0
Roll off 95 %	10 func.		10
Spectral decrease	10 func.		10
Spectral variation	10 func.		10
Perceptual spread	10 func.		10
Perceptual sharpness	10 func.		10
Spectral flatness	10 func.		10
Spectrum		0	60
pecific loudness 0-21 (dB)	10 func.		210
Energy per spectral bands		0	240

10 functionals: min, max, mean, median, std, slope, centroid, spread, kurtosis, skewness

Table 6 Third set of acoustic features (S3-293-LIMSI)

LLD	Functionals	Voiced	Unvoiced	All
ZCR		0	0	0
Local Jitter (Praat)				1
Local Shimmer (Praat)				1
HNR (Praat)				1
Punvoiced (Praat)				1
Voice quality				4
F_0 (st)	Mean, std, max, min	4		
Glissando (st/s)	Moy	1		
Intra/inter F0 (semitones)		2		
Pitch		7		
Total loudness (Bark, dB)	Mean, std	2	2	2
Energy		2	2	2
Roll off 5 % (st)	Mean, std	2	2	2
Roll off 25 % (st)	Mean, std	2	2	2
Roll off 50 % (st)	Mean, std	2	2	2
Roll off 75 % (st)	Mean, std	2	2	2
Roll off 95 % (st)	Mean, std	2	2	2
Total slope (st/Hz)	Mean, std	2	2	2
Slope [0–500] (st/Hz)	Mean, std	2	2	2
Slope [500–1500] (st/Hz)	Mean, std	2	2	2
Barycenter	Mean, std	2	2	2
Spectrum		18	18	18
Bark bands 0–21 (dB)	Mean, std	42	42	42
Harmonic bands 0–5 (dB)	Mean, std	5		
Energy per spectral bands		62	42	42
MFCC 0-12	Mean, std	26	26	26
Cepstrum		52	52	52
F2-F1 (st)	Mean, std, max, min	4		
F3-F2 (st)	Mean, std, max, min	4		
Formants		8		

5 Experiments

For these experiments, we used linear Support Vector Machines (SVM) with data normalization (feature values between -1 and $+1$) and with standard configuration. We decided to use SVM because we obtained better results than with other models such as decision trees [16]. SVM models were trained so that they output a probabilistic decision instead of a positive/negative distance [32]. Probabilistic output are more appropriate for system reliability estimation. Results are given in terms of F-measure (global and specific to each class) and accuracy and are presented for each of corpus/acoustic set combination.

5.1 Auto-coherence and Cross-Validation Results

For these two protocols, the training and test are carried out on the same corpus. The auto-coherence method uses the com-

plete corpus for training and testing. Cross-validation gives normally lower performances because a system is evaluated on unseen data. We used a 3-fold cross-validation protocol for each set of acoustic descriptors. For each run, two folds are used for training data and one fold for test. An average score is then calculated for the three runs.

The goal of an auto-coherence evaluation is to measure the best performance that a system can achieve on seen data. The goal of the cross-validation is to measure the generalization power with new data collected in the same context (data from the same corpus but unseen during training).

Table 7 (first line for each corpus) shows the average auto-coherence results, using three sets of acoustic descriptors (separately) for each corpus. As we can see, even when training data is used for test, the higher performance is around 80 %, which shows the difficulty to model the emotions in realistic context. This is probably due to the high intra-corpus diversity that makes the separation between the four macro-

Table 7 Average corpus results for three sets of acoustic descriptors

Corpus	Experiment	F-measure				
		All	Anger	Joy	Neutral	Sadness
ARMEN	Auto-coh.	64.3 ± 4.3	68.4 ± 2.4	64.5 ± 3.9	68.6 ± 3.8	50.7 ± 8.3
	Coss-val.	49.2 ± 1.8	51.4 ± 2.4	48.9 ± 1.1	57.1 ± 1.7	32.5 ± 2.6
JEMO	Auto-coh.	79.5 ± 3.3	85.4 ± 1.6	76.6 ± 3.9	81.7 ± 2.8	74.3 ± 5.3
	Cross-val.	59.7 ± 0.9	67.2 ± 3.6	54 ± 0.8	66.1 ± 0.5	51.3 ± 1.3
IDV-HR	Auto-coh.	56.9 ± 4	54.1 ± 5	63.2 ± 4.2	57.9 ± 4.3	52 ± 2.9
	Cross-val.	40.5 ± 0.5	36.1 ± 1.4	49.5 ± 2.1	43.1 ± 0.8	32.5 ± 1.3

The first row of each corpus stands for auto-coherence, the second one for cross-validation

Table 8 Cross-corpus results using IDV-HR as a test corpus

Corpus	Acoustic set	F-measure					Accuracy
		All	Anger	Joy	Neutral	Sadness	
ARMEN	S1-384-IS09	34.3	27.4	44.8	45.8	17.8	36.9
	S2-334-LIMSI	34.5	28.7	43.5	43.8	20.7	36.3
	S3-293-LIMSI	34.4	33.0	41.6	44.6	16.7	36.7
JEMO	S1-384-IS09	33.7	32.8	37.8	36.7	26.9	33.7
	S2-334-LIMSI	32.4	30.7	34.2	39.6	24.3	32.4
	S3-293-LIMSI	33.2	42.6	30.9	35.9	22.6	33.3
Fusion of the outputs of 6 systems		36.8	37.0	42.9	46.1	21.1	38.7

classes very difficult. In fact, when using an SVM classifier, many vectors (generally more than 60 %) (i.e. sound files after acoustic features computation) are retained by the training algorithm as part of the model. When using the same data for test, any non-retained vectors could be wrongly classified due to this high intra-corpus diversity. We can also notice the difference between the individual macro-classes performances. Sadness seems to be the most difficult class to identify. This could be explained by the fact that sadness signals are characterized by a low level of energy, and because not all extracted acoustic features carry information. One solution to cope with this problem could be to use rhythmic information [33]. Finally, we mention that the IDV-HR corpus, a realistic and non-acted corpus, had the lowest auto-coherence performance. For this reason IDV-HR will be used as a test corpus in our cross-corpus evaluations.

The performances presented in Table 7 (second line for each corpus) represent an average cross-validation score for three sets of acoustic descriptors. As for the auto-coherence, we notice that the lower performances are obtained with the IDV-HR corpus and that the sadness macro-class is still the most difficult to recognize.

5.2 Cross-Corpus Results : Test on IDV-HR

5.2.1 Independent Systems

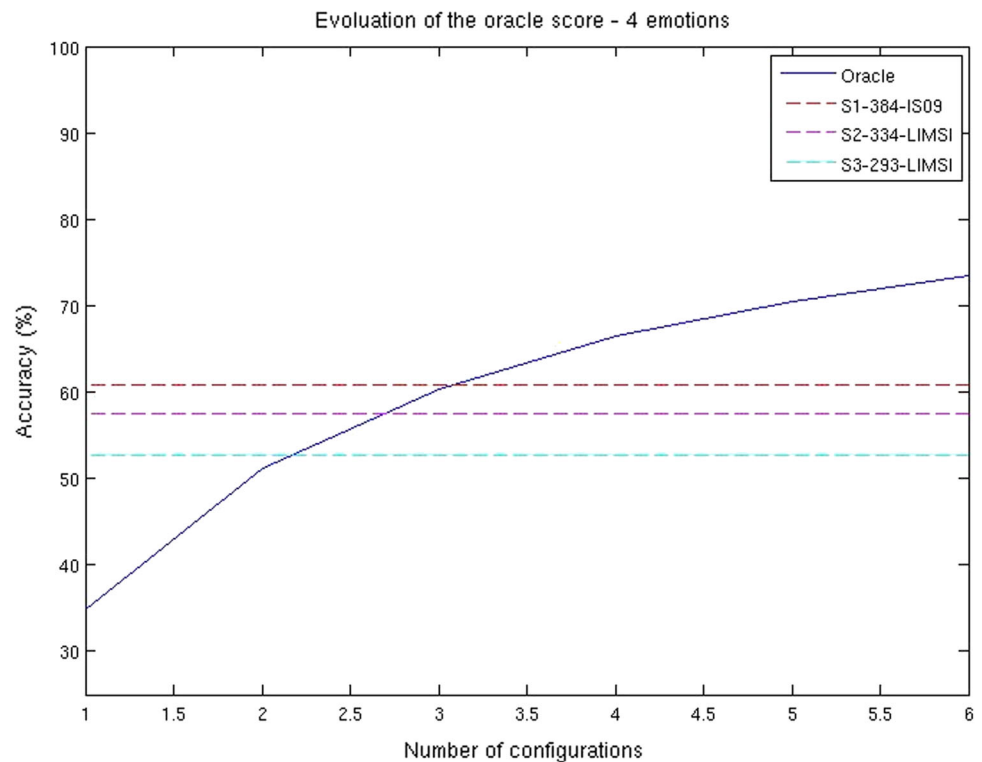
In this section, the results obtained while training the emotional model on JEMO or ARMEN and testing on IDV-HR,

are presented in Table 8 for each of the three corpora and each of the three acoustic sets.

As expected, the emotion recognition performances fall a lot between the auto-coherence and the cross-corpus experiment. It is not surprising to have low results on this complex real-life test set. However, surprisingly, the F-measure is similar in all cases (i.e with different training corpus and acoustic set) and still over the random guess. Sadness has a lower F-measure than other emotions, in auto-coherence as well as in cross-corpus experiment. The S2-334-LIMSI set, training on JEMO corpus seems to be the best configuration for sadness recognition (24.3 %). The S3-293-LIMSI set, training on JEMO seems to be the best configuration for anger recognition (42.6 %). Joy and neutral emotions are better recognized while training on JEMO corpus with the S1-384-IS09 set (44.8 % for joy and 45.8 % for neutral). These results underline a very interesting point: each configuration (training corpus and acoustic set) is best able to recognize a specific emotion. It suggests to combine all these results in order to benefit from this diversity.

To benefit from the diversity of the results obtained while merging training corpora and acoustic sets, the outputs of each system must be combined in a specific way. Both the predicted emotional label and the estimation probability for each class can be useful for this task. The experiment presented in this section consists in using the estimation probabilities obtained by the 6 systems on the four emotional classes. Of course, there exist many ways to combine these outputs. However, the one presented here gives interesting results.

Fig. 1 Evolution of the oracle score—the three *lines* represent the score obtained for IDV-HR in auto-coherence with the three sets



One way to combine the outputs is to multiply the 6 probabilities obtained for the 4 emotional states, thus resulting in four probabilities. The predicted emotion is the one corresponding to the maximum product. Results are reported in the last row of Table 8. Using such a fusion of the six systems the F-measure is improved by about 2.3 % (from 34.5 to 36.8 %).

Although there are many possible methods for system outputs fusion, it is not always easy to find a method which efficiently profits from many system outputs and significantly improves performances. Multiple system outputs can actually be viewed as a set of hypotheses. Among these hypotheses, there may be one or more that correspond to a correct prediction. A good fusion method would most of the time select a good hypothesis (or hypotheses) and discard the wrong ones. A “perfect” fusion method would *always* select a good hypothesis (hypotheses) if there is at least one (hence, it only fails if there is no good hypothesis at all). As this perfect method is unknown, we refer to it as oracle. It represents the theoretically highest performance that the fusion of many system outputs can achieve.

In our experiment, the oracle performance is compared to the auto-coherence performances obtained on the test corpus IDV-HR for each of the three acoustic sets (Fig. 1). The results show that using only one or two systems gives a lower accuracy than the auto-coherence test. However, using 6 systems in parallel gives an oracle accuracy of 73.4 %. Therefore, it would be possible to improve the accuracy on the test corpus

IDV-HR by using a more effective fusion method of the six systems outputs. Alongside to improving the performance of the individual systems, one important issue that emerges is finding a good fusion method.

5.3 Reliability Measurements

In order to improve the reliability of the decision-making system, a protocol to benefit from 6 systems in parallel has been presented. Different rules can be implemented. Among them, the maximum probability product was chosen in this study. A combination of two rules can improve the reliability of the system. The predicted emotion must satisfy:

- Rule 1: the probability product is the maximum,
- Rule 2: at least X systems must have predicted this emotion.

In the Fig. 2, the accuracies obtained when X varies from 1 to 6 are presented. These accuracy are compared to the oracle scores obtained with the second rule. The accuracy obtained with only rule 1 was 36.8 %. The closer accuracy found in Fig. 2 is the one obtained with $X = 3$ annotations. The combination of the two rules lowers the accuracy but improves the reliability. Using 3 annotations allows the system to be as close as possible to the oracle score without being under the random guess and without lowering the accuracy (from 36.8 to 33.3 %).

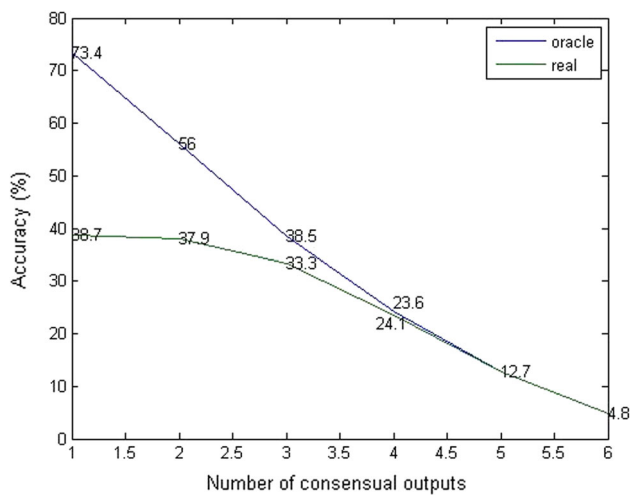


Fig. 2 Oracle scores and real systems accuracy according to the number of annotation required

6 Discussion and Perspectives

6.1 Results Summary

The originality of the experiments presented in this paper lies in the use of six emotional outputs in parallel. These outputs correspond to the combination of two training corpora and three acoustic sets. There is no quite similar work with naturalistic datasets collected in the context of HRI. The two training corpora are one realistic corpus with potential end-users from 18 to 90 years old (ARMEN) collected in the context of human-avatar interaction, and a very expressive corpus (JEMO) obtained as part of a game. The test set remains the IDV-HR corpus in all experiments (a realistic corpus with visually impaired end-users collected during the project ROMEO on Human-Robot interaction). We have used different acoustic libraries to analyse the audio signal. The three acoustic sets differ from one to another: S1-384-IS09 (from the Interspeech 2009 Emotion Challenge) is mainly a cepstral acoustic set; S2-334-LIMSI is mainly a perceptual acoustic set; and S3-293-LIMSI is a hybrid acoustic set (cepstral and perceptual). Each combination of corpus and acoustic set gives significantly different results in auto-coherence. To be as close as possible to the real settings, we use a new real-life database as a test. The cross-corpus results are almost similar in terms of global F-measure, however the configurations differ in terms of F-measure per class showing a different way to see the classification problem.

The main idea of this paper is to combine the outputs of six different system settings, in order to provide information about the reliability of the recognition. As a first result, a very simple fusion method yielded a performance improvement of about 2.3%. We believe that this result is promising and opens the door to experimenting other fusion methods.

The fusion of multiple systems is however a complex problem, and the probability-based fusion in this paper led to a performance of only 38.7 %. These results show that working on spontaneous emotional data is hard, especially with subjects with particular voices (like elderly people). There is a significant gap between the oracle score and the obtained results. Some features sets seem to be better for specific classes for both ARMEN and JEMO: S3-293-LIMSI is the best for anger recognition and S1-384-IS09 is the best for joy. These results may not be dependent only on the features used, but also on the type of data contained in the corpora. A more advanced analysis would be necessary to validate these observations.

The last result presented in this paper consists in an example of how to combine the prediction outputs of the six configurations in parallel. We use a majority vote and a prediction score method. The selection of the emotion prediction which has the maximum probability product over the six configuration, leads to an accuracy of 38 %. So we obtained a gain of 10 % relatively to the score of one unique model. Using the reliability score, we defined a probability threshold, thus discarding some instances as “unable to identify a reliable emotion”. With this configuration, we obtained a higher accuracy of 40.3 %. Improving such outputs combinations could consist in learning weights on estimation probabilities with a development corpus. Another way to improve the reliability of the system could be to adapt these weights from another corpus using techniques such as fuzzy logic.

In the ROMEO2 project which follows the ROMEO project, we have collected a new database with potential real end-users (elderly people). This corpus will be used to improve our results [43].

6.2 Towards Using Reliability in Human-Robot Interactions

The human being uses integrative analytic inference of the expressive cues perceived, as described in the Brunswikian lens model [5]. This model implies that the message emitted by one subject will be modified by several factors, such as his or her goals, social references, experiences, etc., and these same factors in the receiver, will trigger modifications in the perception of the message. Understanding and adapting to social signals is thus highly likely to produce errors when there are missing pieces of contextual information, which is often the case when one deal with an artificially intelligent system. Hence the importance to conceive systems which deal with this uncertainty.

This present work is part of the robotic project ROMEO2 which notably aims for a system able to socially interact with the user, by building a dynamic representation of his or her emotional and interactional tendencies, and selecting the most adapted social behavior in response. Our work focuses

on the detection of affective states in the user's voice, and the way this data can feed the profile of the user. Reliability measures could allow for a better control on each sequence of the decision taking process, and then minimize the spreading of erroneous computations, so as to avoid random or incoherent behaviors from the robot.

Acknowledgments This work was partially funded by the French projects FUI ROMEO and BPI ROMEO2. The authors thank coders and co-workers who participated in elaborating protocols and annotating emotional states.

References

- Alonso-Martin F, Malfaz M, Sequeira J, Gorostiza J, Salichs M (2013) A multimodal emotion detection system during human-robot interaction. *Sensors* 13:15549–15581
- Batliner A, Hacker C, Steidl S, Neth E, D'Arcy S, Russell M, Wong M (2004) "You stupid tin box"—children interacting with the aibo robot: a cross-linguistic emotional speech corpus. In: LREC, Lisbon, pp 171–174
- Batliner A, Schuller B, Seppi D, Steidl S, Devillers L, Vidrascu L, Vogt T, Aharonson V, Amir N (2011) Cognitive technologies. In: The automatic recognition of emotions in speech. Springer, Heidelberg, pp 71–99
- Batliner A, Steidl S, Neth E (2007) Laryngealizations and emotions: how many babushkas? In: Proceedings of the international workshop on paralinguistic speech—between models and data (ParaLing' 07), Saarbrücken, pp 17–22
- Benziger T, Scherer KR (2005) The role of intonation in emotional expressions. *Speech Commun* 46(3–4):252–267
- Brendel M, Zaccarelli R, Devillers L (2010) Building a system for emotions detection from speech to control an affective avatar. In: LREC, Valetta, Malta
- Buendia A, Devillers L (2014) From informative cooperative dialogues to long-term social relation with a robot. In: Mariani J, Rosset S, Garnier-Rizet M, Devillers L (eds) Natural interaction with robots, knowbots and smartphones. Springer, New York, pp 135–151
- Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss I B (2005) A database of german emotional speech. In: Interspeech, Lisbon, pp 1517–1520
- Castellano G, Leite I, Pereira A, Martinho C, Paiva A, McOwan P (2010) Inter-act: an affective and contextually rich multimodal video corpus for studying interaction with robots. In: International ACM conference on multimedia
- Chastagnol C, Clavel C, Courgeon M, Devillers L (2013) Designing an emotion detection system for a socially-intelligent human-robot interaction. In: Jokinen K, Wilcock G (eds) Towards a natural interaction with robots, knowbots and smartphones, putting spoken dialog systems into practice. Springer, New York
- Cordeschi R (2013) Automatic decision-making and reliability in robotic systems: some implications in the case of robot weapons. *AI Soc* 28:431–441
- Dautenhahn K, Werry I (2002) A quantitative technique for analyzing robot-human interactions. In: International conference on intelligent robots and systems, Lausanne
- de Visser E, Parasuraman R (2011) Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload. *J Cognit Eng Decis Mak* 5(2):209–231
- Delaborde A, Devillers L (2010) Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers. In: International Workshop on affective interaction in natural environments (AFFINE), Firenze
- Desai M, Medvedev M, Vázquez M, McSheehy S, Gadea-Omelchenko S, Bruggeman C, Yanco H (2012) Effects of changing reliability on trust of robot systems. In: ACM/IEEE international conference on human-robot interaction, pp 73–80
- Devillers L, Vidrascu L, Lamel L (2005) Challenges in real-life emotion annotation and machine learning based detection. *J Neural Netw Spec Issue Emot Brain* 18(4):407–422
- Devillers L, Martin JC (2008) Coding emotional events in audio-visual corpora. In: LREC, Marrakech
- Devillers L, Vidrascu L, Layachi O (2010) A blueprint for an affectively competent agent, cross-fertilization between emotion psychology, affective neuroscience, and affective computing. In: Automatic detection of emotion from vocal expression. Oxford University Press, Oxford
- Duhaut D (2012) A way to put empathy in a robot. In: ICAI' 10, Las Vegas
- Ekman P (1999) Handbook of cognition and emotion, Wiley, New York, chap Basic emotion
- Engberg IS, Hansen AV, Andersen O, Dalsgaard P (1997) Design, recording and verification of a danish emotional speech database. Eurospeech, Rhodes
- Eyben F, Batliner A, Schuller B, Seppi D, Steidl S (2010) Cross-corpus classification of realistic emotions: some pilot experiments. In: LREC, workshop on EMOTION: corpora for research on emotion and Affect, ELRA, Valetta, pp 77–82
- Fernandez R, Picard RW (2003) Modeling drivers' speech under stress. *Speech Commun* 40:145–159
- Han JG, Gilmartin E, Looze CD, Vaughan B, Campbell N (2012) Speech & multimodal resources: the herme database of spontaneous multimodal human-robot dialogues. In: LREC, Istanbul
- Hegel F, Giesemann S, Peters A, Holthaus P, Wrede B (2011) Towards a typology of meaningful signals and cues in social robotics. In: IEEE RO-MAN, 2011
- Jung M, Lee J, DePalma N, Adalgeirsson S, Hinds P, Breazeal C (2013) Engaging robots: easing complex human-robot teamwork using backchanneling. In: Conference on computer supported cooperative work, San Antonio
- Keizer S, Foster M, Lemon O, Gaschler A, Giuliani M (2013) Training and evaluation of an mdp model for social multi-user human-robot interaction. In: SIGDIAL
- Marchi E, Batliner A, Schuller B (2012) Speech, emotion, age, language, task and typicality: trying to disentangle performance and future relevance. In: Workshop on wide spectrum social signal processing (ASE/IEEE international conference on social computing), Amsterdam
- McKeown G, Valstar M, Cowie R, Pantic M, Schröder M (2012) The semaine database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Trans Affect Comput* 3(1):5–17
- Mower E, Metallinou A, Lee CC, Kazemzadeh A, Busso C, Lee S, Narayanan S (2009) Interpreting ambiguous emotional expressions. In: ACII vol 978(1), Amsterdam, pp 4244–4799
- Ochs M, Sadek D, Pelachaud C (2012) A formal model of emotions for an empathic rational dialog agent. *Auton Agents Multi-Agent Syst* 24(3):410–440
- Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ (ed) Advances in large margin classifiers. MIT Press, Cambridge, pp 61–74
- Ringeval F, Chetouani M, Schuller B (2012) Novel metrics of speech rhythm for the assessment of emotion. In: Proceedings of the interspeech
- Scherer KR (1986) Vocal affect expressions: a review and a model for future research. *Psychol Bull* 99(2):143–165

35. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2010) The interspeech 2010 paralinguistic challenge. In: Interspeech. Makuhari, pp 2830–2833
36. Schuller B, Vlasenko B, Eyben F, Wöllmer M, Stühlsatz A, Wendemuth A, Rigoll G (2010b) Cross-corpus acoustic emotion recognition: variances and strategies. *Trans Affect Comput IEEE* 1(2):119–131
37. Schuller B, Batliner A, Steidl S, Seppi D (2011a) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Commun, Special Issue on (Sensing Emotion and Affect-Facing Realism in Speech Processing)* 53 (9/10):1062–1087
38. Schuller B, Steidl S, Batliner A (2009) The interspeech 2009 emotion challenge. In: Interspeech, Brighton,
39. Schuller B, Steidl S, Batliner A, Nöth E, Vinciarelli A, Burkhardt F, van Son R, Weninger F, Eyben F, Bocklet T, Mohammadi G, Weiss B (2012) The interspeech 2012 speaker trait challenge. In: Interspeech, Portland
40. Schuller B, Steidl S, Batliner A, Schiel F, Krajewski J (2011b) The interspeech 2011 speaker state challenge. In: Interspeech, Firenze
41. Schuller B, Zaccarelli R, Rollet N, Devillers L (2010c) Cinema—a french spoken language resource for complex emotions: facts and baselines. In: LREC, Valetta
42. Schuller B, Zhang Z, Weninger F, Rigoll G (2011c) Using multiple databases for training emotion recognition: to unite or to vote ? In: Interspeech, Florence
43. Sehili M, Yang F, Leynaert V, Devillers L (2014) A corpus of social interaction between nao and elderly people. In: 5th international workshop on emotion, social signals, sentiment & linked open data (ES3LOD2014), LREC
44. Steinfeld A, Fong T, Kaber D, Lewis M, Scholtz J, Schultz A, Goodrich M (2006) Common metrics for human-robot interaction. In: HRI'06, Salt Lake City
45. Sun R, Moore EI (2013) Using rover for multiple databases training at the decision level for binary emotional recognition. In: ICASSP
46. Tahon M, Delaborde A, Devillers L (2011) Real-life emotion detection from speech in human-robot interaction: experiments across diverse corpora with child and adult voices. In: Interspeech, Firenze
47. Walker M, Litman D, Kamm C, Abella A (1997) Paradise: a framework for evaluating spoken dialogue agents. In: EACL '97, Madrid
48. Yagoda RE, Gillian DJ (2012) You want me to trust a robot? the development of a huma-robot interaction trust scale. *Int J Soc Robot* 4:235–248
49. Zhang Z, Weninger F, Wöllmer M, Schuller B (2011) Unsupervised learning in cross-corpus acoustic emotion recognition. In: ASRU, Honolulu

Laurence Devillers is Professor of Computer Science at Paris-Sorbonne IV University. She does her research at LIMSI-CNRS and heads the team on “Affective and Social Dimensions of Spoken Interactions” (<http://www.limsi.fr/tlp/topic2.html>), working on machine analysis of human non-verbal behaviour including audio and multimodal analysis of affective states and social signals, and its applications to Human-Robot Interaction. She participates in BPI ROMEO2 project (2013–2017) which has the main goal of building a social humanoid robot. She leads the European CHISTERA project JOKER (2013–2016), JOKE and Empathy of a Robot/ECA: Towards social and affective relations with a robot. She also contributes to computer ethics, and is member of the working group on the ethics of the research in robotics of the CERN. Prof. Devillers has (co-)authored more than 120 publications. She is a member of the board of AAAC (emotion-research.net) and of the board of the Workgroup on Affects, Artificial Companions and Interactions

(GT-ACAI), member of IEEE, ACL, ISCA and the French Association of Spoken Communication (AFCP).

Marie Tahon graduated in Engineering from the Ecole Centrale de Lyon (France) in 2007 and received the M.S. degree in mechanics, energetics, civil engineering and acoustics from the Ecole Centrale de Lyon, in 2007. She received the Ph.D. degree in informatics and signal processing from the University of Paris-Sud (Orsay, France) in 2012. She has been with the LIMSI-CNRS in the team “Affective and Social Dimensions of Spoken Interactions” since 2009. She has been a Teaching and Research Assistant in acoustics with the Structural Mechanics and Coupled Systems Laboratory (LMSSC), Conservatoire National des Arts et Métiers (Paris, France) from 2012 to 2014. She is currently with the LIMSI-CNRS. Her research interests concern automatic speech processing, i.e. automatic acoustic features extraction for both speech and musical signals. She is a member of the French Association of Spoken Communication (AFCP), of the French Acoustic Association (SFA) and of the Workgroup on Affects, Artificial Companions and Interactions (GT-ACAI).

Mohamed A. Sehili holds a Computer Science Engineering degree from El Ibrahimi University (2008, Bordj Bou Arreridj, Algeria). He received the Master degree at the University of Paris South (France) in 2009. His Ph.D. thesis on environmental sounds recognition was performed at Telecom SudParis and the University of Evry Val D'Essonne (Evry, France) between 2010 and 2013. He is currently a post-doc at the LIMSI Laboratory (Orsay, France) within the Spoken Language Processing group in the team “Affective and Social Dimensions of Spoken Interactions” since 2013. His main research interests are sound and speech detection and analysis, emotion recognition and Human-Robot interaction.

Agnes Delaborde is currently a Research and Teaching assistant at the University of Paris-Sorbonne (France), in the fields of Computer Science and Natural Language Processing. She holds a Master's degree in Natural Language Processing from the University of Grenoble. In 2013, she obtained a Ph.D. in Computer sciences from the University Paris-Sud. She has been working with the LIMSI-CNRS in the team “Affective and Social Dimensions of Spoken Interactions” since 2009, on the modeling of the user's affect and personality in human-machine interaction, in the context of the French robotic project ROMEO (2009–2012). She now contributes to ROMEO2 (2013–2017) on the modeling of the user's social and emotional tendencies in Human-Robot Interaction. She is a member of the Workgroup on Affects, Artificial Companions and Interactions (GT-ACAI).