Mining a Multimodal Corpus for Non-Verbal Signals Sequences Conveying Attitudes

Mathieu Chollet, Magalie Ochs, Catherine Pelachaud

Mining a Multimodal Corpus for Non-Verbal Signals Sequences Conveying Attitudes

Mathieu Chollet[1], **Magalie Ochs**[2], **Catherine Pelachaud**[2]

[1] Institut Mines-Telecom ; Telecom Paristech ; CNRS-LTCI
[2] CNRS-LTCI ; Telecom Paristech
46 rue Barrault, 75013 Paris, France
{mathieu.chollet, magalie.ochs, catherine.pelachaud}@telecom-paristech.fr

Abstract

Interpersonal attitudes are expressed by non-verbal behaviors on a variety of different modalities. The perception of these behaviors is influenced by how they are sequenced with other behaviors from the same person and behaviors from other interactants. In this paper, we present a method for extracting and generating sequences of non-verbal signals expressing interpersonal attitudes. These sequences are used as part of a framework for non-verbal expression with Embodied Conversational Agents that considers different features of non-verbal behavior: global behavior tendencies, interpersonal reactions, sequencing of non-verbal signals, and communicative intentions. Our method uses a sequence mining technique on an annotated multimodal corpus to extract sequences characteristic of different attitudes. New sequences of non-verbal signals are generated using a probabilistic model, and evaluated using the previously mined sequences.

Keywords: Non-verbal behavior, Interpersonal attitudes, Sequence mining

1. Introduction

Embodied Conversational Agents (ECAs) are increasingly used in training and serious games. 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. Such a virtual recruiter should be able to convey different *interpersonal attitudes* (or *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 attitudes are expressed through non-verbal behavior, and to implement the expression of interpersonal attitudes in an ECA.

Most modalities of the body are involved when conveying interpersonal attitudes (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).

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 embarassment (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 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 influence on attitude perception in a systematic fashion.

In previous work, we proposed a framework for analysis and expression of non-verbal behavior, composed of multiple layers focusing on 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 attitude 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 attitudes. 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 still limited. To gather knowledge about this layer, we use a data mining technique to extract sequences of non-verbal signals from the corpus. We then propose a model to generate other sequences and evaluate them using the sequences previously extracted from the corpus.

The paper is organized as follows. In Section 2, we present related models of interpersonal attitude expression for ECAs and their limits. We then introduce our multilayer framework for non-verbal behavior analysis and expression. 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. Finally, Section 6 discusses a method for generating and evaluating behavior sequences using the extracted data.

[1] http://http://www.tardis-project.eu/

2. Related work

Models of interpersonal attitude expression for virtual agents have already been proposed. For instance, in (Ballin and Crabtree, 2004), postures corresponding to a given attitude were automatically generated for a dyad of agents. Lee and Marsella used Argyle's attitude dimensions (see Figure1), 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 attitude and personality. Ravenet *et al.* (Ravenet et al., 2013) proposed a user-created corpus-based methodology for choosing the behaviors of an agent conveying an attitude 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.

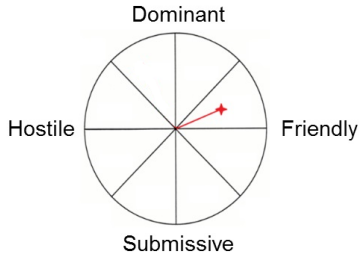


Figure 1: The Interpersonal Circumplex, with Argyle's attitude dimensions. The sample coordinate represents a friendly and slightly dominant interpersonal attitude.

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 attitude expression consider only consider one perspective at a time, with a limited number of modalities. Moreover, no model of attitude 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 framework for the expression of interpersonal attitudes

In previous work (Chollet et al., 2012), we defined a multi-layer framework 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 attitudes at every instant of the interaction, taking into account their behavior, their reactions to other interactants' behaviors, and their global behavior tendencies.

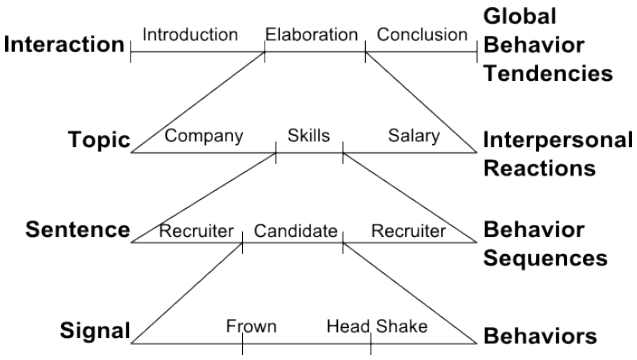


Figure 2: This figure illustrates the multi-layer model in a job interview setting. On the left are represented the layers of the model, and on the right which behavioral features they analyze.

Here is an example of how the different layers work in an interaction. Imagine a recruiter who is annoyed by a candidate because he thinks his foreign language skills do not meet the requirements for a job. The recruiter spreads his right hand towards the candidate while asking the question "You claim to be proficient in English. Can you prove it to me?". The candidate looks down for a while, thinking and hesitating. He looks up at the recruiter and tries an answer with a faint smile, then moving his head to the side. While the candidate is speaking, the recruiter frowns, and then shakes his head as the candidate finishes. All this time, the recruiter kept looking at the candidate.

In the example, the gesture performed by the recruiter is used to show a question is asked and that he gives the speaking floor to the candidate. These two communicative functions are handled by the *Signal* layer. When replying, the candidate smiles and then averts his head away from the recruiter. In that case, the *Sentence* layer considers the sequencing of signals: the smile could have been interpreted as a sign of friendliness at first, however followed by a head

aversion it is a sign of submissiveness. The recruiter behavioral replies to the candidate's answer, the frown and head shake, are analyzed by the *Topic* layer as sign of dominance and hostility. Finally, the fact that the recruiter barely averted gaze during the interaction is a sign of *dominance* revealed by the *Interaction* layer.

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 attitudes 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 attitude 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 attitudes. In order to study how recruiters express interpersonal attitudes, we annotated three videos of job interview enactments, for a total of slighty more than 50 minutes. We consider full body non-verbal behavior, turn-taking, task and interpersonal attitude.

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. The following modalities were considered : gestures (*e.g.* adaptors, deictics), hands rest positions (*e.g.* over or under table, arms crossed), postures (*e.g.* leaning backwards), head movements (*e.g.* nods, head tilted downwards), gaze (*e.g.* looking at interlocutor, downwards), facial expressions (Since the videos were recorded from the side, we only considered simple facial expressions, *e.g.* smiles, eyebrow movements). 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 ($\kappa >= 0.70$), 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.

As the interpersonal attitudes 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 attitude 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. As we focus here on how sequences of non-verbal signals are interpreted in terms of attitudes, we filtered the audio streams to make the speech unintelligible. Indeed, attitudes are also expressed through the choice of words and topics, and through prosody cues. Thus, it is very probable that hearing the recruiter's speech would have influenced the annotators, and having the speech filtered out assures us that the annotators are only influenced by non-verbal behavior. 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 (Chollet et al., 2013).

In a nutshell, the corpus has been annotated at two levels: the non-verbal behavior of the recruiters and their expressed attitudes. Our next step was to identify the correlations between the non-verbal behaviors and the interpersonal attitudes. 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 novel method for extracting knowledge about non-verbal behavior sequences from the multimodal corpora.

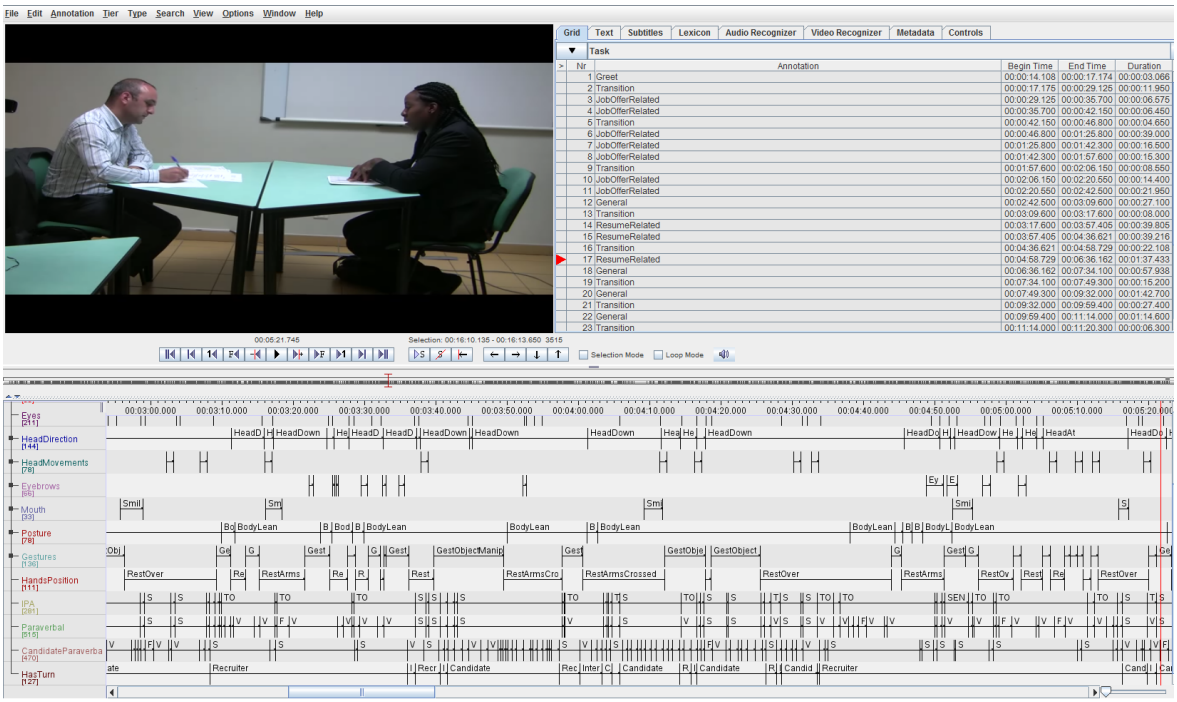


Figure 3: Video of the study in the Elan (Wittenburg et al., 2006) annotation environment

5. Mining non-verbal behavior sequences

It is only fairly recently that the importance of the dynamic features of expressions has been highlighted. Keltner *et al.* found that the sequencing of head aversion, gaze aversion and smile differentiate between embarassment, amusement and shame (Keltner, 1995). With found unique characteristic behavioral sequences for the expression of enjoyment, hostility, embarassment, surprise and sadness (With, 2010). To our knowledge, this work is the first attempt at discovering sequences of non-verbal signals that are characteristic of interpersonal attitudes expression.

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 contigency tables measuring the occurence 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 unrelevant to a particular sequence that can happen in the middle of it). Magnusson proposed the concept of *T-patterns* (Magnusson, 2000), sequences of events occuring 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 Yannakakis, 2011).

In order to extract significant sequences of non-verbal signals conveying interpersonal attitudes from our corpus, we chose to use *sequence mining* techniques. 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 attitudes, 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, smoothe 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' attitude 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).

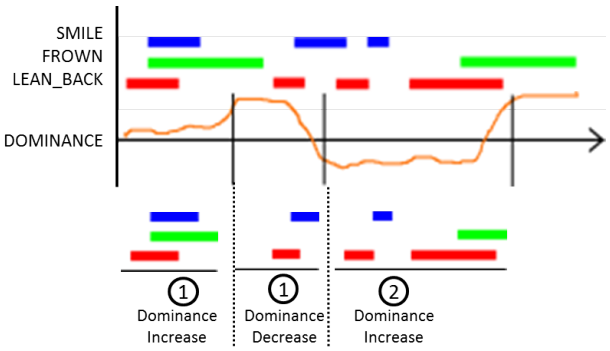


Figure 4: Step 1 through 4 consist of pre-processing the data before performing sequence mining. Attitude variations events are detected and used to segment the non-verbal behavior stream. The result is a set of non-verbal behavior segments for each type of attitude variation event.

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.

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

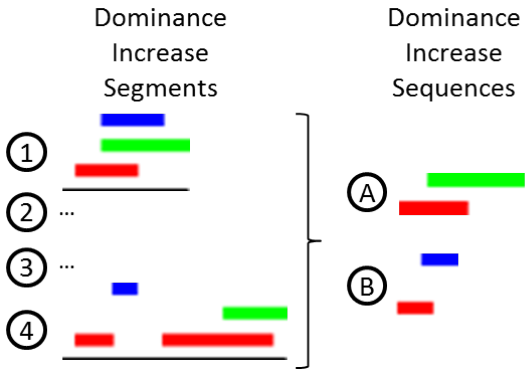


Figure 5: This figure illustrates the data mining process. All the segments for a given type of attitude variation event (here, an increase in dominance) are gathered. The result of the *GSP* algorithm is the set of sequences along with their support

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 measures for every sequence. The confidence represents how frequently a sequence is found before a particular type of attitude variation event. The lift represents how more frequently the sequence occurs before a type of attitude variation event than in other cases (the higher the value, the more likely it is that there is dependence between the sequence and the attitude variation).

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

Table 1: Description of results for each attitude variation type

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 study which signal occurs before an attitude variation perception. For this purpose, we perform Student T-test.

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.

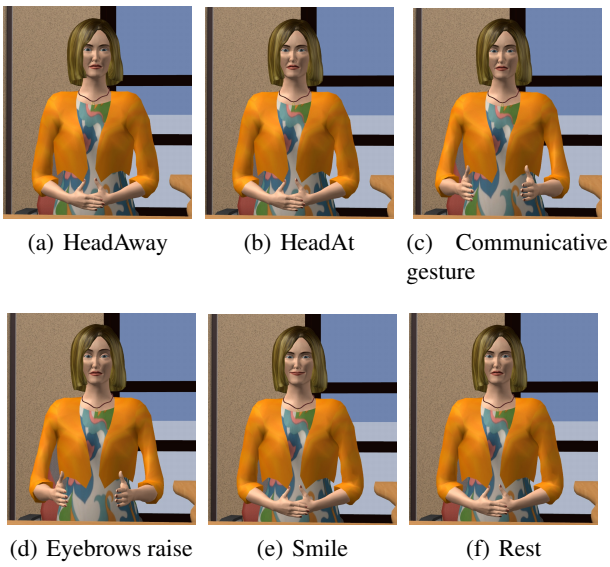
The results show that smiles were significantly more common before large increases in friendliness than in all other cases (Small increase: $p = 0.005 < 0.05$, small decreases $p = 0.001 < 0.05$, large decreases $p = 0.011 < 0.05$). Head nods happened significantly more often before large increases in 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 ($p = 0.044 < 0.05$).

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.

6. Generating new sequences

Our goal is to build a framework for attitude expression by Embodied Conversational Agents considering the sequencing of non-verbal signals, among other features. To that end, we have developed a module for our ECA platform that can animate the agent using the sequences extracted with the mining process detailed in Section 5. An example of such an animation is shown in Figure 6. However, solely relying on the sequences extracted in the previous step limits the variability and expressivity of our characters' behavior: the extracted sequences only consist of a subset of the possible sequences of signals that can be produced. Also, circumstances can occur where the agent has to express communicative intentions that can only be displayed with a particular set of signals (*e.g.* an emotion), and that no extracted frequent sequence contains these signals for the appropriate attitude variation.

In order to cope with these two limitations, we decided to build a probabilistic model for generating sequences. We adopted a Bayesian Network (BN) representation (see Figure 7): in the graphical representation of the network, arrows represent conditional dependence between variables. In our case, the attitude variation influences the types of signal appearing in the sequence, and each signal influences which type of signal appears directly after it. Thus we can note that $P(S_{i+1}|S_i, S_{i-1}, ..., S_1, A) = P(S_{i+1}|S_i, A)$, where S represents signals, i is the index of a signal in the sequence, and A is the chosen attitude variation.



(a) HeadAway (b) HeadAt (c) Communicative gesture

(d) Eyebrows raise (e) Smile (f) Rest

Figure 6: Exemple of the non-verbal sequence $HeadAt \rightarrow GestComm \rightarrow EyebrowUp \rightarrow Smile$ expressed by the virtual agent.

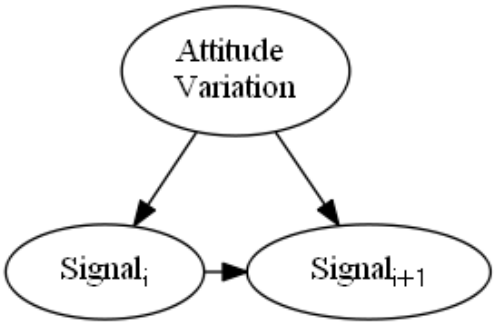


Figure 7: A "rolled" representation of the Bayesian Networks we use for generating new sequences of behavior.

For simplicity, we trained a BN for dominance variations and another BN for friendliness variations. We used the Weka open-source machine learning software (Hall et al., 2009) to train the networks, using the extracted frequent sequences as input data. Choosing a sequence length, we can generate a very large set of sequences using the networks. We can also constrain the generated sequences to contain a particular subsequence of signals, *e.g.* signals that are mandatory to produce some communicative intentions.

Any generated sequence can be evaluated with two criteria. The first one is the probability of a given sequence computed with the Bayesian Networks. Due to the structure of the networks and the way they are trained, this tends to reward the sequences containing the types of signals that

Sequence	Attitude Variation	Sup	$Conf$	$Lift$
BodyStraight -> HeadDown	Friendliness Large Decrease	0.016	0.4	2.74
HeadDown -> HeadAt -> GestComm -> HandsTogether	Friendliness Small Decrease	0.032	0.72	2.33
HeadAt -> HeadSide	Friendliness Small Increase	0.028	0.54	2.02
Smile	Friendliness Large Increase	0.061	0.52	1.88
GestComm -> HeadDown -> HeadAt -> HeadDown	Dominance Large Decrease	0.028	0.42	3.80
HeadDown -> HeadAt -> HeadDown -> HandsTogether	Dominance Small Decrease	0.041	0.75	2.05
HeadAt -> ObjectManipulation -> HandsOverTable	Dominance Small Increase	0.037	0.67	2.21
HeadDown -> EyebrowUp	Dominance Large Increase	0.022	0.45	2.03

Table 2: Top scoring sequences for each attitude variation event

are most frequent (*e.g.* head movements), but not the sequences that could be the most discriminant (*e.g.* smiles, adaptor gestures). In a nutshell, the most likely sequences of a given attitude variation type are rewarded, but not the most discriminant. The second criteria classifies any given sequence into a type of attitude variation using the method described in (Jaillet et al., 2006). This method relies on the concept of *confidence* we described in Section 5. It consists of extracting the k subsequences of the generated sequence that have the best *confidence* score in our data. The generated sequence is then classified using majority voting, *i.e.* the attitude variation type that is the most frequent in the k subsequences is considered to be the class of the sequence.

7. Conclusion

In this paper, we presented a knowledge extraction method for non-verbal behavior sequences based on a data mining technique, and a method for generating and evaluating new behavior sequences. Our next step is to evaluate the extracted and generated sequences by integrating the networks in our Embodied Conversation Agent platform. We will use the results to validate whether the extracted sequences are correctly recognized by users when they are directly reproduced by an agent, and to validate if the generated sequences are correctly classified, that is whether the users agree with the output of the classification method we described in the previous paragraph. Finally, we will tune our model to adjust the importance of the two criteria we proposed for sequence evaluation : how likely is a sequence to be produced, and how confident we are that the sequence is characteristic of the chosen attitude variation.

Acknowledgment

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

8. References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Allwood, J., Kopp, S., Grammer, K., Ahlsen, E., Oberzaucher, E., and Koppensteiner, M. (2007). The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Language Resources and Evaluation*, 41:255–272.

Argyle, M. (1988). *Bodily Communication*. University paperbacks. Methuen.

Bakeman, R. and Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press.

Ballin, D., Gillies M. and Crabtree, B. (2004). A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In *Proceedings of the 1st European Conference on Visual Media Production*, CVMP '04, pages 203–210.

Bevacqua, E., Sevin, E., Hyniewska, S. J., and Pelachaud, C. (2012). A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1-2):27–38.

Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):293–327.

Boersma, P. and Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Burgoon, J. K., Buller, D. B., Hale, J. L., and de Turck, M. A. (1984). Relational Messages Associated with Nonverbal Behaviors. *Human Communication Research*, 10(3):351–378.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., and Valgarðsson, G. S. (2012). First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Proceedings of the 12th international conference on Intelligent Virtual Agents*, IVA'12, pages 67–80, Berlin, Heidelberg. Springer-Verlag.

Carney, D. R., Hall, J. A., and LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2):105–123.

Chollet, M., Ochs, M., and Pelachaud, C. (2012). Interpersonal stance recognition using non-verbal signals on several time windows. In *Workshop Affect, Compagnon Artificiel, Interaction*, WACAI '12.

Chollet, M., Ochs, M., and Pelachaud, C. (2013). A multimodal corpus approach to the design of virtual recruiters. In *Workshop Multimodal Corpora, Intelligent Virtual Agents*, IVA '13, pages 36–41.

Cowie, R., Cox, C., Martin, J.-C., Batliner, A., Heylen, D., and Karpouzis, K. (2011). *Issues in Data Labelling*. Springer-Verlag Berlin Heidelberg.

Ekman, P. and Friesen, V. (1977). *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press.

Escalera, S., Pujol, O., Radeva, P., Vitria, J., and Anguera, M. (2010). Automatic detection of dominance and expected interest. *EURASIP Journal on Advances in Signal Processing*, 2010(1):12.

Ferreira, P. G. and Azevedo, P. J. (2005). Protein sequence classification through relevant sequence mining and bayes classifiers. *Progress in Artificial Intelligence*, 3808:236–247.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Exploration Newsletter*, 11(1):10–18, November.

Jaillet, S., Laurent, A., and Teisseire, M. (2006). Sequential patterns for text categorization. *Intelligent Data Analysis*, 10(3):199–214.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68:441–454.

LaFrance, M. (1982). Posture mirroring and rapport. In Davis, M., editor, *Interaction Rhythms: Periodicity in Communicative Behavior*, pages 279–299. New York: Human Sciences Press.

Lee, J. and Marsella, S. (2011). Modeling side participants and bystanders: The importance of being a laugh track. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, IVA'11, pages 240–247, Berlin, Heidelberg. Springer-Verlag.

Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, Computers*, 32:93–110.

Martínez, H. P. and Yannakakis, G. N. (2011). Mining multimodal sequential patterns: a case study on affect detection. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 3–10, New York, NY, USA. ACM.

Prepin, K., Ochs, M., and Pelachaud, C. (2013). 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)*.

Ravenet, B., Ochs, M., and Pelachaud, C. (2013). From a user-created corpus of virtual agent's non-verbal behaviour to a computational model of interpersonal attitudes. In *Proceedings of the 13th International Conference on Intelligent Virtual Agenst*, IVA '13, Berlin, Heidelberg. Springer-Verlag.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44:695–729.

Srikant, R. and Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology*, 1057:1–17.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

With, S. and Kaiser, W. S. (2011). Sequential patterning of facial actions in the production and perception of emotional expressions. *Swiss Journal of Psychology*, 70(4):241–252.

With, S. (2010). *Structural analysis of temporal patterns of facial actions: Measurement and implications for the study of emotion perception through facial expressions*. Ph.D. thesis, University of Geneva.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference*, LREC '11.