A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances

Mathieu Chollet, Magalie Ochs, Catherine Pelachaud

▶ **To cite this version:**

Mathieu Chollet, Magalie Ochs, Catherine Pelachaud. A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances. IVA 2013 Workshop Multimodal Corpora: Beyond Audio and Video, Sep 2013, Edinburgh, United Kingdom. hal-01074865

HAL Id: hal-01074865
https://hal.science/hal-01074865

Submitted on 15 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances

Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud

Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI
{mathieu.chollet, magalie.ochs, catherine.pelachaud}@telecom-paristech.fr

Abstract. In order to give the capability to a virtual recruiter to convey interpersonal stances through its non-verbal behavior, we propose to use a multimodal corpus approach using videos of job recruitment enactments. The corpus has been annotated at different levels to consider both the participants' non-verbal behavior and the expressed stances. In this article, we present this corpus and our annotations, and discuss issues related to the specificity of the acquired data.

1 Introduction

In the TARDIS project[1], we aim to develop an ECA that acts as a virtual recruiter to train youngsters to improve their social skills. For Scherer [1], interpersonal stances are *"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 that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous)."* A virtual recruiter should be able to convey different interpersonal stances: our goal is to find out how interpersonal stance is expressed through non-verbal behavior, and to implement the expression of interpersonal stance in an ECA. As a representation for interpersonal stance, we use Argyle's attitude dimensions [2], *friendliness* (also called warmth or affiliation) and *dominance* (also called agency). Given an interpersonal stance it should express, i.e. a dominance value and a friendliness value, such an ECA should be able to adapt its behavior to appear with this stance.

Research in behavioral sciences has shown that non-verbal behavior[2] are direct cues of interpersonal stance, for instance a smile is a sign of friendliness [4]. The combinations and sequencing of these signals is also significant: for example, the sequencing of smile, gaze and head aversion can differentiate between amusement, shame and embarassment, therefore expressing different values of dominance [5]. Inter-personal dynamics of behavior also carry meaning: adopting the same postures as your interlocutor can make you seem more friendly towards them [6]. Finally, the global behavior tendencies of a person contribute

[1] http://tardis.lip6.fr/

[2] Non-verbal behavior corresponds to "facial expressions, body language, social touching, vocal acoustics, and interpersonal distance" [3]. Non-verbal behavior may be expressed intentionally (e.g. symbolic gestures) or non-intentionally (e.g. expression of felt emotion).

to the perception of stance: for instance, a high body energy throughout an interaction is correlated with dominance [7]. An agent expressing interpersonal stances should take into account these three different aspects when choosing its non-verbal behavior: non-verbal signals, inter-personal behavior dynamics, and global behavior tendencies. However, though an extensive number of works have studied them [4–7], these studies have mostly been carried out independently of the other aspects.

In the ECA community, Ballin *et al.* [8] proposed a model for automatic posture generation based on interpersonal attitude parameters, but it was only used for agent-agent interactions and did not include all non-verbal modalities. The Laura agent [9] was used to develop long term relationships with users using a menu-based interface, and would adapt the frequency of gestures and facial signals as the relationship with the user grew. However, dominance was not investigated. Prepin *et al.* [10] have investigated how smile alignment and synchronisation can contribute to stance building in a dyad of agents. Although not directly related to dominance or friendliness, Sensitive Artificial Listeners designed in the Semaine project [11] produce feedback and backchannels depending of the personality of an agent, defined by extraversion and emotional stability. These works, even though they don't use full body movement and are not fully interactive, show that single non-verbal signals, inter-personal behavior dynamics and global behavior tendencies of interpersonal stance may convey different ECA's interpersonal stances. These works [8, 9, 11, 12] used hand-coded rules based on literature from behavioral sciences. Another method for designing such agent behavior rules is to extract them from corpora of interactions annotated with non-verbal behavior and interpersonal stance. Multimodal corpora have already been used for tasks such as gesture generation [13] or the analysis of feedback [14]. Automatic dominance in small group meetings has been well studied in recent years, however they mostly rely on global behavioral features (e.g. total visual energy), and have mostly considered the problem of classifying the most/least dominant person in a meeting [7, 15].

In our work, we propose to annotate a corpus of videos of job interviews with non-verbal behavior and interpersonal stance annotations. The objective is to analyse the three different aspects of non-verbal behavior presented earlier (signals, inter-personal dynamics of behavior and global behavior tendencies). In the next section, we present the chosen corpus and the annotation process.

2 Multimodal corpus presentation

In this section, we present our corpus, the coding scheme we used for the annotation of multimodal behavior and *interaction state*, and the process of interpersonal stance annotation.

2.1 Corpus description

As part of the TARDIS project, a study was conducted with practitioners and youngsters from the Mission Locale (a French national association organizing job

interview coaching for youngsters in search for a job). The study consisted in creating a situation of job interviews between 5 practitioners and 9 youngsters. The setting was the same in all videos. The recruiter and the youngster sat on each side of a table. A single camera embracing the whole scene recorded the dyad from the side. This resulted in a corpus of 9 videos of job interview lasting between 15 and 20 minutes each. We discarded 4 videos as the recruiter was not visible due to bad position of the camera. Out of the 5 remaining videos, we have so far annotated 3, for a total of 50 minutes and 10 seconds of video.

2.2 Corpus annotation

The job interviews were annotated on three aspects: the *interaction state*, the *non-verbal behavior* of the recruiter and youngster, and the *interpersonal stance* of the recruiter.

Interaction state annotation - When analysing non-verbal behavior, it is important to know the task being undergone and the turn-taking state. For instance, gaze behaviors in a dyadic interaction change when objects related to the task being undergone are present [16], and the timing of posture shifts is related to conversational turn-taking [17]. We annotated the task state to know if the subject being discussed involved a document or not, and we annotated the turn-taking state to know which interactant is speaking or whether there is an interruption. We refer to these annotations as the interaction state.

Non-verbal behavior annotation - We use the MUMIN multimodal coding scheme [14] and adapt it by removing any types of annotations we cannot extract from the videos (i.e. subtle facial expressions), and selecting the modalities that are implied in the expression of interpersonal stances. We use Praat [18] for the annotation of the audio stream and the Elan annotation tool [19] for the visual annotations. We report here on the annotation of the recruiters' behavior: the youngsters' behavior was annotated by colleagues of the TARDIS consortium using a similar coding scheme.

We include gaze and head movements in the coding scheme, as they are related to interpersonal stance [4, 20]. Because of the camera-dyad distance, we do not try to annotate very complex facial expressions (e.g. action units for facial movements), however we include smiles and eyebrow movements (raised and frown). For the rest of the body, we consider posture and gestures, two important social cues [4, 21]. For gestures, we include object manipulations and adaptor gestures as they can convey nervousness [21]. We also include several hands rest positions (e.g. arms crossed). The para-verbal tags are used to differentiate when a participant is speaking, is silent, laughing. To this category, we add annotations of participants using hesitation words (e.g. "err" or "hmm").

The 3 videos were fully annotated for the recruiter behavior by a single annotator. A month after the annotation process, a double annotation was performed by the same annotator on a video segment of 3 minutes. We computed Cohen's kappa score on this segment for the different annotation tiers to measure the consistency of the coding. It was found to be satisfactory for all modalities ($\kappa >= 0.70$), except for the eyebrow movements ($\kappa = 0.62$), which low score

can be explained by the high camera-dyad distance making detection difficult. The highest scores were for gaze ($\kappa = 0.95$), posture ($\kappa = 0.93$) and gestures ($\kappa = 0.80$). This annotation processes amounted to 8012 annotations for the 3 videos. The para-verbal category has the highest count of annotations, between 483 to 1088 per video. On non-verbal annotations, there were 836 annotations of gaze direction, 658 head directions, 313 gestures, 281 head movements, 245 hands positions, 156 eyebrow movements and 91 smiles.

Important differences in behavior tendencies exist between recruiters: for instance the first recruiter performed many posture shifts: 5.6 per minute, to compare with 2.2 for the second recruiter and 0.6 for the third one. The second recruiter smiles much less than the others: 0.4 smiles per minute versus 2.4 per minute for both the first and third recruiters.

Interpersonal stance annotation - As the interpersonal stance of the recruiters varies through the videos, we chose to use GTrace, successor to FeelTrace [22]. GTrace is a tool that allows for the annotation of continuous dimensions over time. Users have control over a cursor displayed on an appropriate scale alongside a playing video. The position of the cursor is sampled over time, and the resulting sequence of cursor positions is known as *trace data*. We asked 12 persons to annotate the videos. Each annotator had the task of annotating one dimension for one video. To reduce the influence of the actual dialogue content, we filtered the audio tracks to make speech unintelligible. For now, we have collected two annotation files per dimension per video, and the annotation process is still ongoing.

Trace data of socio-emotional dimensions is prone to a certain number of issues, such as scaling, noise, and reliability [22]. Therefore, trace data requires specific methods of pre-treatment and analysis, and classical methods for reliability may not be adequate. A method proposed by Cowie and McKeown [22] is to use stylisation techniques to replace the trace data into three kinds of elements: plateaus, rises and falls. A plateau is a time interval where the annotation of friendliness or dominance remains constant. Rises and falls are intervals when the annotator is moving the cursor, i.e. annotating that he perceives a change in the dominance or friendliness of the recruiter. Annotators might not use the scale in the same way, however they might agree on when a person appears more or less dominant or friendly.

Table 1. Plateaus and slopes average counts, values, durations

Video	Dimension	Plateaus count	Mean plateau value [-1, 1]	Mean plateau duration	Slopes count	Mean slopes value [-1, 1]	Mean slopes duration
1	Dom.	41	0,41	7.9s	40	0.21	7.6s
1	Frd.	46	0.08	7,1s	45	0.25	5.2s
2	Dom.	28	0,19	20,4s	27	0.19	11.9s
2	Frd.	26	-0.33	11.3s	25	0.22	6.5s
3	Dom.	51	0.13	9.8s	50	0.24	9.2s
3	Frd.	47	0.06	6.6s	46	0.29	8.7s

A first analysis of the annotations is presented in Table 1. To reduce noise, we smoothed the data by removing slopes smaller than a twentieth of the scale and removing plateaus shorter than one second, leaving 120 dominance plateaus, 170 dominance slopes, 119 friendliness plateaus and 116 friendliness slopes. Important differences exist between the videos: from the mean values of plateaus, recruiter 1 appears to be very dominant and recruiter 2 very unfriendly. Plateaus and slopes for dominance seem to have a higher duration than for friendliness, which might suggest dominance changes more slowly.

3 Conclusion

In this paper, we presented our ongoing work on the analysis of the relationship between interpersonal stance and non-verbal behavior for the implementation of virtual recruiters. We detailed the process of non-verbal behavior, interaction state, and stance annotation. We discussed problems specific to trace data, such as scaling issues between annotators.

We believe we can use the information on when the perceived interpersonal stance varies or when it remains unchanged. This analysis will require specific techniques, such as those described in [23], to detect possible sequences in intra-personal and inter-personal behavior. Once those sequences are extracted, we will implement them in ECAs and run a perceptive study to validate if users perceive interpersonal stance variations or not. We are also in the process of gathering more interpersonal stance annotation data, in order to have more reliable data and to study inter-gender variations.

Acknowledgment This research has been partially supported by the European Community Seventh Framework Program (FP7/2007-2013), under grant agreement no. 288578 (TARDIS).

References

1. Scherer, K.R.: What are emotions? and how can they be measured? Social Science Information **44** (2005) 695–729
2. Argyle, M.: Bodily Communication. University paperbacks. Methuen (1988)
3. Ambady, N., Weisbuch, M.: Nonverbal behavior. In: Handbook of Social Psychology, John Wiley & Sons, Inc. (2010)
4. Burgoon, J.K., Buller, D.B., Hale, J.L., de Turck, M.A.: Relational Messages Associated with Nonverbal Behaviors. Human Communication Research **10**(3) (1984) 351–378
5. Keltner, D.: Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Journal of Personality and Social Psychology **68** (1995) 441–454
6. LaFrance, M.: Posture mirroring and rapport. In Davis, M., ed.: Interaction Rhythms: Periodicity in Communicative Behavior, New York: Human Sciences Press (1982) 279–299
7. Escalera, S., Pujol, O., Radeva, P., Vitria, J., Anguera, M.: Automatic detection of dominance and expected interest. EURASIP Journal on Advances in Signal Processing **2010**(1) (2010) 12

8. Ballin, Daniel, G.M., Crabtree, B.: A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In: 1st European Conference on Visual Media Production. (2004)
9. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Trans. Comput.-Hum. Interact. **12**(2) (June 2005) 293–327
10. Prepin, K., Ochs, M., Pelachaud, C.: Mutual stance building in dyad of virtual agents: Smile alignment and synchronisation. In: SocialCom/PASSAT. (2012) 938–943
11. Bevacqua, E., Sevin, E., Hyniewska, S., Pelachaud, C.: A listener model: introducing personality traits. Journal on Multimodal User Interfaces **6**(1-2) (2012) 27–38
12. Prepin, K., Ochs, M., Pelachaud, C.: Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents. In: International Conference CogSci (Annual Conference of the Cognitive Science Society). (July 2013)
13. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gestures: How to economically capture timing and form. Journal on Language Resources and Evaluation - Special Issue on Multimodal Corpora **41**(3-4) (2007) 325–339
14. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation **41**(3-4) (2007) 273–287
15. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: A review. Image Vision Comput. **27**(12) (2009) 1775–1787
16. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press (1976)
17. Cassell, J., Nakano, Y.I., Bickmore, T.W., Sidner, C.L., Rich, C.: Non-verbal cues for discourse structure. In: Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics. (2001) 106–115
18. Boersma, P., Weenink, D.: Praat, a system for doing phonetics by computer. Glot International **5**(9/10) (2001) 341–345
19. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: In Proceedings of Language Resources and Evaluation Conference (LREC). (2006)
20. Mignault, A., Chaudhuri, A.: The many faces of a neutral face: Head tilt and perception of dominance and emotion. Journal of Nonverbal Behavior **27**(2) (2003) 111–132
21. Carney, D.R., Hall, J.A., LeBeau, L.S.: Beliefs about the nonverbal expression of social power. Journal of Nonverbal Behavior **29**(2) (2005) 105–123
22. Cowie, R., McKeown, G.: Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme, SEMAINE D6b deliverable (2010)
23. Allwood, J., Kopp, S., Grammer, K., Ahlsen, E., Oberzaucher, E., Koppensteiner, M.: The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. Language Resources and Evaluation **41** (2007) 255–272