Romeo2 Project: Humanoid Robot Assistant and Companion for Everyday Life: I. Situation Assessment for Social Intelligence [1]

Amit Kumar Pandey[1], Rodolphe Gelin[1], Rachid Alami[2], Renaud Viry[2], Axel Buendia[3], Roland Meertens[3], Mohamed Chetouani[4], Laurence Devillers[5], Marie Tahon[5], David Filliat[6], Yves Grenier[7], Mounira Maazaoui[7], Abderrahmane Kheddar[8], Frédéric Lerasle[2], and Laurent Fitte Duval[2]

[1]Aldebaran, A-Lab, France, *akpandey@aldebaran.com; rgelin@aldebaran.com*
[2]CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France;
rachid.alami@laas.fr; frederic.lerasle@laas.fr; renaud.viry@laas.fr; lfittedu@laas.fr
[3]Spirops/CNAM (CEDRIC), Paris; *axel.buendia@cnam.fr; rolandmeertens@gmail.com*
[4]ISIR, UPMC, France; *mohamed.chetouani@upmc.fr*
[5]LIMSI-CNRS University Paris-Sorbonne; *devil@limsi.fr; marie.tahon@limsi.fr*
[6]ENSTA ParisTech - INRIA FLOWERS; *david.filliat@ensta-paristech.fr*
[7]Inst. Mines-Télécom; Télécom ParisTech; CNRS LTCI;
yves.grenier@telecom-paristech.fr; maazaoui@telecom-paristech.fr
[8]CNRS-UM2 LIRMM IDH; *kheddar@gmail.com*

Abstract. For a socially intelligent robot, different levels of situation assessment are required, ranging from basic processing of sensor input to high-level analysis of semantics and intention. However, the attempt to combine them all prompts new research challenges and the need of a coherent framework and architecture.

This paper presents the situation assessment aspect of Romeo2, a unique project aiming to bring multi-modal and multi-layered perception on a single system and targeting for a unified theoretical and functional framework for a robot companion for everyday life. It also discusses some of the innovation potentials, which the combination of these various perception abilities adds into the robot's socio-cognitive capabilities.

Keywords: Situation Assessment, Socially Intelligent Robot, Human Robot Interaction, Robot Companion

1 Introduction

As robots started to co-exist in a human-centered environment, the human awareness capabilities must be considered. With safety being a basic requirement, such robots should be able to behave in a socially accepted and expected manner. This requires robots to reason about the situation, not only from the perspective of physical locations of objects, but also from that of 'mental' and 'physical' states of the human partner. Further, such reasoning should build knowledge with the human understandable attributes, to facilitate natural human-robot interaction.

The Romeo2 project (website [1]), the focus of this paper, is unique in that it brings together different perception components in a unified framework for real-life personal assistant and companion robot in an everyday scenario. This paper outlines our perception architecture, the categorization of basic requirements, the key elements to perceive, and the innovation advantages such a system provides.

[1] This work is funded by Romeo2 project, (http://www.projetromeo.com/), BPIFrance in the framework of the Structuring Projects of Competitiveness Clusters (PSPC)

Fig. 1 shows the Romeo robot and its sensors. It is a *40kg* and *1.4m* tall humanoid robot with *41 degrees-of-freedom*, *vertebral column*, *exoskeleton on legs*, *partially soft torso* and *mobile eyes*.

1.1 An Example Scenario

Mr. Smith lives alone (with his Romeo robot companion). He is elderly and visually impaired. Romeo understands his speech, emotion and gestures, assists him in his daily life. It provides physical support by bringing the 'desired' items, and cognitive support by reminding about medicine, items to add in to-buy list, playing memory games, etc. It monitors Mr. Smith's activities and calls for assistance if abnormalities are detected in his behaviors. As a social inhabitant, it plays with Mr. Smith's grandchildren visiting him.

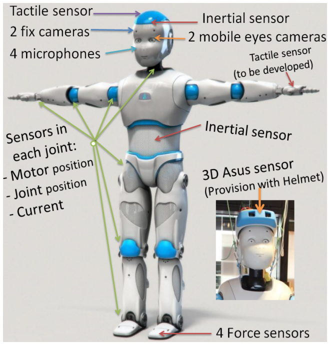


Fig. 1. Romeo robot and sensors.

This outlined partial target scenario of Romeo2 project (also illustrated in fig. 2), depicts that being aware about human, his/her activities, the environment and the situation are the key aspects towards practical achievement of the project's objective.

Fig. 2. Romeo2 Project scenario: A Humanoid Robot Assistant and Companion for Everyday Life.

1.2 Related Works and the main Contributions

Situation awareness is the ability to perceive and abstract information from the environment [2]. It is an important aspect of day-to-day interaction, decision-making, and planning, so as important is the domain-based identification of the *elements* and *attributes*, constituting the state of the environment. In this paper, we will identify and present such elements from companion robot domain perspective, sec. 2.2. Further, three levels of it have been identified (Endsley *et al.* [9]): ***Level 1 situation awareness***: To perceive the state of the *elements* composing the surrounding environment. ***Level 2 situation awareness***: To build a goal oriented understanding of the situation. Experience and comprehension of the meaning are important. ***Level 3 situation awareness***: To project on the future. Sec. 2.1 will present our sense-interact perception loop and map these levels.

Further, there have been efforts to develop integrated architecture to utilize multiple components of situation assessment. However, most of them are specific for a particular task like navigating [21], intention detection [16], robot's self-perception [5], spatial and temporal situation assessment for robot passing through a narrow passage [1], laser data based human-robot-location situation assessment, e.g. human entering, coming closer, etc. [12]. Therefore, they are either limited by the variety of perception attributes, sensors or restricted to a particular perception-action scenario loop. On the other hand, various projects on Human Robot Interaction try to overcome perception limitations by different means and focus on high-level semantic and decision-making. Such as, the detection of objects is simplified by putting tags/markers on the objects, in the detection of people no audio information is used, [6], [14], etc. In [10], different layers of perception have

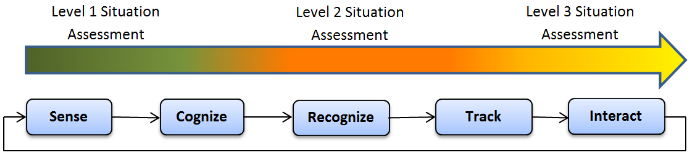

Level 1 Situation Assessment
Level 2 Situation Assessment
Level 3 Situation Assessment
Sense
Cognize
Recognize
Track
Interact

Fig. 3. A generalized perception system for sense-interact in Romeo2 project, with five layers functioning in a closed loop.

been analyzed to build representations of the 3D space, but focused on eye-hand coordination for active perception and not on high-level semantics and perception of the human.

In the Romeo2 project, we are making effort to bring a range of multi-sensor perception components within a unified framework (Naoqi, [18]), at the same time making the entire multi-modal perception system independent from a very specific scenario or task, and explicitly incorporating reasoning about human, towards realizing effective and more natural multi-modal human robot interaction. In this regard, to the best of our knowledge, Romeo2 project is the first effort of its kind for a real world companion robot. In this paper, we do not provide the details of each component. Instead, we give an overview of the entire situation assessment system in Romeo2 project (sec. 2.1). Interested readers can find the details in documentation of the system [18] and in dedicated publications for individual components, such as [4], [11], [19], [15], [3], [24], [17], [23], etc. (see the complete list of publications [1]). Further, the combined effort to bring different components together helps us to identify some of the innovation potentials and to develop them, as discussed in section 3.

2 Perceiving Situation in Romeo2 Project

2.1 A Generalized Sense-Interact Perception Architecture for HRI

We have adapted a simple yet meaningful, sensing-interaction oriented perception architecture, by carefully identifying various requirements and their interdependencies, as shown in fig. 3. The roles of the five identified layers are:

(i) Sense: To receive signals/data from various sensors. Depending upon the sensors and their fusion. This layer can build 3D point cloud world; sense stimuli like touch, sound; know about the robot's internal states such as joint, heat; record speech signals; etc. Therefore, it belongs to *level 1* of situation assessment.

(ii) Cognize: Corresponds to the 'meaningful' (human-understandable level) and relevant information extraction, e.g. learning shapes of objects; learning to extract the semantics from 3D point cloud, the meaningful words from speech, the meaningful parameters in demonstration, etc. In most of the perception-action systems, this *cognize* part is provided a priori to the system. However, in Romeo2 projects we are taking steps to make *cognize* layer more visible by bringing together different learning modules, such as to learn objects, learn faces, learn the meaning of instructions, learn to categorize emotions, etc. This layer lies across *level 1* and *level 2* of situation assessment, as it is building knowledge in terms of attributes and their values and also extracting some meaning for future use and interaction.

(iii) Recognize: Dedicated to recognizing what has been 'cognized' earlier by the system, e.g. a place, face, word, meaning, emotion, etc. This mostly belongs to *level 2* of situation assessment, as it is more on utilizing the knowledge either learned or provided a priori, hence 'experience' becomes the dominating factor.

Table 1. Identification and Classification of the key situation assessment components

(I) Perception of Human	
(i) People Presence	(ix) Perspective Taking
(ii) Face Detection	(x) Emotion Recognition
(iii) Face Characteristics	(xi) Speaker Localization
(iv) Gaze Analysis	(xii) Speech Recognition
(v) Face Recognition	(xiii) Speech Rhythm Analysis
(vi) Face and Person Tracking	(xiv) User Attention Detection
(vii) Posture Characterization	(xv) User Profile Analysis
(viii) Waving Detection	(xvi) Intention Analysis

(II) Perception of Robot Itself
(i) Battery Status
(ii) Body Temperature
(iii) Foot Status
(iv) Robot Posture
(v) Fall Detection
(vi) Self Collision Detection

(III) Perception of Object	(IV) Perception of Environment	(V) Perception of Stimuli
(i) 3D Segmentation	(i) Landmark Detection	(i) Sound Detection
(ii) Barcode Reader	(ii) Darkness Detection	(ii) Chest Button Interpretation
(iii) Close Object Detection	(iii) Place Recognition	(iii) Movement Detection
(iv) Object Recognition	(iv) Location Tracker	(iv) Sound Localization
(v) Object Tracker	(v) Sound Tracker	(v) External Collision Detection
(vi) Semantic perception	(vi) Semantic Perception (place)	(vi) Contact Observer

(iv) Track: This layer corresponds to the requirement to track something (sound, object, person, etc.) during the course of interaction. From this layer, *level 3* of situation assessment begins, as tracking allows to update in time the state of the beforehand entity (person, object, etc.), hence involves a kind of 'projection'.

(v) Interact: This corresponds to the high-level perception requirements for interaction with the human and the environment. E.g. activity, action and intention prediction, perspective taking, social signal and gaze analyses, semantic and affordance prediction (e.g. pushable objects, sitable objects, etc.). It mainly belongs to *level 3* of situation assessment, as involves 'predicting' side of perception.

Sometimes, practically there are some intermediate loops and bridges among these layers, for example a kind of loop between tracking and recognition. Those are not shown for the sake of making main idea of the architecture better visible.

Note the *closed loop* aspect of the architecture from *interaction* to *sense*. As shown in some preliminary examples in section 3, such as *Ex1*, we are able to practically achieve this, which is important to facilitate natural human-robot interaction process, which can be viewed as: *Sense* → *Build knowledge for interaction* → *Interact* → *Decide what to sense* → *Sense* →*...*

2.2 Basic Requirements, Key Attributes and Developments

In Romeo2 project, we have identified the key attributes and elements of situation assessment, to be perceived from companion robotics domain perspective, and categorized along five basic requirements as summarized in table 1. In this section, we describe some of those modules. See Naoqi [18] for details of all the modules.

I. Perception of Human

People presence: Perceives presence of people, assign unique ID to each detected person. **Face characteristics**: To predict age, gender and degree of smile on a detected face. **Posture characterization (human)**: To find position and orientation of different body parts of the human, shoulder, hand, etc. **Perspective taking**: To perceive reachable and visible places and objects from the human's perspective, with the level of effort required to see and reach. **Emotion recognition**: For basic emotions of anxiety, anger, sadness, joy, etc. based on multi-modal audio-video signal analysis. **Speaker localization**: Localizes spatially the speaking person. **Speech rhythm analysis**: Analyzing the characterization of speech rhythm by using acoustic or prosodic anchoring, to extract social signals such as engagement, etc. **User profile**: To generate emotional and interactional profile of the interacting user. Used to dynamically interpret the emotional behavior as well as to build behavioral model of the individual over a longer period of time. **Intention analysis**: To interpret the intention and desire of the user through conversation in order to provide context, and switch among different topics to talk. The context also helps other perception components about what to perceive and where to focus. Thus, facilitates closing the interaction-sense loop of fig. 3.

II. Perception of Robot Itself

Fall detection: To detect if the robot is falling and to take some human user and self-protection measures with its arms before touching the ground.

Other modules in this category are self-descriptive. However it is worth to mention that, such modules also provide symbolic level information, such as *battery nearly empty*, *getting charged*, *foot touching ground*, symbolic posture *sitting*, *standing*, *standing in init pose*, etc. All these help in achieving one of the aims of Romeo2 project: sensing for natural interaction with human.

III. Perception of Object

Object Tracker: It consists of different aspects of tracking, such as moving to track, tracking a moving object and tracking while the robot is moving. **Semantic perception (object)**: Extracts high-level meaningful information, such as object *type* (*chair*, *table*, etc.), *categories and affordances* (*sitable*, *pushable*, etc.)

IV. Perception of Environment

Darkness detection: Estimates based on the lighting conditions of the environment around the robot. **Semantic perception (place)**: Extracts meaningful information from the environment about places and landmarks (a kitchen, corridor, etc.), and builds topological maps.

V. Perception of Stimuli

Contact observer: To be aware of desired or non-desired contacts when they occur, by interpreting information from various embedded sensors, such as accelerometers, gyro, inclinometers, joints, IMU and motor torques'.

3 Results and Discussion on Innovation Potentials

We will not go in detail of the individual modules and the results, as those can be found online [18]. Instead, we will discuss some of the advantages and innovation potentials, which such modules functioning on a unified platform could bring.

Ex1: The capability of multi-modal perception, combining input from the interacting user, the events triggered by other perception components, and the centralized memorization mechanism of robot, help to achieve the goal of closing the interact-sense loop and dynamically shaping the interaction.

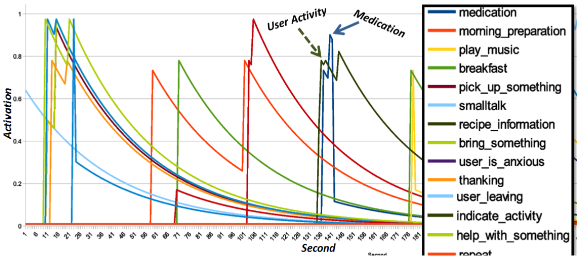
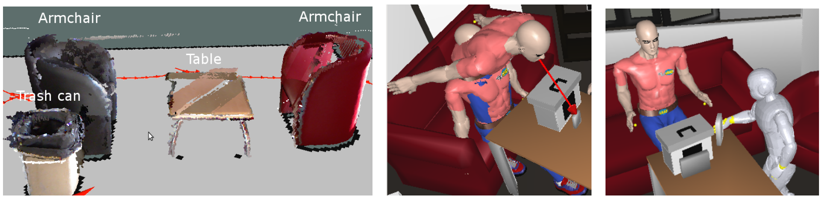


Fig. 4. Subset of interaction topics (right), and their dynamic activation levels based on multi-modal perception and events.



(a) (b) (c)

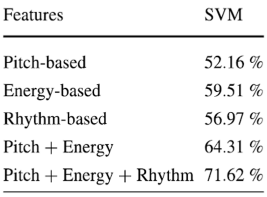
Fig. 5. High-level situation assessment. (a) The semantics map of the environment. (b) Effort and Perspective taking based situation assessment. (c) Combining (a) and (b), the robot will be able to make the object accessible to the human.

To demonstrate, we programmed an extensive dialogue with 26 topics that shows the capabilities of the Romeo robot. During this dialogue the user often interrupts Romeo to quickly ask a question, this leads to several 'conflicting' topics in the dialogue manager. The activation of different topics during an interaction over a period is shown in fig. 4. The plot shows that around *136th* second the user has to take his medicine, but the situation assessment based memory indicates that the user has ignored and not yet taken the medicine. Eventually, the system results the robot urging the user to take his medication (pointed by blue arrow), making it more important than the activity indicated by the user during the conversation (to engage in reading a book, pointed by dotted arrow in dark green). Hence, a close loop between the perception and interaction is getting achieved in a real time, dynamic and interactive manner.

Ex2: Fig. 5(a) shows situation assessment of the environment and objects at the level of semantics and affordances, such as there is a 'table' recognized at position X, and this belongs to an affordance category on which something can be put. Fig. 5(b) shows situation assessment by perspective taking, in terms of abilities and effort of the human. This enables the robot to infer that the sitting human (as shown in fig. 5(c)) will be required to stand up and lean forward to see and take the object behind the box. Thanks to the combined reasoning of (a) and (b), the robot will be able to make the object accessible to the human by placing it on the table (knowing that something can be put on it), at a place reachable and visible by the human with least effort (through the perspective taking mechanism), as shown in fig. 5(c).

In Romeo2 we also aim to use this combined reasoning about abilities and efforts of agents, and affordances of the environment, for autonomous human-level understanding of task semantics through interactive demonstration, for the development of robot's proactive behaviors, etc. as suggested the feasibility and advantages in some of our complementary studies in those directions, [19], [20].

Ex3: Analyzing verbal and non-verbal behaviors such as head direction (e.g. on-view or off-view detection) [15], speech rhythm (e.g. on-talk or self-talk) [22], laugh detection [8], emotion detection [24], attention detection [23], and their dynamics (e.g. synchrony [7]), combined with acoustic analysis (e.g. spectrum) and prosodic analysis altogether greatly allows to improve social engagement characterization of the human during interaction.



Features	SVM
Pitch-based	52.16 %
Energy-based	59.51 %
Rhythm-based	56.97 %
Pitch + Energy	64.31 %
Pitch + Energy + Rhythm	71.62 %

Fig. 6. Self-talk detection

To demonstrate, we collected a database of human-robot interaction during sessions of cognitive stimulation. The preliminary result with *14* users shows that on a *7* level evaluation scheme, the average scores for questions, "*Did robot show any empathy?*", "*Was it nice to you?*" and "*Was it polite?*" were *6.3*, *6.2* and *6.4* respectively. In addition, the multi-modality combination of the rhythmic, energy and pitch characteristics seems to be elevating the detection of self-talk (known to reflect the cognitive load of the user, especially for elderly) as shown in table of fig. 6.

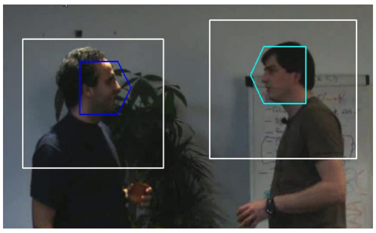


Fig. 7. Face, shoulder and face orientation detection of two interacting people.

Ex4: Inferring face gaze (as illustrated in fig. 7), combined with sound localization and object detection, altogether provides enhanced knowledge about who might be speaking in a multi-people human-robot interaction, and further facilitates analyzing the attention and intention. To demonstrate this, we conducted an experiment with two speakers, initially speaking at the different sides of the robot and then slowly moving towards each other and eventually separate away. Fig. 8 shows the preliminary result for the sound source separation by the system based on beamforming. The left part (BF-SS) shows when only the audio signal is used. When the system uses the visual information combined with the audio signals, the performance is better (AVBF-SS) in all the three types of analyses: signal-to-interference ratio (SIR), signal-to-distortion ratio(SDR) and signal-to-artifact (SAR) ratio.

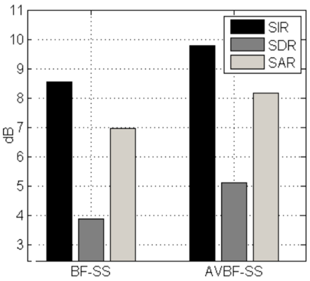


Fig. 8. Sound source separation, only audio based (BF-SS) and audio-video based (AVBF-SS).

Ex5: The fusion of rich information about visual clues, audio speech rhythm, lexical content and the user profile is also opening doors for automated context extraction, helping for better interaction and emotion grounding and making the interaction interesting, like doing humor [13].

4 Conclusion and Future Work

In this paper, we have provided an overview of the rich multi-modal perception and situation assessment system within the scope of Romeo2 project. We have presented our sensing-interaction perception architecture and identified the key perception components requirements for companion robot. The main novelty lies in the provision for rich reasoning about the human and practically closing the sensing-interaction loop. We have pointed towards some of the work in progress innovation potentials, achievable when different situation assessment components are working on a unified theoretical and functional framework. It would be interesting to see how it could serve as guideline in different context than companion robot, such as robot co-worker.

References

1. Beck, A., Risager, C., Andersen, N., Ravn, O.: Spacio-temporal situation assessment for mobile robots. In: Int. Conf. on Information Fusion (FUSION) (2011)

2. Bolstad, C.A.: Situation awareness: Does it change with age. vol. 45, pp. 272–276. Human Factors and Ergonomics Society (2001)
3. Buendia, A., Devillers, L.: From informative cooperative dialogues to long-term social relation with a robot. In: Natural Interaction with Robots, Knowbots and Smartphones, pp. 135–151 (2014)
4. Caron, L.C., Song, Y., Filliat, D., Gepperth, A.: Neural network based 2d/3d fusion for robotic object recognition. In: Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2014)
5. Chella, A.: A robot architecture based on higher order perception loop. In: Brain Inspired Cognitive Systems 2008, pp. 267–283. Springer (2010)
6. CHRIS-Project: Cooperative human robot interaction systems. http://www.chrisfp7.eu/
7. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. Affective Computing, IEEE Transactions on 3(3), 349–365 (July 2012)
8. Devillers, L.Y., Soury, M.: A social interaction system for studying humor with the robot nao. In: ICMI. pp. 313–314 (2013)
9. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors: Journal of the Human Factors and Ergonomics Society 37(1), 32–64 (1995)
10. EYESHOTS-Project: Heterogeneous 3-d perception across visual fragments. http://www.eyeshots.it/
11. Filliat, D., Battesti, E., Bazeille, S., Duceux, G., Gepperth, A., Harrath, L., Jebari, I., Pereira, R., Tapus, A., Meyer, C., Ieng, S., Benosman, R., Cizeron, E., Mamanna, J.C., Pothier, B.: Rgbd object recognition and visual texture classification for indoor semantic mapping. In: Technologies for Practical Robot Applications (2012)
12. Jensen, B., Philippsen, R., Siegwart, R.: Narrative situation assessment for human-robot interaction. In: IEEE ICRA. vol. 1, pp. 1503–1508 vol.1 (Sept 2003)
13. JOKER-Project: Joke and empathy of a robot/eca: Towards social and affective relations with a robot. http://www.chistera.eu/projects/joker
14. Lallee, S., Lemaignan, S., Lenz, A., Melhuish, C., Natale, L., Skachek, S., van Der Zant, T., Warneken, F., Dominey, P.F.: Towards a platform-independent cooperative human-robot interaction system: I. perception. In: IEEE/RSJ IROS. pp. 4444–4451 (Oct 2010)
15. Le Maitre, J., Chetouani, M.: Self-talk discrimination in human-robot interaction situations for supporting social awareness. J. of Social Robotics 5(2), 277–289 (2013)
16. Lee, S., Baek, S.M., Lee, J.: Cognitive robotic engine: Behavioral perception architecture for human-robot interaction. In: Human Robot Interaction (2007)
17. Mekonnen, A.A., Lerasle, F., Herbulot, A., Briand, C.: People detection with heterogeneous features and explicit optimization on computation time. In: ICPR (2014)
18. NAOqi-Documentation: https://community.aldebaran-robotics.com/doc/2-00/naoqi/index.html/
19. Pandey, A.K., Alami, R.: Towards human-level semantics understanding of human-centered object manipulation tasks for hri: Reasoning about effect, ability, effort and perspective taking. Int. J. of Social Robotics pp. 1–28 (2014)
20. Pandey, A.K., Ali, M., Alami, R.: Towards a task-aware proactive sociable robot based on multi-state perspective-taking. J. of Social Robotics 5(2), 215–236 (2013)
21. Pomerleau, D.A.: Neural network perception for mobile robot guidance. Tech. rep., DTIC Document (1992)
22. Ringeval, F., Chetouani, M., Schuller, B.: Novel metrics of speech rhythm for the assessment of emotion. Interspeech pp. 2763–2766 (2012)
23. Sehili, M., Yang, F., Devillers, L.: Attention detection in elderly people-robot spoken interaction. In: ICMI WS on Multimodal Multiparty real-world HRI (2014)
24. 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. pp. 3121–3124 (2011)