Investigating Non-Verbal Behaviors Conveying Interpersonal Stances
Mathieu Chollet, Magalie Ochs, Catherine Pelachaud

To cite this version:
Mathieu Chollet, Magalie Ochs, Catherine Pelachaud. Investigating Non-Verbal Behaviors Conveying Interpersonal Stances. First European Symposium on Multimodal Communication, Sep 2013, La Valetta, Malta. <hal-01074877>

HAL Id: hal-01074877
https://hal.archives-ouvertes.fr/hal-01074877
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.
Investigating Non-Verbal Behaviors Conveying Interpersonal Stances

Mathieu Chollet¹, Magalie Ochs² and Catherine Pelachaud²

¹Institut Telecom; Telecom Paristech; CNRS-LTCI
²CNRS LTCI; Telecom ParisTech
37 rue Dareau, 75014, Paris, France
{mchollet,ochs,pelachaud}@telecom-paristech.fr

Abstract

Interpersonal stances are expressed by non-verbal behaviors on a variety of different modalities. The perception of these behaviors is influenced by the context of the interaction, how they are sequenced with other behaviors from the same person and behaviors from other interactants. In this paper, we introduce a framework considering the expressions of stances on different layers during an interaction. This framework enables one to reason on the non-verbal signals that an Embodied Conversational Agent should express to convey different stances. To identify more precisely humans’ non-verbal signals conveying dominance and friendliness attitudes, we propose in this paper a methodology to automatically extract the sequences of non-verbal signals conveying stances. The methodology is illustrated on an annotated corpus of job interviews.

Keywords

Interpersonal stance; Non-verbal behaviors; Sequence mining

1 Introduction

Embodied Conversational Agents (ECAs) are increasingly used in training and serious games. In the TARDIS project¹, we aim to develop an ECA that acts as a virtual recruiter to train youngsters to improve their social skills. Such a virtual recruiter should be able to convey different interpersonal stances, that can be defined as “spontaneous or strategically employed affective styles that colour interpersonal exchanges (Scherer, 2005)”. Our goal is to find out how interpersonal stances are expressed through non-verbal behavior, and to implement the expression of interpersonal stances in an ECA.

Most modalities of the body are involved when conveying interpersonal stances (Burgoon et al., 1984). Smiles can be signs of friendliness (Burgoon et al., 1984), performing large gestures may be a sign of dominance, and a head directed upwards can be interpreted with a dominant stance (Carney et al., 2005). A common representation for interpersonal stance is Argyle’s bi-dimensional model of attitudes (Argyle, 1988), with an affiliation dimension ranging from hostile to friendly, and a status dimension ranging from submissive to dominant (see Figure 1).

Figure 1: The Interpersonal Circumplex, with Argyle’s attitude dimensions. The sample coordinate represents a friendly and slightly dominant interpersonal stance.

A challenge when interpreting non-verbal behavior is that every non-verbal signal can be interpreted with different perspectives: for instance, a smile is a sign of friendliness (Burgoon et al., 1984); however, a smile followed by a gaze and head aversion conveys embarrassment (Keltner, 1995). Non-verbal signals of a person in an interaction should also be put in perspective to non-verbal signals of the other participants of the interaction: an example is posture mimicry, which

¹http://http://www.tardis-project.eu/
can convey friendliness (LaFrance, 1982). Finally, the global behavior tendencies of a person, such as performing large gestures in general, are important when interpreting their stance (Escalera et al., 2010). These different perspectives have seldom been studied together, and this motivates the use of multimodal corpora of interpersonal interactions in order to analyze their impact in a systematic fashion.

We propose a model for non-verbal behavior analysis, composed of multiple layers analyzing a particular perspective of non-verbal behavior interpretation on time windows of different lengths. To build this model, we annotated a corpus of job interview enactment videos with non-verbal behavior annotations and interpersonal stance annotations. In this paper, we focus on a layer of the model which deals with how sequences of non-verbal signals displayed while speaking can be interpreted as the expression of dominance and friendliness stances. While it has been proved that the sequencing of non-verbal signals influences how they are perceived (With and Kaiser, 2011), the literature on the topic is limited. To gather knowledge about this layer, we use a data mining technique to extract sequences of non-verbal signals from the corpus.

The paper is organized as follows. In section 2, we present related models of interpersonal stances for ECAs and their limits. We then introduce our multi-layer model. Section 4 describes the multimodal corpus and how it was annotated. Section 5 details a data mining method we propose to gather knowledge about how sequences of non-verbal behavior are perceived.

2 Related work

Models of interpersonal stances expression for virtual agents have already been proposed. For instance, in (Ballin et al., 2004), postures corresponding to a given stance were automatically generated for a dyad of agents. Lee and Marsella used Argyle’s attitude dimensions, along with other factors such as conversational roles and communicative acts, to analyze and model behaviors of side participants and bystanders (Lee and Marsella, 2011). Cafaro et al. (Cafaro et al., 2012) conducted a study on how smile, gaze and proximity cues displayed by an agent influence the first impressions that the users form on the agent’s interpersonal stance and personality. Ravenet et al. (Ravenet et al., 2013) proposed a user-created corpus-based methodology for choosing the behaviors of an agent conveying a stance along with a communicative intention. These models, however, only consider the expression of a few signals at a given time, and do not consider longer time spans or sequencing of signals.

Other works have gone further by also considering global behavior tendencies and reactions to the interactants’ behaviors: the Laura agent (Bickmore and Picard, 2005) was used to develop long term relationships with users, and would adapt the frequency of gestures and facial signals as the relationship with the user grew. However, dominance was not investigated, and the users’ behaviors were not taken into account as they used a menu-based interface. Prepin et al. (Prepin et al., 2013) 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 (Bevacqua et al., 2012) produce feedback and backchannels depending of the personality of an agent, defined by extraversion and emotional stability.

Even though different perspectives of interpretation of non-verbal behavior we mentioned have been integrated in models of ECAs, the existing models of interpersonal stances expression consider only consider one perspective at a time, with a limited number of modalities. Moreover, no model of stance expression seems to consider how non-verbal signals are sequenced. In the next section, we present a theoretical model to the integration of these different perspectives.

3 A multi-layer approach to the expression of interpersonal stances

In (Chollet et al., 2012), we defined a multi-layer model to encompass the different non-verbal behavior interpretation perspectives (See figure 2). The Signal layer looks at the interpretation of signals in terms of communicative intentions (e.g. a hand wave means greeting someone). In the Sentence layer, we analyze the sequence of signals happening in a dialogue turn (e.g. a smile followed by a head aversion means embarrassment). The Topic layer focuses on the inter-personal behavior patterns and tendencies (e.g. adopting the
same posture as the interlocutor is a sign of friendliness). Finally, the Interaction layer encompasses the whole interaction and looks at global behavior tendencies (e.g. smiling often is a sign of friendliness). These different layers allow to interpret interactants' interpersonal stances at every instant of the interaction, taking into account their behavior, their reactions to other interactants’ behaviors, and their global behavior tendencies.

In order to build a model for each layer, our approach consists of automatically extracting knowledge from a multimodal corpus of interactions during which interpersonal stances are expressed. In this paper, we focus on the Sentence layer: it is known that the sequencing of non-verbal signals influence how these behaviors are perceived (With and Kaiser, 2011), however since relatively little accounts exist on this phenomenon, automated methods of knowledge extraction are particularly relevant for this layer. In the next section, we present our multimodal corpus and its annotation process.

4 Multimodal corpus of interpersonal stance expression

As part of the TARDIS project, a study was conducted with practitioners and youngsters from the Mission Locale Val d’Oise Est, a French job coaching association. The study consisted in creating a situation of job interviews between 5 practitioners and 9 youngsters. The setting was the same in all videos (see Figure 3). 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. From this study was gathered a corpus of 9 videos of job interview lasting approximately 20 minutes each. We decided to use these videos to investigate the sequences of non-verbal signals the recruiters use when conveying interpersonal stances. In order to study how recruiters express interpersonal stances, we annotated three videos of job interview enactments, for a total of slightly more than 50 minutes. We consider full body non-verbal behavior, turn-taking, task and interpersonal stance.

Numerous coding schemes exist to annotate non-verbal behavior in multimodal corpora. A widely used system for facial expressions is the Facial Action Coding System (Ekman and Friesen, 1977). A very exhaustive coding scheme for multimodal behavior is the MUMIN multimodal coding scheme, that was used for the analysis of turn-taking and feedback mechanisms (Allwood et al., 2007). For the non-verbal behavior annotation, we adapted the MUMIN multimodal coding scheme.
to our task and our corpus (e.g. by removing any types of annotations we cannot extract from the videos, such as subtle facial expressions). We used Praat (Boersma and Weenink, 2001) for the annotation of the audio stream and the Elan annotation tool (Wittenburg et al., 2006) for the visual annotations. A single annotator annotated the three videos. To measure the reliability of the coding, three minutes of video were randomly chosen and annotated a second time one month after the first annotation effort, and we computed Cohen’s kappa score between the two annotations. It was found to be satisfactory for all modalities ($\kappa \geq 0.70$), except for the eyebrow movements ($\kappa \geq 0.62$), which low score can be explained by the high camera-dyad distance making detection difficult. The highest scores were for gaze ($\kappa \geq 0.95$), posture ($\kappa \geq 0.93$) and gestures ($\kappa \geq 0.80$). 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.

As the interpersonal stance of the recruiters varies through the videos, we chose to use GTrace, successor to FeelTrace (Cowie et al., 2011). 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 adapted the software for the interpersonal stance dimensions we considered. Though the software allows for the annotation of two dimensions at a time using a bi-dimensional space, we constrained it to a single dimension to make the annotation task slightly easier. We asked 12 persons to annotate the videos. Each annotator had the task of annotating one dimension for one video, though some volunteered to annotate more videos. As the videos are quite long, we allowed them to pause whenever they felt the need to. With this process, we collected two to three annotation files per attitude dimension per video. While evaluating inter-rater agreement is a simple task when analyzing discrete labels (e.g. two people assign the same class to an item), it is not as straight-
forward when dealing with trace data (Metallinou and Narayanan, 2013), though recently new approaches to this problem have been proposed (Cowie and McKeown, 2010). Similarly to previous experiments on trace data annotation of emotions, we found that raters agreed more on how attitude values varied (i.e. when attitudes raise or falls), than on actual absolute values.

Similarly to (Cowie and McKeown, 2010), we averaged attitude values in bins of 3 seconds. We then computed the reliability of different annotations by computing Cronbach’s $\alpha$, using the variation values from one bin to the next. Cronbach’s $\alpha$ value was found to be generally average ($\alpha = 0.489$), with the highest video scoring $\alpha = 0.646$. We believe these values to be acceptable for our purposes, considering Cronbach’s $\alpha$ is likely to produce lower scores on annotations continuous both in time and in value, and that the sequence mining process we propose (described in Section 5) provides a natural way of discarding the time segments where annotators were not agreeing. Indeed, the non-verbal signals sequences contained in these segments will be distributed for different types of attitude variations, and therefore will not be very frequent before any particular attitude variation. However, the sequence mining algorithm we use relies on frequency to extract meaningful non-verbal signals sequence, which means that the time segments where annotators do not agree will not contribute to making some non-verbal signals sequences more frequent.

In a nutshell, the corpus has been annotated at two levels: the non-verbal behavior of the recruiters and their expressed stances. Our next step was to identify the correlations between the non-verbal behaviors and the interpersonal stances. As a first step, we have focused on the non-verbal signals sequences expressed by the recruiters when they are speaking (i.e. at the Sentence level, Section 3). In the next section, we describe a method for extracting knowledge about non-verbal behavior sequences from the multimodal corpora.

5 Investigating non-verbal behavior sequences

A number of tools and techniques exist for the systematic analysis of sequences of events in sequential data. Traditional sequence analysis (Bakeman and Quera, 2011) techniques typically revolve around the computation of simple contingency tables measuring the occurrence of one type event of event after another one. Such methods are not well suited to longer sequences of events (i.e. made of more than 2 events) and to cases where noise can happen (i.e. behaviors irrelevant to a particular sequence that can happen in the middle of it). Magnusson proposed the concept of *T-patterns* (Magnusson, 2000), sequences of events occurring in the same order with “relatively invariant” temporal patterns between events. The THEME software automatically detects *T-patterns* and was used in (With and Kaiser, 2011) to detect characteristic sequences of signals for emotion expression. Finally, *sequence mining* techniques have been widely used in task such as protein classification (Ferreira and Azevedo, 2005), and recent work has used this technique to find sequences correlated with video game players’ emotions such as frustration (Martínez and Yanakakis, 2011).

In order to extract significant sequences of non-verbal signals conveying interpersonal stances from our corpus, we use a frequent sequence mining technique. To the best of our knowledge, this technique has not yet been applied to analyse sequences of non-verbal signals. In the following part, we describe the procedure used to mine frequent sequences in our corpus, and we then describe the result of applying this procedure on our data.

5.1 Applying sequence mining to our multimodal corpus

To apply the frequent sequence mining technique to our data, we proceed through the following six steps.

The first step consists of parsing the non-verbal annotations files, coded in the ELAN format, filtering the annotation modalities and time segments to investigate (e.g. we only consider here behavior sequences while speaking, therefore we discard the segments when the recruiter is listening) and converting every interaction’s annotations into a list containing all the non-verbal behaviors in a sequence.

The second step’s objective is to find events to segment the interactions: indeed, frequent sequence mining techniques require a dataset of sequences. In our case, our data consists of 3 continuous interactions. Since we investigate which
sequences of signals convey stances, we decide to segment the full interactions with attitude variation events: attitude variation events are the timestamps where an attitude dimension begins to vary. To this end, we parse the attitude annotations files, smooth them and find the timestamps where the annotated attitude dimension starts to vary. More details can be found in (Chollet et al., 2013).

We found that the attitude variation events in our data came with a wide range of values, *i.e.* in some cases the annotators moved the cursor a lot, indicating he annotators perceived a strong change in the recruiters’ stance from the recruiter’s behavior, while sometimes the cursor movements were more subtle. We chose to differentiate between small and strong attitude dimension variations, therefore we used a clustering technique to identify the 4 clusters corresponding to small increases, strong increases, small decreases and strong decreases. To this end, we used a K-means clustering algorithm with $k = 4$.

The fourth step consists of segmenting the full interaction sequences with the attitude variations events obtained from step 2. Following this procedure, we obtain 219 segments preceding dominance variations and 245 preceding friendliness variations. We found dominance segments to be longer in duration, averaging at 12.7 seconds against 8.3 for friendliness segments. These two sets are split further depending on which cluster the attitude variation event belongs to. For instance, we have 79 segments leading to a large drop in friendliness, and 45 segments leading to a large increase in friendliness (see Table 1).

Step five consists of applying the frequent sequence mining algorithm to each set of segments. We used the commonly used Generalized Sequence Pattern (GSP) frequent sequence mining algorithm described in (Srikant and Agrawal, 1996). The GSP algorithm requires as an input a minimum support, *i.e.* the minimal number of times that a sequence has to be present to be considered frequent, and its output is a set of sequences along with their support. For instance, using a minimum support of 3, every sequence that is present at least 3 times in the data will be extracted. The GSP algorithm based on the Apriori algorithm (Agrawal and Srikant, 1994): first, it identifies the frequent individual items in the data and then extends them into larger sequences iteratively, pruning out the sequences that are not frequent enough anymore.

Figure 4: Step 1 through 4 consist of preprocessing the data before performing sequence mining. Attitude variations events are detected and used to segment the non-verbal behavior stream. The result is a set of non-verbal behavior segments for each type of attitude variation event.

Figure 5: This figure illustrates the data mining process. All the segments for a given type of attitude variation event (here, an increase in dominance) are gathered. The result of the GSP algorithm is the set of sequences along with their support.

However, the support is an insufficient measure to analyse how a sequence is characteristic of a type of attitude variation event. For instance, having the gaze move away and back to the interlocutor happens very regularly in an interaction. Thus it will happen very often before all types of attitude variation events (*i.e.* it will have a high support), even though it is not sure that it characteristic of any of them. The objective of step 6 is to compute quality measures to assess whether a sequence is really characteristic of a type of attitude variation events. Based on (Tan et al., 2005), we choose to compute confidence and lift quality mea-
Table 1: Description of results for each attitude variation type

<table>
<thead>
<tr>
<th>Variation type</th>
<th>Cluster Center</th>
<th>Segment Count</th>
<th>Frequent Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendliness Large Increase</td>
<td>0.34</td>
<td>68</td>
<td>86</td>
</tr>
<tr>
<td>Friendliness Small Increase</td>
<td>0.12</td>
<td>66</td>
<td>72</td>
</tr>
<tr>
<td>Friendliness Small Decrease</td>
<td>-0.11</td>
<td>77</td>
<td>104</td>
</tr>
<tr>
<td>Friendliness Large Decrease</td>
<td>-0.32</td>
<td>36</td>
<td>67</td>
</tr>
<tr>
<td>Friendliness Total</td>
<td>247</td>
<td></td>
<td>329</td>
</tr>
<tr>
<td>Dominance Large Increase</td>
<td>0.23</td>
<td>49</td>
<td>141</td>
</tr>
<tr>
<td>Dominance Small Increase</td>
<td>0.09</td>
<td>66</td>
<td>244</td>
</tr>
<tr>
<td>Dominance Small Decrease</td>
<td>-0.13</td>
<td>80</td>
<td>134</td>
</tr>
<tr>
<td>Dominance Large Decrease</td>
<td>-0.34</td>
<td>24</td>
<td>361</td>
</tr>
<tr>
<td>Dominance Total</td>
<td>219</td>
<td></td>
<td>879</td>
</tr>
</tbody>
</table>

In the next part, we describe the sequences we extracted when applying this procedure to our corpus.

5.2 Results

As a first step, we wanted to get a glimpse at which kinds of non-verbal signals were more frequent in the extracted sequences of the different attitude variation types. For this purpose, we performed Student T-tests, comparing the number of occurrences of each signal type for the different types of attitude variations. Note that this is not meant as a complete analysis of the extracted sequences, but rather as an exploration of the types of signals most present in these sequences.

We found smiles to be significantly more common before large increases in friendliness than in all other cases (Small increase: \( p = 0.005 < 0.05 \), small decreases \( p = 0.001 < 0.05 \), large decreases \( p = 0.011 < 0.05 \)). Head nods happened significantly more often before large increases in friendliness than large decreases (\( p = 0.026 < 0.05 \)). The same was found for head shakes, which appeared more before large increases in friendliness than small decreases (\( p = 0.023 < 0.05 \)) or large decreases (\( p = 0.024 < 0.05 \)). Leaning towards the candidate was found to be more common before small increases in dominance than large decreases (\( p = 0.013 < 0.05 \)). Similarly, adopting a straight posture was more common before small increases in dominance, compared to small decreases (\( p = 0.040 < 0.05 \)) and large decreases (\( p = 0.001 < 0.05 \)). A head averted sideways was found to be more common before small increases in dominance than before large decreases (\( p = 0.019 < 0.05 \)). The same was found for crossing the arms (\( p = 0.044 < 0.05 \)).

To obtain a reasonable number of potentially relevant sequences, we have chosen to only identify the sequences present in our corpus at least 10 times (using a large minimum support would yield very few sequences, while a small minimum support would yield a very large number of sequences). The output of the GSP algorithm with a minimal support of 10 occurrences is a set of 879 sequences for dominance variations, and a set of 329 sequences for friendliness variations (see table 1). In average we found friendliness sequences to contain 2,91 signals, and dominance sequences to contain 3,58 signals.

In table 2 we show the top scoring (i.e. highest Lift score) extracted sequences for every attitude variation type found using this process. The Sup column corresponds to the support of the sequence and the Conf column to the confidence of the sequence. We have integrated the extracted sequences in an animation module for our ECA platform. Our next step consists of conducting user perceptive tests to validate that the sequences displayed by the virtual agent convey the expected attitude.

6 Conclusion

The complexity of non-verbal behavior expression and interpersonal stance perception in specific contexts motivates the use of a framework that considers all perspectives of behavior interpretation, and of a multimodal corpus as ground truth. We have proposed a multi-layer framework to handle the complexity of interpersonal stances expression and we annotated videos of job interview enactments. We presented a knowledge extraction method for non-verbal behavior sequences based on a data mining technique. Our future work consists of validating that the extracted sequences convey the appropriate interpersonal stance when
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Attitude Variation</th>
<th>Sup</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>BodyStraight -&gt; HeadDown</td>
<td>Friendliness Large Decrease</td>
<td>0.016</td>
<td>0.4</td>
<td>2.74</td>
</tr>
<tr>
<td>HeadDown -&gt; HeadAt -&gt; GestComm</td>
<td>Friendliness Small Decrease</td>
<td>0.032</td>
<td>0.72</td>
<td>2.33</td>
</tr>
<tr>
<td>HeadAt -&gt; HeadSide</td>
<td>Friendliness Small Increase</td>
<td>0.028</td>
<td>0.54</td>
<td>2.02</td>
</tr>
<tr>
<td>Smile</td>
<td>Friendliness Large Increase</td>
<td>0.061</td>
<td>0.52</td>
<td>1.88</td>
</tr>
<tr>
<td>GestComm -&gt; HeadDown -&gt; HeadAt</td>
<td>Dominance Large Decrease</td>
<td>0.028</td>
<td>0.42</td>
<td>3.80</td>
</tr>
<tr>
<td>HeadDown -&gt; HeadAt -&gt; HandsTogether</td>
<td>Dominance Small Decrease</td>
<td>0.041</td>
<td>0.75</td>
<td>2.05</td>
</tr>
<tr>
<td>HeadAt -&gt; ObjectManipulation -&gt; HandsOverTable</td>
<td>Dominance Small Increase</td>
<td>0.037</td>
<td>0.67</td>
<td>2.21</td>
</tr>
<tr>
<td>HeadDown -&gt; EyebrowUp</td>
<td>Dominance Large Increase</td>
<td>0.022</td>
<td>0.45</td>
<td>2.03</td>
</tr>
</tbody>
</table>

Table 2: Top scoring sequences for each attitude variation event

expressed by a virtual agent.

Acknowledgment

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

References


R. Cowie and G. McKeown. 2010. Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme.


B. Ravenet, M. Ochs, and C. Pelachaud. 2013. From a user-created corpus of virtual agent’s non-verbal behaviour to a computational model of interpersonal attitudes. In *International Conference on Intelligent Virtual Agent (IVA2013)*.


