

Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the “socio-affective glue”

Yuko Sasa¹, Véronique Aubergé¹

¹ LIG-lab CNRS, Grenoble University, France

Yuko.Sasa@imag.fr, Veronique.Auberge@imag.fr

Abstract

The aim of this preliminary study of feasibility is to give a glance at interactions in a Smart Home prototype between the elderly and a companion robot that is having some socio-affective language primitives as the only vector of communication. Through a Wizard of Oz platform (EmOz), a robot is introduced as an intermediary between the technological environment and some elderly who have to give vocal commands to the robot to control the Smart Home. The robot vocal productions increases progressively by adding prosodic levels: (1) no speech, (2) pure prosodic mouth noises supposed to be the “glue’s” tools, (3) lexicons with supposed “glue” prosody and (4) subject’s commands imitations with supposed “glue” prosody. The elderly subjects’ speech behaviors confirm the hypothesis that the socio-affective “glue” effect increase towards the prosodic levels, especially for socio-isolated people.

Index Terms: socio-affective “glue”, human-robot interaction, affective prosody, elderly, Smart Home.

1. Introduction

Prosody carries emotional, socio-affective and interactional information where each language has its own values [10]. This communicative information appears in different prosodic levels as in non-lexical sounds. Those can be non-phonetic sounds like grunts, affect bursts or mouth noises [20,24], phonological as fillers, mind markers or interjections [25], or onomatopoeia, widely studied in Japanese [26]. These sounds that we can consider as pure prosodic tools, were studied for specific and supposed emotional [5] and pragmatic [1] functions, as well as moods, emotions, intentions, attitudes, cognitive processes and mental states also known as “Feeling of Thinking” [14]. Moreover lexicons, sentences and paraphrases prosodic form also support various socio-affective values [29]. These cues can be extended from simple sounds to sentences produced in a same context, which have been tested in synthesis [12,15,17]. Lately, the prosody carrying this communicative information was introduced as a way to develop “socio-affective glue”, terms introduced for the first time in [2]. In face-to-face interaction, the communication channel used by the speakers depends of the context and their social role, giving clues on how people have to talk to each other. Moreover, the “glue” refers to the fact that the interlocutors build dynamically their relation and adjust constantly the way they converge or not, globally changing the basic communicative channel firstly introduced by the context and their role. However, this kind of dynamic process can be more difficult to handle for some isolated-person, typically the elderly [3] but also younger person, like the Japanese hikikomori [11] or autistic people [6]. Furthermore, imitation has been studied as a basic process to create the same kind of “glue” in children language acquisition [8] or as a primitive of

robots learning [23]. By the way, since the 90’s, Affecting Computing and multidisciplinary communities have been focusing their work on the face-to-face interactions, especially on facial, gestural and vocal expressions using virtual agents and robots as in various studies in social computing [4,16]. It is interesting to see that when a robot is not explicitly humanoid, human creates by himself a socio-affective relationship with this device toward its « pet » stance [7]. Because all these different prosodic levels have not been studied together particularly to see their functions in the “socio-affective glue” building, our work will test them progressively thanks to a robot interacting with elderly towards gradual vocal productions: (1) no speech, (2) pure prosodic mouth noises supposed to be the “glue’s” tools, (3) lexicons with supposed “glue” prosody and (4) subject’s commands imitations with supposed “glue” prosody. This will be tested with a Wizard of Oz protocol in a Smart Home prototype that has been used to study speech in a natural context environment with vocal commands, including elderly speech recognition [28].

2. Methodology

First of all, this study is testing the feasibility of our protocol to observe the gradual prosody productions effect on the “socio-affective glue” develop or not between elderly who are giving Smart Home’s vocal commands and a robot “controlling” this Smart Home. The spontaneous corpus briefly analyzed in this study has been collected thanks to the EMOX robot (www.awabot.com/) interacting with the elderly in a Smart Home prototype named DOMUS. A Wizard of Oz platform called EmOz [2] was developed to control both the robot and the environmental system. The Wizards followed an accurate script describing which robot vocal productions they have to use and when they need to be produced. The Wizards also controlled the Smart Home each time the subjects gave a vocal command addressed to EMOX as we told them it was the robot that controlled DOMUS. This protocol is associated to a scenario with a specific recruiting pretext we were able to play thanks to elderly’s caregivers (www.bienalamaison.com). The caregivers who were accompanying the subjects during the experiment were aware of the Wizard of Oz trick and our scientific goal being the researchers accomplices.

2.1. Experimental tools

2.1.1. DOMUS, a voice commanded Smart Home

As concept of living labs [19], DOMUS is designed like a 40m² flat with a kitchen, a bedroom and a living room equipped with two cameras and two microphones in each room, and a shower room with a microphone. Here, we selected a few possible actions in DOMUS to propose 31 voice commands that we can simulate from a control room. E.g.: *Mettre/Eteindre la lumière (fr)* – *To turn on/off the light*

(en); *Monter/descendre les stores* (fr) - *To up/down blinds*; *Moins/plus fort la télé* (fr) - *Lower/louder TV* (en).

2.1.2. EMOX robot with gradual vocal productions

The EMOX robot used in this study doesn't look like a human or an animal which morphology can induce the way we picture this tool and would create artifacts that we cannot control, and so the only way to find some anthropomorphism for this robot is its speech. In primarily study we attempted to change the Fundamental Frequency (F0) of spontaneous vocal micro-expressions, without modifying the prosodic contours to find a coherent voice for the robot [22]. From this database, we selected some of non-lexical sounds as part of the robot speech (e.g. *euh*, and *laughs*). We also recorded extra-sounds (see *table1a*), lexicons/interjections (see *table1b*) and commands imitations (see *table1c*) all with specific prosodies that we suppose to have an effect on the “socio-affective glue”. In order to have a homogeneous voice for all our stimuli, we used Voxal (www.nchsoftware.com/voicechanger), voice conversion software. For parameterization, we increased the pitch by 1.52 from the original female speaker's voice, who recorded the sounds and so the robot has got a pitched “cartoon-like voice”. So the voice esthetics has a reduced anthropomorphism, as we only want to observe information on the communicative functions carried by the chosen prosodic forms. Table1 shows the 30 vocal stimuli used by EMOX. 16 other sentences (also with supposed “glue prosody”) were also pre-recorded, used only if it is necessary reacting to keep the subject motivation [e.g.: *Bonjour, je suis Emox(fr)*= *Hello, I am Emox(en)*; *Oui, je suis là, j'écoute(fr)*=*Yes, I am here, I am listening* (en); *Oh pardon(fr)*=*Oh sorry(en)*].

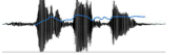





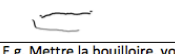

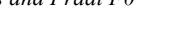
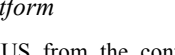
Sounds levels	Sounds nature	Examples of F0 contours
(a) Mouth noises	3 types of laughs	
	Various prosody : euh - hum1 - hum2	E.g. Hum2 
	prosody 1: humhum	
	2 onomatopoeia « woup » (associated to up/down blinds movements)	
(b) Interjections	prosody 1 : d'accord1 - ok1 - oui1	
	prosody 2 : d'accord2 - ok2 - oui2	
	ça va - comme ça - ça y est	
	voilà1, voilà2	
	ouais, oops	
(c) Commands imitations	Commands in infinitive form + (b) interjections or (a) onomatopoeia	E.g. Mettre la bouilloire, voilà 

Table 1. *Emox robot stimuli types and Praat F0 contours*

2.1.3. EmOz – the Wizard of Oz Platform

In order to operate EMOX and DOMUS from the control room, we developed a Java interface associated to the robot Urbi system and the Smart Home protocols [2]. This interface

is filled with buttons and each one of them controls one speech act of EMOX or one DOMUS command (see *Figure1*). To move the robot we use a game controller so the robot can move around to follow the subjects, turn on itself, move its head (top/down, right/left) and go forward/backward. We also have the possibility to record our voice (changed through Voxal) and live play it if needed, to manage a coherent answer when the subjects' reactions were unpredictable.

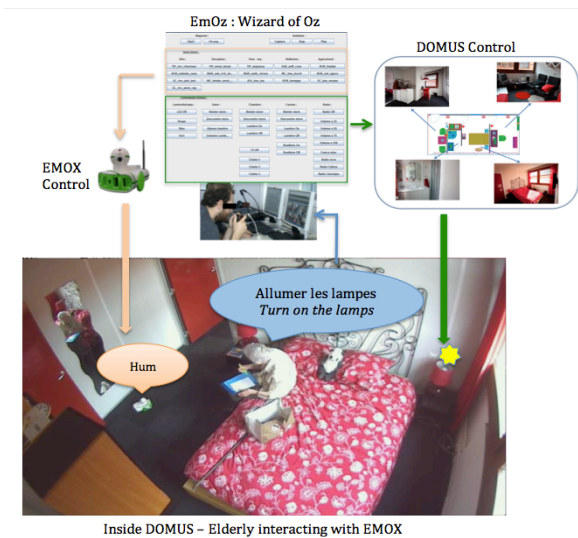


Figure1. *Scheme of EmOz wizard of Oz platform*

2.2. Experimental scenario

2.2.1. Subjects and caregivers

This study hired 4 French elderly subjects from 68 to 89 years old (one man and three women) accompanied by a caregiver/home helper they know well. The need of a helper was a criterion to select a “fragile but not sick subject”, as it is difficult to find this kind of people and their degree of frailty corresponds to the scale 6 and 5 of the GIR grid [21], which is a French reference to evaluate an elderly ability. Moreover, these helpers were our accomplices, knowing the real goal of our study with the robot. Their role allowed us to reinforce the safety of the “fragile subjects”. They also helped us to observe to observe the relation that the elderly had been creating with the robot. Furthermore, the helpers gave us an excuse to leave the subjects alone with the robot during one experiment step.

2.2.2. Experiment context and consent

To bring the subjects in DOMUS, we used a pretext to motivate the elderly to come, giving them a false task to do during the experiment. We did not mention the presence of the robot at first so they can interact spontaneously with it. Every subject was told to be the first participant of the experiment in order to keep a coherent scenario. In term of consent, the subjects had to sign three documents. Before the experiment we gave to the elderly a “pre-experimental consent” based on the pretext. Once they finished, they signed a “post-experimental consent” and an “image rights document” revealing the purpose of the Wizard of Oz. The helpers had their own single consent as the experimenters' accomplices.

2.2.3. *Pretext of recruitment*

We proposed a fake object of study to recruit the elderly. The subjects were asked to come in our Smart Home prototype because we were developing new technologies for elderly to allow them to live as long as possible in their own house, thinking of their safety and their welfare. Nevertheless we are currently not able to equip directly their home yet, so we need the subjects to come in our flat prototype. However, the problem we observed in “previous study” was that when elderly change their living environment (e.g. move into a retirement/nursing home or a hospital) they mostly have difficulties to accustom to this new place, especially when it is based on technologies. One of our “so-called hypotheses” was that if people bring some personal items (e.g. books, trinkets, decorative objects...etc.), which can be benchmarks as in Alzheimer disease, and they arrange their living environment with these items, people get used to the place more easily. Finally we asked the subjects to bring around ten items they choose and care about to come and spend some time (an hour or two) to evaluate our flat equipment, at the same time they personalize it with their own items. For their security, we asked them to be accompanied by one of their caregivers, who were also the experimenters’ accomplices.

2.2.4. *Experimenters roles*

To play our experiment’s scenario, we need two experimenter-actors interacting with the subjects:

- A “student-recruiter” who explains the pretext to the subjects and organizes the arrangement with the helper. He pretends not knowing what kind of technologies the flat is equipped with because it is his advisor who communicated with the engineering staff.
- An “engineer” who does not know the recruiter and was only asked by the student’s advisor to explain how the Smart Home works. He is not aware of the study’s aim.

Technically, we also needed at least two other experimenters to manipulate the EmOz platform from the control room: the first one used the java interface (see Figure 1) producing the EMOX vocal reactions (*in orange*) and the DOMUS commands (*in green*). The second one was coordinating the robot movements. All these reactions were listed within a specific order regarding the voice commands that can be produced by the subjects, and these two Wizards followed this experimental script.

2.2.5. *Experiment steps*

The experiment scenario itself is divided into six steps:

- Step 1: the engineer welcomes the subject, the helper and the recruiter to present the Smart Home. He does not mention at this time the voice commands use and the robot is hidden. Very quickly, the helper receives a fake emergency call simulated from the control room, (before showing EMOX). The helper is asked to act like he has an urgent mission and needs to leave DOMUS for a short time, but she will come back very soon. The helper has not got any mean of transportation so the recruiter offers the ride, while the engineer explains to the elderly how the Smart Home works, waiting for their return. The helper and the recruiter use this excuse to go to the control room. Before leaving, the recruiter asks the subject to start arranging the flat with the personal items.

- Step 2: the engineer introduces EMOX to control the Smart Home. He presents a 31 commands list that can possibly be “activated”. To work on a better quality speech signals, the engineer places a lapel microphone on the subject, justifying it by the fact that the robot vocal recognition system is not good enough without it. For the robot to be able to recognize the subject’s voice, the elderly is asked to try once all the voice commands one by one. When this training period is done, the engineer leaves the subject pretending to have some work to do.
- Step 3: It is an improvisation stage. The subject is alone with the robot, setting down his things in DOMUS.
- Step 4: After about 10 to 20 minutes, we ask the helper to go back into DOMUS and let the subject explain how the Smart Home works. If asked, the caregiver says that the recruiter was caught to talk about administrative stuffs and will return soon. EMOX does not realize the helper’s commands and it only obeys to the subject.
- Step 5: When the subject has showed most of the voice commands to the helper, it is the recruiter turn to ask how DOMUS is working. If the subject himself tells something about the robot, the recruiter listens and asks about it, just out of curiosity. Nevertheless, the recruiter real task is to remind the pretext, so he starts a “fake interview” to know how the subject got used to DOMUS thanks to the items he brought.
- Step 6: During this debriefing stage, the engineer comes back and asks what the subject has thought about the voice commands technologies, without mentioning the robot at first. Here we want to know if the subject will tell about the robot and how he will talk about it. Once EMOX is evoked, we ask about the robot’s voice, morphology, use, ways of speaking, functions, ability, role and personality. After this interview, the experimenters tell the real purpose of the study, revealing the Wizard of Oz tricks.

3. Discussions

3.1. The Elderly Emox Expressions Corpus

This corpus is composed by 4 experiments lasting from 1.5 to 2 hours each. For each subject, we have six videos (two per rooms) and an audio file collected by the subjects’ lapel microphone. We then selected 167 interactions between EMOX and the elderly (from 43 to 52 per subject), throughout the interactions held in step 2 to 5, while the subjects: (a) are learning the commands with the engineer, (b) are alone with Emox, (c) are explaining how DOMUS and Emox work to the helper and then to the recruiter. Each interaction lasts about 10 to 50 seconds, showing a sequence of exchanges around one voice command. In this corpus, we observed two types of subjects’ vocal productions of the “socio-affective glue”: the commands form produced by the subjects and some of their answers after EMOX feedback. As the robot was introduced like a simple vocal commands receptor, there is no reason for the subjects to talk or ask other things to the robots. Moreover, as they had a specific list of voice commands to control DOMUS, they had no reason to produce other forms of commands as well as react to simple automation. Then, if they change their way to address their commands to EMOX or try to talk to him, we suppose that something in the robot’s speech

prosody lead them to change their way to interact with it. The purpose of this for preliminaries analyses

3.2. Results

3.2.1. Debriefing global impressions

While debriefing, the main question we ask is “what opinion do you have concerning the system and the way you can control the Smart Home by voice?” (without mentioning at all the robot). For all the subjects, the first answer they gave us was: “the robot is a good company”, so the technical support of the vocal commands is not mentioned first, meanwhile we introduced it in this way. Moreover, when we ask the subjects if they prefer the same commands but address directly to the Smart Home, they assert again that the robot is useful but not only: “we don’t feel lonely, he talks to us”. So we can say that some “socio-affective glue” was established between the robot and the elderly. We then want to know how by observing: (a) the commands form, particularly the prosody associated to them; and (b) the reactions of the subjects after EMOX feedbacks with gradual prosody levels.

3.2.2. Subjects commands tendency

During the interactions held in step 2 to 5, there are of course some variations concerning the subject’s behaviors during the experiment first steps, but some common main characteristics emerged as features of the “glue” building toward different steps resumed in Figure 2: (a) declarative commands without paraphrasing; (b) the same original form commands but with a positive attitude prosody (in particular fundamental frequency arise which systematically appears at the end of the sentences, with a breathy voice); (c) commands paraphrased variations (used in synergy with a “we”, illustrating the same kind of joint attention that we can observe in a mother-child dyad in an early language development context [27]) with a globally high fundamental frequency and a great arise at the end of the sentences; and finally (d) multiple prosodic focuses of support terms with a higher fundamental frequency. These phenomena are observed as well as a voice quality becoming more and more breathy. This elderly’s voice quality breathiness seems to vary particularly while the robot produced a feedback based on pure prosodic vocal micro-sounds.

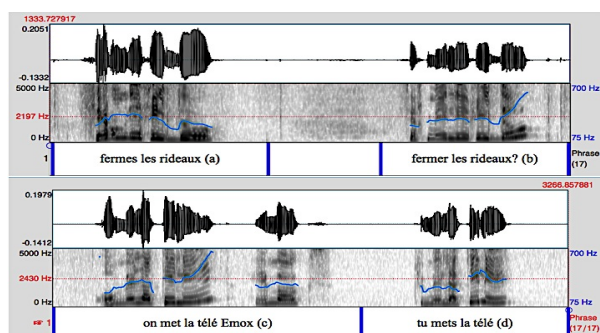


Figure 2. Commands prosody: Praat spectrogram and F0 contours of some S1 subject's commands

3.2.3. Subjects reactions to EMOX feedbacks

For every subject, nothing happened while there was no EMOX speech reaction. The subjects just continue to read commands.

But after the first appearance of pure prosodic mouth noises (see Table 1.a), we always notice specific subjects reactions as showed in Table 2. Moreover, the voice breathiness seems to be amplified after mouth noises. Once the relation is established, the robots’ lexicons and mouth noises become the base of the subjects’ imitations and echolalia as answer to EMOX feedbacks. This can be a clue on the degree of “glue” [13] to know when to introduce imitations that increase the “glue”. The timing of EMOX’s reactions and the silence duration are also an important factor in the “socio-affective glue” process that is also a major key for the interaction synchrony [9], but it also seems to be the moment that we can see if the “glue” is established. Indeed, even if EMOX (in fact the Wizards) makes mistakes or is slow, the subjects do not blame the robot but spend this response time justifying the commands, being politer or heartening it with compliments.

Types of subjects feedbacks	Quantity
Commands associated to politeness and compliments	5
Commands justifications**	13
Echolalia	10
Imitation	45
Kind reproaches	9
Mouth noises, interjections and laughs	49
Politeness forms**	11
Punctual compliments**	16
Spatial proximity	28

**During silence or timelag after a vocal command

Table 2. Subjects’ reactions distribution as answer to EMOX feedbacks

4. Conclusion

The elderly’s speech behaviors confirm that the effect of socio-affective “glue” increases towards the prosodic levels. This starts by the pure prosodic mouth noises that seem to be essential to initiate the relationship between the robot and a human. Once the link is created, the supposed prosodic “glue” tends to reinforce the bond through lexicons echolalia, imitations supported by a breathy voice quality. To have a cleaner corpora collection, the way to introduce the robot becomes essential as it induces the basic communication channel, where the glue is adding its effects after. To allow a precise control of the robot reactions timing and order, we need an efficient interface so the cognitive effort of the Wizard of Oz experimenter is the same as the effort the robot “seems to produce” to execute the commands. This will be one of these study perspectives, in addition to expanding this experiment with more various stimuli and subjects, on a revised and longer scenario to confirm those tendencies.

5. Acknowledgements

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the Interobot project (Investissements d’Avenir, DGCIS) collaboration with Awabot (robotics) company. We thank *Bien à la Maison* company (elderly personal services) for their active participation to held the corpus. Best thanks to Tim Robert, Nicolas Bonnefond and Brigitte Meillon, for their precious technical contribution and their organization support for the experiments. Finally thanks to Leandra Batista for her advice on the results analyses and to Gilles De Biasi for his reviewing.

6. References

- [1] Ameka, F (1992) "*Interjections: The universal yet neglected part of speech*", Journal of Pragmatics, 18, 101-118, 1992.
- [2] Aubergé V., Sasa Y., Robert T., Bonnefond N., Meillon B. (2013) "*Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot*". In proceedings of WASSS 2013, Grenoble, France.
- [3] Bayles K.A. and Kaszniak A. (1987) "*Communication and Cognition: Normal Aging and Dementia*", Little, Brown, Boston.
- [4] Breazeal C. and Aryananda L. (2002) "*Recognition of affective communicative intent in Robot-Directed speech*", Autonomous Robots 12, pp 83-104.
- [5] Campbell N. (2004) "*Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language*", Languages Resources and Evaluation, 39, 109-118.
- [6] Chaby L., Chetouani M., Plaza M., Cohen D. (2012) "*Exploring multimodal social-emotional behaviors in autism spectrum disorders*". In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust; IEEE Computer Society: Washington, DC, USA, 2012, pp. 950–954.
- [7] Darling K. (2012). "*Extending Legal Rights to Social Robots*". We Robot Conference, University of Miami, April 2012.
- [8] Decety J. (2007). "*A social cognitive neuroscience model of human empathy*". In E. Harmon-Jones & P. Winkelman (Eds.), Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior (pp. 246-270). New York: Guilford Publications.
- [9] Delaherche E., Chetouani M., Mahdhaoui A., Saint-Georges C., Viaux S., and Cohen D. (2012) "*Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines*". IEEE Transactions on Affective Computing 3(3), pp. 349-365.
- [10] Fonagy P., & Target M. (1997). "*Attachment and reflective function: Their role in self-organization*", Development and Psychopathology, 9, pp. 679-700.
- [11] Furlong A. (2008) "*The Japanese hikikomori phenomenon: acute social withdrawal among young people*", The Sociological Review, 56(2), pp. 309-325.
- [12] Greenberg Y., Tsuzaki M., Kato H. and Sagisaka Y. (2006) "*A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech*". In Proceedings of Speech Prosody 2006, pp. 37-40.
- [13] Ladd D.R. and Cuttler A. (1983) "*Models and measurements in the study of prosody*". In A. Cutler & D. R. Ladd (Eds.), Prosody: Models and measurements, Heidelberg: Springer-Verlag, pp. 1-10.
- [14] Loyau F., Aubergé V. (2006). "*Expressions outside the talk turn: ethograms of the Feeling of Thinking*", 5th LREC, pp.47-50.
- [15] Mac D., Castelli E, Aubergé V. (2012). "*Modeling the prosody of Vietnamese attitudes for expressive speech synthesis*". Workshop of Spoken Languages Technologies for Under-resourced Languages (SLTU 2012), Cape Town, South Africa.
- [16] Mairesse F., Walker A., Mehl Matthias M. and Moore R.K. (2007) "*Using linguistic cues for the automatic recognition of personality in conversation and text*". In Journal of Artificial Intelligence Research 30, pp. 457-500.
- [17] Morlec Y., Bailly G. and Aubergé V. (2001). "*Generating prosodic attitudes in French: data, model and evaluation*". Speech Communication, 33(4), pp.357-371.
- [18] Morenc L.P (2010) "*Modeling human communication dynamics*", Social Sciences. Signal Processing Magazine, IEEE, 27(5), pp. 112 -116.
- [19] Niitamo V.-P., Kulkki S., Eriksson M., Hribernik K. A. (2006) "*State-of-the-art and good practice in the field of living labs*". In Proceedings of the 12th International Conference on Concurrent Enterprising: Innovative Products and Services through Collaborative Networks, Milan, Italy, pp.349-357.
- [20] Poggi I. (2008) "*The language of interjections*". In COST 2012 School, pp.170–186.
- [21] Renault S. (2004). "*Du concept de fragilité et de l'efficacité de la grille AGGIR*". In Gérontologie et société. 2004, n° 109, pp.83-107.
- [22] Sasa Y., Aubergé V., Franck P., Guillaume L., Moujtahid S. (2012). "*Des micro-expressions au service de la macro-communication pour le robot compagnon EMOX*", WACAI 2012, Grenoble, pp.54-59.
- [23] Schaal S. (1999) "*Is imitation learning the route to humanoid robots?*" Trends Cognit. Sci.3, 233-242.
- [24] Scherer K.R., "*Affect bursts*" (1994). In S.H.M. van Goozen, N. E. van de Poll & J.A. Sergeant (Eds.), Emotions, Hillsdale (NJ, USA), Lawrence Erlbaum, 161-193
- [25] Schröder M., Heylen D., Poggi I. (2006) "*Perception of non-verbal emotional listener feedback*". In Speech Prosody, R. Hoffmann & H. Mixdorff, Eds., 2006, pp. 1–4..
- [26] Shibatani M. (1990) "*The languages of Japan*". Cambridge: Cambridge University Press.
- [27] Tomasello M., Carpenter M., Call J., Behne T. & Moll H. (2005) "*Understanding and sharing intentions: The origins of cultural cognition*". Behavioral and Brain Sciences 28(5), pp. 675–91.
- [28] Vacher M., Fleury A., Portet F., Serignat J.-F., Noury N. (2010) "*Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living*", New Developments in Biomedical Engineering, Intech Book, pp. 645-673.
- [29] Wichmann A. (2000) "*The attitudinal effects of prosody, and how they relate to emotion*". In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.