Recognising the Social Attitude in Natural Interaction with Pedagogical Agents

Berardina De Carolis, Stefano Ferilli, Nicole Novielli,
Fabio Leuzzi, Fulvio Rotella
Dipartimento di Informatica, University of Bari, Italy
{decarolis, ferilli, novielli}@di.uniba.it
{fabio.leuzzi, fulvio.rotella}@uniba.it

Pedagogical Conversational Agents (PCAs) have the advantage of offering to students not only task-oriented support but also the possibility to interact with the computer media at a social level. This form of intelligence is particularly important when the character is employed in an educational setting. This paper reports our initial results on the recognition of users' social response to a pedagogical agent from the linguistic, acoustic and gestural analysis of the student communicative act.

1. Introduction

The concept of "Pedagogical Conversational Agent" (PCA) refers to the capability of this software to combine the capability to fullfill pedagogical goals and to interact with the user through a natural dialog by appropriately mixing verbal and non verbal expressions [Cassell, 2001]; moreover, a PCA has to show an autonomous and intelligent behaviour (a typical feature of software agents). In general, Embodied Conversational Agents (ECAs) are used in Human Computer Interaction (HCI) since people interact with computers as if they were social actors [Reeves and Nass, 1996]. For this purpose, an ECA must be able to: i) recognize and answer to verbal and non-verbal inputs, ii) generate verbal and non verbal outputs, iii) handle typical functions of human conversations, with particular emphasis on social aspects. Due to these features, many learning environments have been integrated with PCAs [D'Mello et al, 2008; Johnson et al, 2000; Baylor and Kim 2005], aimed at increasing the ability of the learning systems to engage and motivate the learners. Thus, they should be able not only to adapt their behavior to the cognitive skills and capabilities of the learner, but also to tune their social behavior for accommodating critical situations from the social point of view (e.g. closure and negative attitudes). [Kim and Baylor, 2008] argue that digital pedagogical characters may improve the quality of the learning task by providing situated social interaction, that traditional computer-based learning environments often fail to provide. In this case, it is important to endow the character with the capability of recognizing also the learner social attitude in order to interleave task and domain-oriented conversation with a more socially-oriented one by establishing a social relation with the students [Bickmore, 2003]. To this aim,

besides modeling cognitive ingredients of the user's mental state, a conversational interface of this kind should consider also extra-rational factors such as *empathy* [Paiva, 2004; Hoorn and Konijn, 2003] *engagement, involvement*, *sympathy or distance*.

This paper describes a preliminary study aimed at building a multimodal framework for the recognition of the social response of users to a PCA in the context of a system aimed at providing useful concepts about a correct diet and healthy eating. In our framework the agent is projected on the wall and the user interacts with it using an ambient microphone and Microsoft Kinect [MK, 2011]. Since the combination of speech and gesture is a natural way for humans to interact, we decided to extend our model of social attitude recognition presented in [De Carolis et al, 2012]. In the basic model the signals of social attitude are derived from the linguistic analysis of the user move. In the new version, we propose a framework that integrates the analysis of the linguistic component of the user's communicative act with the analysis of the acoustic features of the spoken sentence and of the gestures. The underlying idea is that the combination of these different input modalities may improve the recognition of multimodal behaviours that may denote the openness attitude of the users towards the embodied agent. The framework was built as a Dynamic Bayesian Network, due to the ability of this formalism in representing uncertainty and graduality in building an image of the user cognitive and affective state of mind. Dynamic recognition during the dialogue of individual signs of social signalling in language, prosody and gestures, not only enables an estimation of the overall social attitude value, but also allows the agent to adapt its dialogue plan accordingly. To this aim we designed an experimental setting to collect a corpus of multimodal conversations with an ECA in a Wizard of Oz (WoZ) simulation study [Clarizio et al, 2006]. Then, after carefully examining our corpus and considering suggestions from the studies about verbal and non-verbal expression of social attitude [Andersen and Guerrero, 1998; Polhemus et al, 2001; Swan, 2002; Pease, 2006], we annotated the user moves in the corpus according to the social attitude conveyed by users in each multimodal dialogue move. Then, we tested our model on the resulting dataset.

The paper is structured as follows: in Section 2 we provide a brief description of the conceptual framework, by addressing the issues related to the modelling of social attitude using linguistic, acoustic and body features; in Section 3 we describe the dynamic modelling approach used for integrating the results of the multimodal analysis; then, in Section 4 we describe the experiment for collecting the corpus. Section 5 provides a sample dialogue to demonstrate the functioning of our framework. Finally, conclusions and future work directions are reported in Section 6.

2. Signs of Social Attitude

After several forms of 'anthropomorphic behavior' of users towards technologies were demonstrated [Reeves and Nass, 1996], various terms and concepts have been employed to denote this behavior and describe it. [Paiva, 2004] talks about *empathy*. [Hoorn and Konijn, 2003] address the concept of

engagement, involvement, *sympathy* and their contrary, *distance*. [Cassell and Bickmore, 2003] adopt the Svennevig's theory of *interpersonal relations*. We refer to Scherer's concept of *interpersonal stance* as a category which is *"characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in this situation (e.g. being polite, distant, cold, warm, supportive, contemptuous)".* In particular, in referring to the social response of users to ECAs, we distinguish warm/open from cold/close/distant *social attitude*, according to the Andersen and Guerrero's definition of interpersonal warmth as "*the pleasant, contented, intimate feeling that occurs during positive interactions with friends, family, colleagues and romantic partners*".

As described in details in [De Carolis et al, 2012], the multimodal indicators of social attitude, that we employ in our approach, concern signs deriving from linguistic, acoustic and gesture analysis.

The signs of social attitude in the linguistic part of the student communicative act are recognized according to the approach described in [Novielli et al, 2010]. In particular, we defined a taxonomy of signs for analyzing social communication in text-based interaction which employs *affective, cohesive and interactive indicators* (i.e. talking about self, expressing feelings, using a friendly style, expressing positive or negative comments, and so on). Then, the recognized linguistic signs are included in the multimodal framework described in this paper.

However, according to several studies [Litman et al, 2003; Sundberg et al, 2011], linguistic analysis is not enough to properly interpret the real user's communicative intention and his attitude towards an embodied agent. For instance, the user can pronounce the same sentence with different emotional attitudes in order to convey different communicative intents and show then a different attitude [Bosma and Andre', 2004; De Carolis and Cozzolongo, 2009].

Research in emotional speech has shown that acoustic and prosodic features can be extracted from the speech signal and used to develop models for recognising emotions and attitudes [Vogt et al, 2008; Sundberg et al, 2011]. In fact, the effects of emotion in speech tend to alter the pitch, timing, voice quality, and articulation of the speech signal, and reliable acoustic features can be extracted from speech that vary with the speaker's affective state. Then, in order to classify the social attitude of the user from speech, we decided to use prosodic features of the spoken utterance for recognising the valence of the sentence in terms of a *negative/cold* style vs. a *positive/friendly* one (in a five-points scale) and the arousal from *low* to *high* in a three-points scale. Recognising the value of only these two dimensions is justified since the valence indicates a failure/success in the achievement of the user's goal and, if it is related to the arousal, it allows to distinguish for instance a negative/cold attitude towards the agent from sadness related to a personal mental state. Therefore, a *positive* valence is a sign of positive feedback, comment, agreement towards the agent while a *negative* one indicates a disagreement or a negative comment. Moreover, a classification model based on simple features allows handling online analysis of the user's attitude [Vogt et al, 2008].

In order to classify the attitude of the user expressed through speech we used the corpus annotated in terms of valence and arousal collected in the previously mentioned project (see [De Carolis and Cozzolongo 2009] for more details). In parallel with the annotation process, the audio files relative to the moves in the corpus were analyzed using Praat functions [Boersma and Weenink, 2007] in order to perform a macro-prosodic or global analysis and to extract from the audio file of each move features related to:

- variation of the fundamental frequency (f0): pitch minimum, mean, maximum and standard deviation, slope;
- variation of energy (RMS): min, max and standard deviation.
- variation of harmonicity: min, max and standard deviation.
- Spectrum Central Moment, Standard Deviation, Gravity centre, Skeweness and Kurtosis.
- Speech rate

At present, our classifier exploits the NNge algorithm [Martin, 1995] and recognizes the valence with an accuracy of 89%, evaluated on a dataset of 4 speakers and 748 user moves overall, and validated using a *10 Fold Cross Validation technique.*

In order to endow an embodied agent with the ability of recognizing the social attitude also from gestures, according to the literature we considered those, involving arms and hands position. Arms are quite reliable indicators of mood and feeling, especially when interpreted with other signals. For example, crossed arms act as defensive barriers, indicating closure; using an arm across the body denotes nervousness or a negative attitude. Conversely, arms in open positions (especially combined with open palms) indicate feelings of openness and security. However, since we perform gesture recognition using Microsoft Kinect, we had to consider only a subset of gestures compatible with the nodes in the skeleton that the Kinect SDK is able to detect.

Hands also very expressive parts of the body as well, used a lot in signalling consciously - as with emphasizing gestures - or unconsciously - as in a wide range of unintentional movements that indicate otherwise hidden feelings and thoughts. Since at present Kinect skeleton does not includes nodes for detecting the position of fingers, we are able to recognize only simple hands gestures like hands picking nose, denoting social disconnection or stress, neck scratching, expressing doubt or disbelief, running hands through hair indicating vexation or exasperation.

Even if the position of the legs cannot be considered as a part of gesture, in evaluating the social attitude we take into account whether the legs are crossed or not, to support the corresponding arms signals (in conjunction with crossed arms they indicate a strong closure or rejection or insecurity).

3. Dynamic Modeling of the User Social Attitude

The user modeling procedure integrates (i) language analysis for linguistic cues extraction [Novielli et al, 2010], (ii) prosodic analysis and (iii) gesture recognition into a Dynamic Belief Network (DBN) [Jensen, 2001]. The DBN

formalism is particularly suitable for representing situations which gradually evolve from a dialog step to the next one since time slices (local belief networks) are connected through temporal links to constitute a full model. The DBN (Figure 1) is used to infer how the user's social attitude evolves during the dialog in relation to the dialog history according to signals expressed in the verbal and non-verbal part of the communication. Social attitude (SAtt) is the variable we want to monitor, which depends on *observable* ones, i.e. the recognized significative signals in the user move deriving from the linguistic, acoustic and gestural analysis. These nodes may correspond to a simple variable, as in the case of SAttVocal, or to a nested belief network as in the case of the SAttLing and SAttGesture whose probability distribution is calculated by the corresponding belief networks.

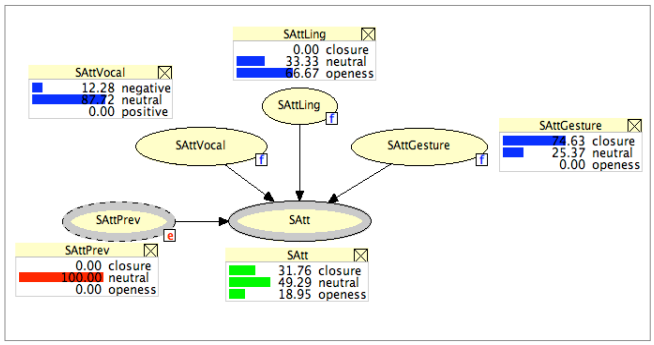

Fig.1 DBN modelling the user social attitude.

At the beginning of interaction, the model is initialized; at every dialog step, knowledge about the evidence produced by the multimodal analysis is entered and propagated in the network: the model revises the probabilities of the social attitude node. The new probabilities of the signs of social attitude are used for planning the next agent move, while the probability of the social attitude node supports revising high-level planning of the agent behavior.

As to voice analysis, our model classifies the spoken utterance according to the recognized valence and arousal. From these two values, using simple rules, we set the evidence of variable SAttVocal in the model. For instance, when the valence is very negative and the arousal is high the system sets the attitude to negative (anger).

Figure 2 shows a generic time-slice of the BN for modeling the user's social attitude from gestures. In particular, the gestures recognized by Kinect become the evidences of the root nodes of this model. These evidences are then propagated in the net and the probability of the SAttGesture node is computed given the probabilities of intermediate nodes, Hands, Arms and CrossedLegs, denoting the social attitude expressed by each of them. For instance the skeleton in Figure 3 is recognized as crossed_arms and corresponds to evidence for the node *crossed* in the model in Figure 2.

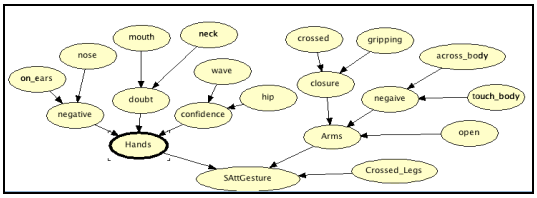

Fig. 2.User Model for the Social Attitude from Gestures, a generic time-slice

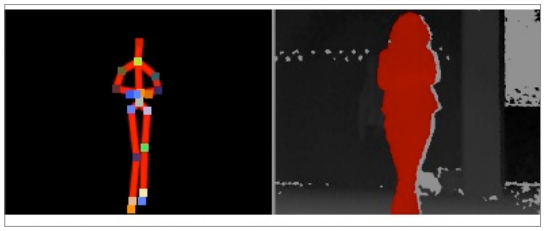

Fig. 3. KinectDTW recognition of crossed arms.

4. Collecting a corpus of interactions.

In order to perform an evaluation of the model, we start an experiment for collecting new dialogues for tuning the probabilities of our model.

As in [Clarizio et al, 2006], we performed a Wizard of Oz simulation study in which we collected multimodal dialog moves consisting of linguistic, acoustic and gesture data. Participants involved in the study were Italian students aged between 16 and 25, equally distributed by gender, consistently with the subjects involved in the previous experiment. They were divided in two groups, composed by 5 people each. We assigned to each group the same goal of information seeking:

Getting information about a correct diet in order to stay in shape.

To obtain this information subjects could dialogue with the PCA playing the role of an expert in nutrition. Before starting the experiments we administered to each subject a simple questionnaire aimed at collecting some personal data (age and gender) and at understanding their background (department, year of course, Artificial Intelligence background). Subjects were told they had to both provide a final evaluation of the agent and to answer a series of questions about the degree of recalling, to test the effectiveness of the interaction with the PCA with respect to both the engagement in the interaction and the information provision task. When finished, the subject had to compile a questionnaire about the topic.

Following the described approach we collected a corpus of about 300 moves. Each move was recorded using a wireless microphone whose output was sent to the speech processing system. We assigned to each utterance a unique identifier, the correspondent transcript and the related 16-bit single channel, 8 kHz signal (in .wav format). For recording gestures, we used Microsfot Kinect and Dynamic Time Warping (DTW) [KinectDTW 2011] for the recognition of signals concerning arms, hands and legs.

Each move was annotated by a human rater (three researchers in the field of human-computer interaction in total) with respect to the perceived user social attitude, conveyed by the turn. Each move received a final label for the social attitude using a majority agreement criterion. Then we automatically annotated the single moves using the modules defined for the language and speech analysis and the DTW for the gestures. These evidences constitute the features describing each user move, according to the signals variables in the previously described models. According to the result of the annotation experiments (both manual and automatic) we conducted a preliminary evaluation for tuning the structure and the probabilities of our multimodal framework, shown in Figure 1.

4. An Example of Dialogue

In this section we show a sample dialogue (translated from Italian) from our corpus and the recognized social attitude for significant moves. In this dialogue the agent and user moves are denoted respectively by Vi (generated by our dialogue simulator) and Uj (performed by the user). Where no annotation is provided, our models did not recognize any sign of social attitude.

V0: Hi, my name is Valentina! I'm here to give you information about correct eating habits. What's your name?
U0: OK, my name is Carla. (positive valence, medium arousal)
V1: Hi Carla, nice to meet you. How can I help you?
U1: I would like to learn more about a correct diet! (positive valence, medium arousal - open arms)
V2: Good. Let's start with a pratical example. What did you eat for breakfast this morning?
U2: Coffee and some cookies.
V3: And for lunch?
U3: Just a sandwich
V4: What did you eat for dinner yesterday night?
U4: Meat and salad.
V5: Good. Your diet is not so bad. However, like you, most of people forget about the importance of eating more fruit and vegetables. International Research about foods proves that fruit and vegetables are very important in a healthy diet. They provide fibre and vitamins that are very important …(benefits of vitamins and fibers). Now, try to think again about your diet, is it healthy?
U5: Well, I should change something.
V6: Good, let's try to do something then. Did you ever think to follow a correct diet?
U6: Yes, sometimes I did it and I did not get very good results (negative valence, medium arousal).
V7: Why?
U7: I don't know (negative valence, low arousal - crossed arms).
V8: Could you tell me why, so that I can help you?
U8: I do not want to answer (crossed arms, crossed legs, negative valence, medium arousal)
V9: OK I understand that sometimes talk about self can be difficult….

In the present prototype, linguistic cues of Friendly Style and Talks about Self are detected and evidence about these signals contributes to increase the overall likelihood of observing a warm social attitude of the user from the

linguistic point of view. Valence and arousal are detected from acoustic parameters using a classifier and gestures are detected using Kinect and a library for recognition of signals related to gestures based on DTW are given as evidences to the model for recognising a more closed or open attitude.

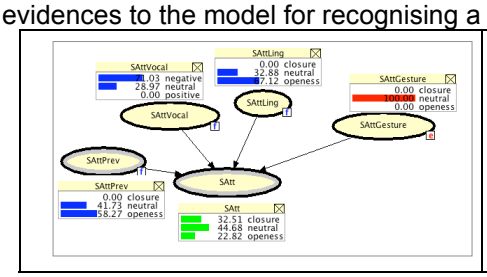

Fig 4. Social Attitude Recognition for Move U6.

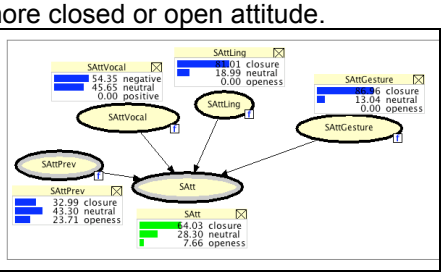

Fig 5. Social Attitude Recognition for Move U7.

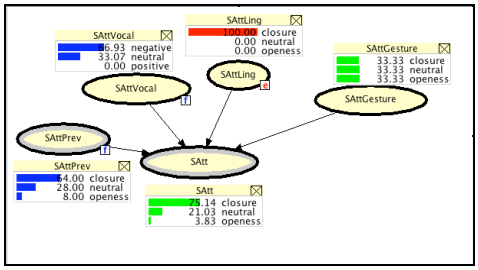

Fig. 6. Social Attitude Recognition for Move U8.

For instance, in move U6, linguistic analysis provides an evidence of Talk_about_Self to allow the recognition of a warm social attitude. However, acoustic analysis classifies the valence as negative and the arousal as medium, thus denoting a negative/cold attitude. Figure 4 shows the model results in this case. Then the agent, in move V7, asks which is the reason of this result. In the next move U7 the user says that she does not want to answer and crosses her arms. These signals provide evidences of a negative/cold social attitude (Figure 5). The subsequent move V8, in which the agent keeps asking for the reason, causes a further closure since the user move, U8, is recognized linguistically as a Negative Comment, Acoustically as a negative valence and from gesture and legs position as a strong closure (Figure 6).

4.Conclusions

This research builds on prior work on affect modeling and dialog simulation. In this paper we enrich a model for recognizing social attitude with the analysis of signals regarding non-verbal communication: prosody and gesture in particular. The two approaches to social attitude modeling using speech and language analysis have been validated in our previous research, with satisfying results. Here we have proposed an extension of the multimodal analysis to gesture modeling, according to the meanings that psycholinguistic researchers attach to gestures in conversations. We plan to improve gesture recognition since the new Kinect should allow for a better hand recognition. Moreover, we

will perform more evaluation studies in order to test the robustness of our framework for social attitude recognition in different scenarios and with respect to different interaction modalities with both ECAs and Robots.

Bibliography

[Andersen and Guerrero, 1998] Andersen, P.A. and Guerrero, L.K., 1998. Handbook of Communication and Emotions. Research, theory, applications and contexts. Academic Press.

[Baylor and Kim, 2005] Baylor, A., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. International Journal of Artificial Intelligence in Education, 15(1), 95–115.

[Bickmore, 2003] Bickmore, T. (2003). Relational agents: Effecting change through human-computer relationships. PhD Thesis, Media Arts & Sciences, Massachusetts Institute of Technology.

[Boersma and Weenink, 2007] Boersma, P. and Weenink, D. (2007). Praat: doing phonetics by computer (version 4.5.15) [computer program]. http://www.praat.org/. Retrieved 24.02.2007.

[Bosma and André, 2004] W.E. Bosma and E. André, “Exploiting Emotions to Disambiguate Dialogue Acts”, in Proc. 2004 Conference on Intelligent User Interfaces, January 13 2004, N.J. Nunes and C. Rich (eds), Funchal, Portugal, pp. 85-92, 2004.

[Cassell, 2001] Cassell, J. (2001) "Embodied Conversational Agents: Representation and Intelligence in User Interface" AI Magazine, Winter 2001, 22(3): 67-83.

[Cassell and Bickmore, 2003] Cassell, J. and Bickmore, T., 2003. Negotiated collusion: modelling social language and its relationship effects in intelligent agents. User Modelling and User-Adapted Interaction, 13, 1-2, 2003.

[Clarizio et al, 2006] Clarizio, G., Mazzotta, I., Novielli, N. and de Rosis, F., 2006. Social Attitude Towards a Conversational Character. In Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication. RO-MAN 2006, 2-7.

[D'Mello et al, 2008] D'Mello, S. K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., and Graesser, A. C. (2008). Automatic Detection of Learner’s Affect from Conversational Cues. User Modeling and User-Adapted Interaction, 18(1-2), 45-80.

[De Carolis and Cozzolongo, 2009] De Carolis B., Cozzolongo G., 2009. Interpretation of User’s Feedback in Human-Robot Interaction. Journal of Physical Agents, 2, 2.

[De Carolis et al, 2012] De Carolis B., Ferilli S., Novielli N. Towards a Model for Recognising the Social Attitude in Natural Interaction with Embodied Agents. In Proceedings of the 5th International Workshop on Intelligent Interfaces for Human-Computer Interaction. Palermo 2012.

[Hoorn and Konijn, 2003] J.F. Hoorn, and E.A. Konijn, Perceiving and Experiencing Fictitional Characters: An integrative account. Japanese Psychological Research, 45, 4, 2003.

[MK, 2011] http:// kinectforwindows.org

[Jensen, 2001] Jensen, F.V., 2001. Bayesian Networks and Decision Graphs. Springer

[Johnson et al, 2000] Johnson, W., Rickel, J., Lester, J., 2000. Animated pedagogical agents: face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education, 11, 47–78

[Kim and Baylor, 2008] Kim, C. & Baylor, A.L. (2008). A Virtual Change Agent: Motivating Pre-service Teachers to Integrate Technology in Their Future Classrooms. Educational Technology & Society, 12(2), 309-321.

[KinectDTW 2011] KinectDTW: http://kinectdtw.codeplex.com/

[Litman et al, 2003] D. Litman, K. Forbes, S. Silliman, “Towards emotion prediction in spoken tutoring dialogues”. Proceedings of HLT/NAACL, 2003.

[Martin, 1995] Brent Martin, (1995) "Instance-Based learning : Nearest Neighbor With Generalization", Master Thesis, University of Waikato, Hamilton, New Zealand

[Moreno et al, 2001] R. Moreno, R. E. Mayer, H. Spires and J. Lester, The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?, Cognition and Instruction, 19, pp. 177-213, 2001.

[Nakamura, 2003] S. Nakamura, Toward Heart-to-Heart Speech Interface with Robots. ATR UptoDate. Summer 2003.

[Novielli et al, 2010] N. Novielli, F. de Rosis and I. Mazzotta, User Attitude Towards an Embodied Conversational Agent: Effects of the Interaction Mode, in Journal of Pragmatics, Volume 42, issue 9 (September, 2010), p. 2385-2397. Elsevier Science.

[Paiva, 2004] A. Paiva, (Ed): Empathic Agents. Workshop in conjunction with AAMAS’04. 2004.

[Pease, 2006] Pease A and B.. 2006. The Definitive Book of Body Language. Bantam Books.

[Polhemus et al, 2001] Polhemus, L., Shih, L-F and Swan, K., 2001. Virtual interactivity: the representation of social presence in an on line discussion. Annual Meeting of the American Educational Research Association.

[Reeves and Nass, 1996] Reeves, B., and Nass, C. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge: Cambridge University Press.

[Swan, 2002] Swan, K., 2002. Immediacy, social presence and asynchronous discussion, in: J. Bourne and J.C.Moore (Eds.): Elements of quality online education. Vol3, Nedham, MA. Sloan Center For Online Education.

[Sundberg et al, 2011] Sundberg, J., Patel, S., Björkner, E., & Scherer, K.R. (2011). Interdependencies among voice source parameters in emotional speech. IEEE Transactions on Affective Computing, 2(3),

[Vogt et al, 2008] Vogt,T, Andre’ E., and Bee N. 2008.EmoVoice- A Framework for Online Recognition of Emotions from Voice.In Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems,Springer-Verlag, Heidelberg,188-199.