From Non-verbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression

Mathieu Chollet, Magalie Ochs, Catherine Pelachaud

From Non-verbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression

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

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

Abstract. In this paper, we present a model and its evaluation for expressing attitudes through sequences of non-verbal signals for Embodied Conversational Agents. To build our model, a corpus of interpersonal job interview interactions has been annotated at two levels: the non-verbal behavior of the recruiters as well as their expressed attitudes was annotated. Using a sequence mining method, sequences of non-verbal signals characterizing different interpersonal attitudes were automatically extracted from the corpus. From this data, a probabilistic graphical model was built. The probabilistic model is used to select the most appropriate sequences of non-verbal signals that an ECA should display to convey a particular attitude. The results of a perceptive evaluation of sequences generated by the model show that such a model can be used to express some interpersonal attitudes.

1 Introduction

Embodied Conversational Agents (ECAs) are increasingly used in training and social coaching, in applications such as science teaching [17], education against bullying [5]. Empathetic ECAs have been proposed [28] and it was shown that they can be successful in regulating the emotional state of users in a learning context, effectively affecting the learning outcome of the users [29].

In the TARDIS project[1], we aim at building a virtual recruiter to train job seekers to improve their social skills. Such a virtual recruiter should be able to convey different *interpersonal attitudes* (or *interpersonal stances*). Interpersonal attitudes can be defined as "*spontaneous or strategically employed affective styles that colour interpersonal exchanges*" [34]. A common representation for interpersonal stances is Argyle's bi-dimensional model of attitudes [3], with an affiliation dimension ranging from hostile to friendly, and a status dimension ranging from submissive to dominant.

Most modalities of the body are involved when conveying interpersonal attitudes [9]. Smiles can be signs of friendliness [9], performing large gestures may

[1] www.tardis-project.eu

be a sign of dominance, and a head directed upwards can be interpreted with a dominant stance [11]. However, when interpreting non-verbal behavior, the sequencing of non-verbal signals can be significant: for instance, while a smile is a sign of friendliness, a smile followed by a gaze and head aversion conveys embarassment [21]. While it has been observed that the sequencing of non-verbal signals influences how they are perceived [37], the literature on the topic is still limited. In this paper, our goal is to build a model for non-verbal behavior generation, which computes a sequence of non-verbal signals that an ECA should display given an input attitude to express and an input text that the ECA should say.

To build a model that takes the sequencing of signals into account, we use a data mining technique to extract sequences of non-verbal signals from a corpus of job interviews we annotated at two levels: the non-verbal behavior and the attitude of the recruiter. The generation model uses a probabilistic framework to compute a set of candidate sequences and then selects the best sequence for expressing the given attitude using a classification method based on the frequent sequences previously extracted from the corpus. The model was evaluated with an online study.

The paper is organized as follows. In Section 2, we present related models of interpersonal attitude expression for ECAs and their limitations. We then describe in Section 3 the multimodal corpus we collected and how it was annotated. Section 4 details the data mining process we used to gather knowledge about how sequences of non-verbal behavior are perceived. Section 5 discusses a method for generating and selecting behavior sequences using the extracted data. In Section 6, we describe the study we conducted to evaluate whether the generated sequences convey the appropriate attitude. Finally, the results of the evaluation study are discussed in Section 7.

2 Related work

Models of interpersonal attitude expression for virtual agents have already been proposed. For instance, in the Demeanour project [6], postures corresponding to a given attitude 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 [23]. Cafaro *et al.* [10] 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.* [33] proposed a user-created corpus-based methodology for choosing the behaviors of an agent conveying an attitude along with a communicative intention. The *Laura* agent [7] 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.* [32] have investigated how smile alignment and synchronisation can contribute to stance building in a dyad of agents. These models,

however, only consider the expression of a few signals at a given time, and do not consider how signals are sequenced.

The importance of the dynamic features of expressions has been highlighted in previous work. Keltner *et al.* [21] found that the sequencing of head aversion, gaze aversion and smile differentiate between embarassment, amusement and shame. With [37] found unique characteristic behavioral sequences for the expression of enjoyment, hostility, embarassment, surprise and sadness. Recently, models for displaying emotions for ECAs as sequences of signals have been proposed. Niewiadomski *et al.* [30] propose a representation for multimodal sequences of signals based on temporal constraints between signals, for instance $signal_1 \quad precedes \quad signal_2$. They were able to express several emotions from annotated videos using their representation scheme. Pan *et al.* [31] proposed another approach that makes use of motion graphs. In such graphs, arcs are motion clips (possibly containing facial expressions and/or head movements) and nodes are transitions between them. They trained a motion graph using video clips labelled with mental states (*e.g.* interest), and appropriate paths in the graph for each mental state are selected using dynamic programming. Finally, Lee and Marsella [24] proposed a model of head nods prediction based on Hidden Markov Models (HMM). The input of these HMMs is a sequence of words with associated linguistic features (*e.g.* part of speech, emotion label, noun phrase start...). Using an annotated corpus, they trained the prediction of head nods on trigrams, *i.e.* sequences of three words. Though they effectively adopted a sequential representation for their model, the sequential relationship between different head behaviors was not modelled (only linguistic features), and their work was limited to head movements.

Even though models of interpersonal attitude expression for ECA have already been proposed, they typically do not consider how the sequencing of signals influence attitudes, and only consider a limited number of modalities. In the next section, we present the corpus of job interview interactions we collected and annotated, and which was used subsequently to extract frequent sequences of non-verbal signals expressing interpersonal attitudes.

3 Multimodal corpus

As part of the TARDIS[1] project, a corpus of simulation of job interviews between human resources practitioners and youngsters was collected. We decided to use these videos to investigate the sequences of non-verbal signals the recruiters use when conveying interpersonal attitudes. The non-verbal behavior of the recruiters, their perceived attitudes and the turn taking were then annotated manually on 3 videos, for a total of slightly more than 50 minutes. Our sequence mining method (see Section 4) being relatively simple, we found this amount of data to be sufficient for our purpose. Of course, more data would allow for more precision and to use more complex models.

For the non-verbal behavior annotation, we adapted the MUMIN multimodal coding scheme [2] 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). Full details on the coding scheme can be found in [12]. We used Praat [8] for the annotation of the audio stream and the Elan annotation tool [38] for the visual annotations. A single annotator fully annotated the non-verbal behavior for the three videos. A second annotation on 10% of the total annotated video length was performed one month after the initial annotation to measure the reliability of the coding. Cohen's Kappa measures were computed across the two annotations and were found to be mostly satisfactory: *e.g.* $\kappa = 0.80$ for gestures, $\kappa = 0.93$ for postures. The lowest score was found for eyebrow movements ($\kappa = 0.62$), which we had anticipated considering the video setup.

As the interpersonal attitudes of the recruiters vary through the videos, we chose to use GTrace, successor to FeelTrace [14]. GTrace is a tool that allows for the annotation of continuous dimensions over time. We adapted the software for the interpersonal attitude dimensions we considered. The speech was rendered unintelligible, as we focus on non-verbal behavior and did not want the content of the recruiters' utterances to affect the annotators' perception of attitudes. We asked 12 persons to annotate the videos with this tool. Each annotator had the task of annotating one dimension for one video, though some volunteered to annotate more videos. With this process, we collected two to three annotation files per attitude dimension per video. More details about the attitude annotation can be found at [13]

In a nutshell, the corpus has been annotated at two levels: the non-verbal behavior of the recruiters and their perceived attitudes. Our next step was to identify which sequences of non-verbal signals characterize interpersonal attitudes. As a first step, we have focused on the non-verbal signals sequences expressed by the recruiters when they are speaking. In the next section, we describe a method for extracting frequent non-verbal signals sequences from the multimodal corpus.

4 Mining frequent sequences characterizing attitudes

In order to extract significant sequences of non-verbal signals conveying interpersonal attitudes from our corpus, we chose to use a *sequence mining* technique. Such techniques have been widely used in tasks such as protein classification [15], and they have been recently used in computer-human interaction to find sequences of video game players' key presses correlated with affects such as frustration [27]. To the best of our knowledge, this technique has not yet been applied to analyse sequences of non-verbal signals.

Frequent sequence mining techniques require a dataset of sequences. Since we investigate which sequences of signals convey attitudes, we decided to segment the non-verbal behavior data using the timestamps in the annotations files where an attitude dimension begins to vary. We call these instants *attitude variation events*. Once the data was segmented with these events, we kept only the

	Large Decrease	Small Decrease	Small Increase	Large Increase
Friendliness	0.34 / 68 / 86	0.12 / 66 / 72	-0.11 / 77 / 104	-0.32 / 36 / 67
Dominance	0.23 / 49 / 141	0.09 / 66 / 244	-0.13 / 80 / 134	-0.34 / 24 / 361

Table 1. Cluster centers, segment counts per cluster and frequent sequences per cluster

segments where the recruiter is speaking. Since we found that the attitude variation events came with a wide range of values, we chose to differentiate between small and strong attitude displays. Therefore we used a K-means clustering algorithm with $k = 4$ to identify clusters corresponding to small increases, strong increases, small decreases and strong decreases. The amount of segments per attitude variation type and the associated clusters are described in table 1.

The next step consisted of applying a frequent sequence mining algorithm to each set of segments. We used the Generalized Sequence Pattern (GSP) frequent sequence mining algorithm described in [35]. This algorithm extracts sequences without temporal information, *i.e.* it only represents that behaviors happened after another. It is not be able to differentiate between short and long gestures. It also cannot represent simultaneous events (*e.g.* a smile and a nod happening simultaneously). More recent sequence mining techniques exist that take temporal information into account [16], [18]. However, as a first step, we decided to choose a simpler model and focus on the sequential representation, as a higher model complexity would require more data to learn and would be harder to apply to our generation problem. Our model could potentially be complemented by related works considering simultaneous signals, such as [33]. The GSP algorithm requires as an input a minimum support, *i.e.* the minimal number of times that a sequence has to be present in the corpus 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 [1] follows two steps: first, it identifies the frequent individual items in the data and then extends them into larger sequences by iteratively adding other items, pruning out the sequences that are not frequent enough. Having acquired a set of frequent sequences for each type of attitude variation, we can characterize each of these sequences with several *quality measures*: *Support*, that is how many times the sequence appears in the data ($[0; \infty] \in \mathbb{N}$) ; *Confidence*, which represents the proportion of a sequence's occurrences that happen before a particular type of attitude variation ($[0; 1] \in \mathbb{R}$, 1 meaning this sequence only occurs before this attitude variation) ; *Lift*, which can be seen as how strong the confidence of a sequence is, compared to the random co-occurrence of sequence and the attitude variation, given their individual support ($[0; \infty] \in \mathbb{R}$, higher representing a stronger association).

In Table 2 we show examples of extracted sequences. The *Sup* column corresponds to the support of the sequence and the *Conf* column to the confidence of the sequence. Using a minimum support of 10, we extracted a set of 879 sequences for dominance variations and 329 for friendliness variations. In the next section, we describe an algorithm for generating non-verbal signals sequences

Sequence	Attitude Variation	*Sup*	*Conf*	*Lift*
BodyStraight -> ObjectManip	Friendliness Large Decrease	13	0.31	2.09
HeadNod -> Smile	Friendliness Large Increase	32	0.59	2.09
HeadNod -> RestHandsTogether -> Smile	Dominance Large Decrease	13	0.31	2.90
EyebrowsUp -> RestOverTable	Dominance Large Increase	21	0.33	1.54

Table 2. Example sequences obtained with the sequence mining process.

conveying attitudes, that makes use of the frequent sequences we presented in this section.

5 Model of non-verbal signals sequences generation for expressing attitudes

Given an input attitude that an ECA should express and an input utterance tagged with communicative intentions that the ECA should say, the objective of our model is to generate a sequence of non-verbal signals that conveys the appropriate attitude. We place ourselves within the SAIBA framework [36], where our model fulfils the role of the *Behavior Planner* module, whose role is to translate communicative intentions into multimodal behaviors and to schedule them. Input utterances and intentions are defined in the Functional Markup Language (FML) [25], and output sequences of scheduled non-verbal signals are defined in the Behavior Markup Language (BML) format [36].

In a nutshell, our algorithm follows three steps, which are detailed in the following subsections. First, for each communicative intention contained in the input FML message, we retrieve all the signals that can express this intention, and build all the possible combinations of signals that can express the input's communicative intentions (Section 5.1). Secondly, for each of these combinations, the algorithm then finds all the time intervals where additional signals can be inserted, and builds a set of larger sequences by inserting additional signals in the available time intervals using a probabilistic framework (Section 5.2). These signals will enable the agent to display its interpersonal attitude. The third step (Section 5.3) consists of selecting the best sequence out of all these candidate sequences, by using a classification method trained on the frequent sequences that were extracted using the method described in Section 4.

5.1 Building minimal sequences expressing the input FML

In a conversation, communicative intentions can be expressed through non-verbal behavior as well as through speech. For instance, in Western culture, it is possible to convey uncertainty by squinting the eyelids, tilting the head, or performing a particular hesitation gesture. When emphasizing a word, it is common to make a quick head movement downwards, and to raise one's eyebrows.

The FML language [25] represents such communicative intentions. The first step in our algorithm consists of retrieving all the possible non-verbal signals that

can be used to express the intentions contained in the input FML message. For this purpose, we used Mancini's framework [26], in which each communicative intention is characterized by a *behavior set*. A behavior set is the specification of the different non-verbal signals that can be displayed by an ECA to express a communicative intention. We can build a non-verbal signals sequence expressing an input message by selecting one signal in the behavior set of each communicative intention of the input message. Such a resulting sequence is called a *minimal sequence*. In our model, we only consider communicative intentions for altering the speech prosody (*i.e.* pitch accents and boundaries) and communicative intentions related to the speech semantics (*i.e.* spatio-temporal information which will trigger deictic gestures, particular meaning which will trigger iconic gestures, and performatives such as asking a question). Once all the minimal sequences have been computed (*i.e.* all the different combinations of signals from the behavior sets have been collected), the next step consists of enriching these sequences with additional signals to convey the interpersonal attitude.

5.2 Generating new sequences

For every minimal sequence obtained in the previous step, we start by looking at all the time intervals where it is possible to insert other signals. For instance, if there is enough time between two head signals, we might insert a head nod, or a head shake. Since the signals chosen for the minimal sequences are only related to speech prosody or to certain speech semantics, we make the hypothesis that the inserted signals will not conflict with the original communicative intentions, and that only the inserted additional signals will contribute to expressing attitudes.

For this purpose, we represent the extracted frequent sequences (Section 4) with a probabilistic model: a Bayesian Network (BN). The nodes of the network represents the non-verbal signals and the interpersonal attitudes (Figure 1). The edges define a conditional dependence between two variables. The Bayesian Networks enable us to represent the causal and non deterministic relation of the attitudes on the signals (*e.g.* there might be more smiles for friendliness increases, or more arms crossed for friendliness decreases) and the sequences of signals (*e.g.* hands rest pose changes typically appear often after a gesture). In our case, we 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. Note that some paths are impossible (*e.g. HeadAt* $\rightarrow$ *HeadAt or BodyStraight* $\rightarrow$ *BodyStraight*), and we made sure that such paths do not exist in the network.

An interesting feature of this model is that non-verbal signals sequences that did not occur in our data can still be generated, and their likelihood can be evaluated. Indeed, the representation of the sequences might lead to new sequences in the network. These new sequences are valuable as they can help improving the variability of the recruiter's behavior beyond the sequences that were observed in the corpus. To evaluate a new sequence, we can use the method described in [20] which classifies a new sequence with a majority-voting technique using its k sub-sequences contained in the new sequence that have the highest confidence

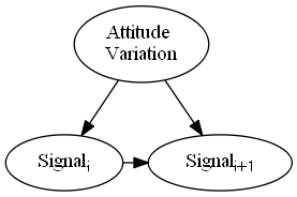

Fig. 1. A "rolled" representation of the Bayesian Networks we use for generating new sequences of behavior.

score in the corpus. We trained a BN for dominance variations and another BN for friendliness variations. As a first step, we consider the attitudes of friendliness and dominance independently. A next step will consist in analysing how to combine attitude variations in the two dimensions simultaneously. We used the Weka open-source machine learning software [19] to train the networks, using our multimodal corpus as input data. The constructed models are then used to compute the sequences of non-verbal signals conveying a particular attitude.

The generation of the sequences starts with the minimal sequences obtained after the previous step (Section 5.1), and uses the Bayesian Networks to add new signals in the available intervals. Thus, it is ensured that every generated sequence contains signals that express every input communicative intentions. Also, the maximum sequence length (*i.e.* how far we "unroll") is the amount of time intervals contained in the original FML message. In order to reduce computing time and to sort out sequences that are too unlikely, we compute the overall probability of every generated sequence, and only keep those whose probability is above a certain threshold λ. For our evaluation, we chose λ to be equal to $P(minimalsequence) * \alpha$ where $P(minimalsequence)$ is the probability of the original minimal sequence and α is a coefficient, which we set to 0.005 after trial-and-error showed it to be an adequate compromise between the amount of generated sequences and computing time. The generated sequences that are left after this pruning process are called *candidate sequences*. Having computed all the candidate sequences, the final step consists of selecting the one that is most likely to convey the input attitude.

5.3 Selecting the final sequence

For selecting the final sequence, we compromise between having a sequence with a high likelihood to appear in the data, and a high confidence for conveying the appropriate attitude. We defined a score variable of a candidate sequence s as $Sc(s) = P(s) * Conf(s)$, where $P(s)$ is the probability of s computed by the appropriate Bayesian Network (BN for dominance or friendliness depending on the input attitude). If the sequence s has been extracted in the frequent sequence mining process, then $Conf(s)$ is equal to the sequence's confidence (see Section 4). If not, we compute $Conf(s')$ for every subsequence s' contained in s, and we define $Conf(s) = \Sigma Conf(s')/n$ where n is the number of subsequences of s. Finally, we select the sequence with the highest score Sc. Note that for a

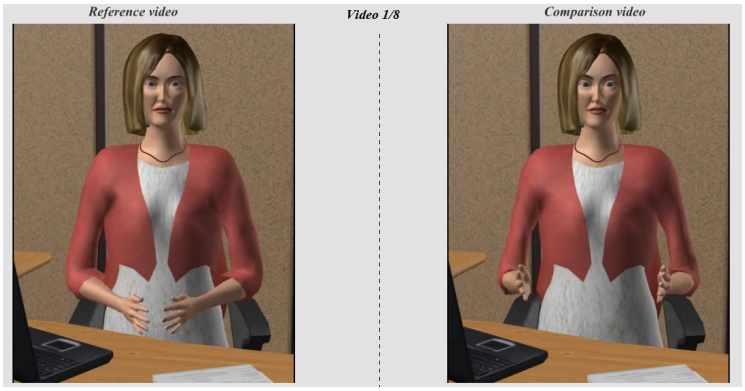

Fig. 2. The main screen of the online study.

particular input utterance, the chosen sequence will always be the same with this method. In the future, we plan to add a factor to weigh down sequences very similar to previously played sequences.

In the next section, we present an evaluation study we realized to assess whether the generated sequences convey the expected attitudes.

6 Evaluation

In order to evaluate our model, we conducted a study to verify that non-verbal signals sequences generated by the model with a certain input attitude are perceived as conveying the same attitude , and are perceived with the same intensity. In the following sections, we describe the study design and we then report the results of the study.

6.1 Study design

The study was conducted online. The platform of the study was developed using Adobe Flash technology. Participants were asked to compare 8 pairs of videos of a virtual character acting as a job recruiter expressing non-verbal signals when speaking (see Figure 2). For every pair of videos, the virtual recruiter said a different job interview question (*e.g.* "In your previous professional experiences, did you ever have to deal with difficult situations?"). The 8 different questions were always presented to the users in the same order. The character's speech was identical in both videos, and was produced in English with the Cereproc Text-To-Speech engine [4]. The non-verbal behavior of the recruiter was however different in the two videos of every pair. Since the speech content was identical in both videos, and since we asked user the difference in attitude between both videos, we considered the utterances' content did not have an impact on the results.

Each of these 8 pairs of videos corresponded to a testing condition, namely one of the 8 following attitudes: *high dominance, low dominance, low submissiveness, high submissiveness, high friendliness, low friendliness, low hostility,*

high hostility. On the right video (Figure 2), the ECA displayed sequences of non-verbal signals generated with our model. To generate a sequence for a given condition, we used as an input the corresponding attitude variation, for instance for *low hostility* our input was *small friendliness decrease*. On the left video, the virtual character's behavior was generated with a neutral attitude. For this purpose, our model only went through the first step of our algorithm (Section 5.1), and selected randomly one of the minimal sequences for the input question. While such a sequence could effectively express an attitude, we hypothesised that the attitude expressed in these videos would be considered more neutral than the attitude expressed in the videos generated with the attitude model. All in all, 64 distinct sequences were evaluated with our study (8 questions said by the ECA $*$ 8 attitudes), and 72 videos were generated for this purpose: (64 + 8 neutral). For every pair of videos, the participants answered the following questions: *Q1: "Compared to the Reference Video (left), the character on the Comparison video (right) is:"*, with the possible answers being : "Much less dominant", "Less dominant", "Equivalent", "More dominant", "Much more dominant", "Undecided" (resp. friendly). If they had not chosen "Undecided", they would then be asked to give their opinion on the next question: *Q2: "The intensity of the expressed attitude on the Comparison video (right) is:"*, with the possible answers being : "Very low", "Low", "Medium", "High", "Very high". In the following section, we present the results of this study.

6.2 Results

Eighty-one participants took part in our study (43 Female, 38 Male). The participants were mostly French (88%), and the mean age of the population was 32.4 years old (*StdDev*: 12.8).

In Table 3, we report the frequency table for participants' answers for *Q1*. The *Mean* values are computed by considering the answers are on an ordinal scale (Much less friendly = 1, Much more friendly = 5, *etc.*). To assess the statistical significance of our results, we performed χ^2 tests for every condition, and all were found to be significant (for all conditions $\chi^2 > 23.5$, $p < 0.0001$).

For *Q2*, we performed Student's T-tests between pairs of conditions of the same type (*i.e.* increase or decrease of dominance or friendliness) but different intensity (*i.e.* small or large). There was only a significant difference between perceived intensity of large ($Mean = 2.97$) and small ($Mean = 3.31$) decreases in friendliness ($p = 0.016 < 0.05$). Differences between increases in friendliness ($p = 0.62$), decreases in dominance ($p = 0.48$) and increases in dominance ($p = 0.73$) were not found to be significant.

7 Discussion

Our evaluation study aimed at assessing whether our model can generate sequences of non-verbal signals that convey the appropriate attitude, and if the generated sequences are perceived with the appropriate intensity.

	Friendliness Decrease	Friendliness Increase	Dominance Decrease	Dominance Increase
Much less (1)	3.73%	3.68%	2.26%	0.78%
Less (2)	**45.5**%	24.3%	24.1%	14.7%
Equivalent (3)	24.6%	**38.2**%	**39.1**%	20.9%
More (4)	20.9%	25.7%	27.1%	**52.7**%
Much more (5)	3.73%	8.09%	5.26%	9.30%
Undecided	1.04%	0%	2.26%	1.55%
Mean	2.71	3.10	3.02	3.50
$\chi^2(4)$	120.7	91.8	98.9	146.0

Table 3. Percentages table for attitude rankings for the 8 conditions.

The results of *Q1* indicate that expressions of dominance were indeed perceived as such. However, sequences for submissive attitudes were perceived as equivalent to the neutral expression. Expressions of hostile attitudes were perceived as less friendly than the neutral one. Moreover, the expressions of friendliness were perceived as conveying a more friendly attitude than the neutral non-verbal behavior. In other words, the results of the study validate partially our model. Indeed, our model seems to generate appropriate non-verbal signals sequences for the expression of dominance, friendliness and hostility. However, the model cannot be used to convey submissiveness.

For *Q2*, the only significant difference was found between intensity of large and small decreases in friendliness, however the videos generated for smaller variations were found to be more intense than the videos generated for larger variations. Therefore, it seems that our model cannot simulate attitudes of different intensities.

The analysis of the results brings some interesting considerations. One factor that might have influenced the results of *Q1*, is that speaking in a job interview context can be viewed as a form of asserting control over the interaction. Therefore it might be argued that a virtual recruiter cannot express submissiveness while speaking. Similarly, interactions are not a one-way exchange, and the behaviors of the recruiter when the interviewee is speaking are certainly critical to express friendliness. For instance, it is known that mimicking the behaviors of an interlocutor is a sign of friendliness [22]. However, our evaluation protocol did not allow us to study this effect. Moreover, while the non-verbal signals and their sequencing were different between compared and reference videos, there was no difference in behavior expressivity (*e.g.* gesture amplitude, smile intensity). Also, the notion of *intensity* mentioned in *Q2* could have been interpreted in other ways that we had anticipated (*e.g.* intensity of behaviors, whereas we studied intensity of attitudes). These two factors might have been influenced the participants in rating the videos with similar intensities. Finally, one caveat in our evaluation protocol is that our model considers attitude variations, whereas our study could only compare differences in attitude expression with neutral behavior. To really assess whether the model can express attitude variations, we

need to measure participants' appraisals of our virtual recruiter's attitude in full length, uninterrupted job interviews.

8 Conclusion

In this paper, we presented a corpus-based model for expression of attitudes by Embodied Conversational Agents and an evaluation study. From an annotated corpus of job interview, frequent sequences for different types of attitude expressions were extracted using a sequence mining technique. These were then used as data for building our sequence generation model based on Bayesian Networks.

The evaluation study validated that our model can generate non-verbal signals sequences for appearing friendly, hostile and dominant. However, the model was not able to express different attitude intensities. In future work, we plan on taking our model one step further by considering the listening behavior of the recruiter, and how the recruiter should react to behaviors of the interviewee. We also want to investigate how behavior expressivity, such as gesture amplitude or smile duration, is related to expressions of attitude and implement this in our model. Extensions of the sequence mining method considering temporal information will be considered. We will also extend our sequence generation and selection model to allow for simultaneously express dominance and friendliness variations. Finally, we plan on evaluating our model in full-length, uninterrupted simulated job interviews, to assess whether our model can express attitude variations through the course of an interpersonal interaction. To this end, we will define a measure of similarity between sequences, which will be used when selecting a new sequence to ensure that the behavior remains varied.

Acknowledgements

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

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Very Large Data Bases. pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)
2. 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, 255–272 (2007)
3. Argyle, M.: Bodily Communication. University paperbacks. Methuen (1988)
4. Aylett, M., Pidcock, C.: The CereVoice Characterful Speech Synthesiser SDK. In: Pelachaud, C., Martin, J.C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) Intelligent Virtual Agents, Lecture Notes in Computer Science, vol. 4722, pp. 413–414. Springer Berlin Heidelberg (2007)

5. Aylett, R., Paiva, A., Dias, J., Hall, L., Woods, S.: Affective agents for education against bullying. In: Affective Information Processing, pp. 75–90. Springer (2009)
6. Ballin, D., G.M., Crabtree, B.: A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In: First European Conference on Visual Media Production. pp. 203–210 (2004)
7. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Transactions in Computer-Human Interaction 12(2), 293–327 (2005)
8. Boersma, P., Weenink, D.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
9. Burgoon, J.K., Buller, D.B., Hale, J.L., de Turck, M.A.: Relational Messages Associated with Nonverbal Behaviors. Human Communication Research 10(3), 351–378 (1984)
10. Cafaro, A., Vilhjálmsson, H.H., Bickmore, T., Heylen, D., Jóhannsdóttir, K.R., Valgarðsson, G.S.: First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In: Intelligent Virtual Agents. pp. 67–80. IVA'12, Springer-Verlag, Berlin, Heidelberg (2012)
11. Carney, D.R., Hall, J.A., LeBeau, L.S.: Beliefs about the nonverbal expression of social power. Journal of Nonverbal Behavior 29(2), 105–123 (2005)
12. Chollet, M., Ochs, M., Clavel, C., Pelachaud, C.: A multimodal corpus approach to the design of virtual recruiters. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. pp. 19–24. ACII '13 (2013)
13. Chollet, M., Ochs, M., Pelachaud, C.: A multimodal corpus approach to the design of virtual recruiters. In: Workshop Multimodal Corpora, Intelligent Virtual Agents. pp. 36–41. IVA '13 (2013)
14. Cowie, R., Cox, C., Martin, J.C., Batliner, A., Heylen, D., Karpouzis, K.: Issues in Data Labelling. Springer-Verlag Berlin Heidelberg (2011)
15. Ferreira, P.G., Azevedo, P.J.: Protein sequence classification through relevant sequence mining and bayes classifiers. Progress in Artificial Intelligence 3808, 236–247 (2005)
16. Fricker, D., Zhang, H., Yu, C.: Sequential pattern mining of multimodal data streams in dyadic interactions. In: Development and Learning (ICDL), 2011 IEEE International Conference on. vol. 2, pp. 1–6 (2011)
17. Graesser, A., Chipman, P., King, B., McDaniel, B., D'Mello, S.: Emotions and learning with autotutor. In: Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work. pp. 569–571. IOS Press, Amsterdam, The Netherlands (2007)
18. Guillame-Bert, M., Crowley, J.L.: Learning temporal association rules on symbolic time sequences. In: ACML. pp. 159–174 (2012)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Expl. News. 11(1), 10–18 (Nov 2009)
20. Jaillet, S., Laurent, A., Teisseire, M.: Sequential patterns for text categorization. Intelligent Data Analysis 10(3), 199–214 (2006)
21. Keltner, D.: Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Journal of Personality and Social Psychology 68, 441–454 (1995)
22. LaFrance, M.: Posture mirroring and rapport. In: Davis, M. (ed.) Interaction Rhythms: Periodicity in Communicative Behavior. pp. 279–299. New York: Human Sciences Press (1982)

23. Lee, J., Marsella, S.: Modeling side participants and bystanders: The importance of being a laugh track. In: Intelligent Virtual Agents. pp. 240–247. Springer-Verlag, Berlin, Heidelberg (2011)
24. Lee, J., Marsella, S.C.: S.c.: Predicting speaker head nods and the effects of affective information. IEEE Transactions on Multimedia pp. 552–562 (2010)
25. Mancini, M., Pelachaud, C.: Dynamic behavior qualifiers for conversational agents. In: Intelligent Virtual Agents. pp. 112–124. Springer-Verlag, Berlin, Heidelberg (2007)
26. Mancini, M., Pelachaud, C.: The FML - APML language. In: The First FML workshop, AAMAS'08. Estoril, Portugal (May 2008)
27. Martínez, H.P., Yannakakis, G.N.: Mining multimodal sequential patterns: a case study on affect detection. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 3–10. ACM, New York, NY, USA (2011)
28. McQuiggan, S.W., Robison, J.L., Phillips, R., Lester, J.C.: Modeling parallel and reactive empathy in virtual agents: An inductive approach. In: Proceedings of 7th International Conference on Autonomous Agents and Multiagent Systems. pp. 167–174. AAMAS '08, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2008)
29. Moridis, C.N., Economides, A.A.: Affective learning: Empathetic agents with emotional facial and tone of voice expressions. IEEE Transactions on Affective Computing 3(3), 260–272 (2012)
30. Niewiadomski, R., Hyniewska, S.J., Pelachaud, C.: Constraint-based model for synthesis of multimodal sequential expressions of emotions. IEEE Transaction on Affective Computing 2(3), 134–146 (2011)
31. Pan, X., Gillies, M., Sezgin, T.M., Loscos, C.: Expressing complex mental states through facial expressions. In: Paiva, A., Prada, R., Picard, R.W. (eds.) Affective Computing and Intelligent Interaction. Lecture Notes in Computer Science, vol. 4738, pp. 745–746. Springer (2007)
32. Prepin, K., Ochs, M., Pelachaud, C.: Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents. In: Proceedings of COGSCI 2013 (2013)
33. Ravenet, B., Ochs, M., Pelachaud, C.: From a user-created corpus of virtual agent's non-verbal behaviour to a computational model of interpersonal attitudes. In: Intelligent Virtual Agents. Springer-Verlag, Berlin, Heidelberg (2013)
34. Scherer, K.R.: What are emotions? and how can they be measured? Social Science Information 44, 695–729 (2005)
35. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. Advances in Database Technology 1057, 1–17 (1996)
36. Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., Welbergen, H., Werf, R.J.: The behavior markup language: Recent developments and challenges. In: Intelligent Virtual Agents. pp. 99–111. Springer-Verlag, Berlin, Heidelberg (2007)
37. With, S.: 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 (2010)
38. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: Language Resources and Evaluation (2006)