HAL
open science

A Multimodal Corpus Approach to the Design of Virtual Recruiters

Mathieu Chollet, Magalie Ochs, Chloé Clavel, Catherine Pelachaud

To cite this version:

Mathieu Chollet, Magalie Ochs, Chloé Clavel, Catherine Pelachaud. A Multimodal Corpus Approach to the Design of Virtual Recruiters. Affective Computing and Intelligent Interaction, Sep 2013, Genève, Switzerland. pp.19 - 24, 10.1109/ACII.2013.10 . hal-01074861

HAL Id: hal-01074861

https://hal.science/hal-01074861v1

Submitted on 15 Oct 2014

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

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

A multimodal corpus approach to the design of virtual recruiters

Mathieu Chollet, Magalie Ochs, Chloé Clavel, Catherine Pelachaud
Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI
37 rue Dareau, 75014, Paris
{mathieu.chollet, magalie.ochs, chloe.clavel, catherine.pelachaud}@telecom-paristech.fr

***Abstract*—This paper presents the analysis of the multimodal behavior of experienced practitioners of job interview coaching, and describes a methodology to specify their behavior in Embodied Conversational Agents acting as virtual recruiters displaying different interpersonal stances. In a first stage, we collect a corpus of videos of job interview enactments, and we detail the coding scheme used to encode multimodal behaviors and contextual information. From the annotations of the practitioners' behaviors we observe specificities of behavior across different levels, namely monomodal behavior variations, inter-modalities behavior influences, and contextual influences on behavior. Finally we propose the adaptation of an existing agent architecture to model these specificities in a virtual recruiter's behavior.**

I. Introduction

This work is part of the TARDIS project, whose goal is to create a scenario-based serious-game simulation platform for young people to improve their social skills. The scenarios that are used for this purpose are job interview enactments, where an Embodied Conversational Agent (ECA) acts as a virtual recruiter. However, in human-human real job interviews, some recruiters may try to make the candidate as comfortable as possible, while other may act very distant and dominant.

In order to propose various job interview experiences to users of the serious game platform, we want to model virtual recruiters that can express different interpersonal stances through their non-verbal behavior. According to Scherer [1], interpersonal stances are *"characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous)"*

The relationship between interpersonal stance and non-verbal behavior has been widely studied: a head tilted aside can convey submissiveness [2], while a head directed upwards can be interpreted with a dominant attitude [2]. High amounts of gaze directed at the interlocutor, smiling and forward body lean are signs of immediacy [3]. Individuals who perform larger gestures are generally considered more dominant [4]. Multimodality can have effects in the perception of interpersonal stance: in [3], the combination of body leaning backward and of the absence of smiles was found to indicate much more disengagement than either of the cues alone. It is also the case that non-verbal behavior is influenced by the context of the interaction: for instance, there is a significant influence of the presence of a complex object relevant to the task being undergone by interactants on their gaze behavior [5].

Therefore, to allow for the definition of models of non-verbal behavior influenced by interpersonal stance, it should be possible to specify behavioral influences of the three following types: monomodal behavior variations between recruiters (i.e. inter-recruiters variation), inter-modalities behavior influences, and contextual influences on behavior.

As a preliminary step before the analysis of how virtual recruiters should express interpersonal stances, the aim of this paper is to analyze how these three different behavioral aspects appear in the behavior of recruiters, and how they can be specified in an ECA architecture.

For this purpose, we use a set of videos of job interview enactments performed by experienced practitioners of job interview coaching at a French local association. In order to obtain a representation of the non-verbal behavior in these videos, we define a coding scheme for the annotation of contextual information and multimodal behavior, and we proceed with the annotations of these videos. We then analyze the corpus of annotations regarding inter-character behavior variations, inter-modality influences on behavior, and contextual influences on behavior, using specific modalities as examples. Finally, we propose methodologies to specify these behavior influences in an agent architecture.

This paper is structured as follows. We start by presenting related works on Embodied Conversational Agents whose non-verbal behavior model is based on literature or multimodal corpora analysis. Then, we present the videos corpus we used in this study and the annotation coding scheme. In section IV, we discuss the corpus analysis we conducted on the three aforementioned types of behavioral influence. Finally, we present our methodology to extract specification parameters from the corpus.

II. Related work

Embodied conversational agents have already been used in a wide variety of domain-specific applications. Steve [6] is an ECA used as a mentor for team practice in a reproduction of a real learning environment, and is able to use deictic (e.g. pointing) gestures and gaze to guide attention and display work procedures, as well as managing turn taking. The Real State Agent Rea [7] is another example of a multimodal conversational agent whose behavior supports a specific task-domain. Audiovisual input from the user is interpreted into communicative functions, to which Rea responds with hard-coded non-verbal responses. More sophisticated non-verbal

behavior generation models from text have been proposed and can support generic speaking behavior [8]–[10].

Efforts have been made to expand the expressivity of ECA behaviors and to allow to specify distinctive behavior for different agent types or personalities. Mancini and Pelachaud propose a model for expressive distinctive behavior generation: every agent would be specified with a lexicon, a baseline and behavior qualifiers [11] configuration files. The lexicon is a repository of signals that express communicative intentions. The baseline is a configuration file that defines expressivity parameters for every modality: for instance, an agent could be configured to perform gestures less often than another agent. Finally, behavior qualifiers can be used to define how communicative intentions or emotional states influence even further the behavior tendency of an agent, for instance an agent in the "sad" state could perform smaller gestures.

The SEMAINE project aimed at building Sensitive Artificial Listeners (SAL) whose goal was to put the user into different emotional states. Four SALs with different personalities and appearances were defined. In particular a listening behavior model for virtual agents is proposed in [12]. It allows to display backchannels, a special kind of acoustic or visual signals that are used by the listener in the speaker turn to provide information such as engagement or understanding [13]. According to the agent's personality, different backchannel selection and triggering rules are defined using the backchannel literature.

The full definition of such agent behaviors can be difficult to do by hand: though literature on non-verbal behavior and its perception is extensive [2]–[5], [14], they are rarely operationalized and, understandably, no account systematically takes on the full range of monomodal and multimodal behavioral influences. Another method for synthesizing behavior is to use a corpus of behavioral data. Rehm and Andre [15] suggest that there are two main ways to go about this task: either using the data to directly specify behaviors of particular agents, making their behavior exactly reflect the content of the corpus, or finding relationships between low-level behavior and higher-order dimensions, such as expressivity parameters, personality, attitudes.

Contextual influence is modelled in [16], where Kipp proposes a method for partially automating the authoring of non-verbal actions based on a dialogue script. A corpus of dialogue scripts enriched by human authors with non-verbal actions is collected. The authors can also manually define rules for gesture generation. Using machine learning, gesture generation rules are extracted, using features on the verbal content of the utterance and the context (e.g. turn-taking, first utterance in the scene...). All theses rules are then used on new utterances to generate many possible non-verbal behavior occurences. As the generated behavior occurences are too numerous, they are then filtered out using a display rate variable. This is refered in [15] as a, "overgenerate and filter" approach. Foster and Oberlander propose a corpus-based model for generation of head and eyebrow movements while speaking [17]. As an input, this model uses the speech content of the sentences, as well as contextual features such as the pragmatic content of the sentences.

In Buisine *et al.* [18], cooperation between modalities is investigated and implemented in an embodied conversation agent. The LEA agent is a 2D agent made out of images that are combined in order to produce a resulting pose. For instance, the set of right arm images comprises a greeting pose (arm held high with an open palm oriented towards the user), a pose where it lies along the body, and a few deictic poses. To point at a specific object, the LEA agent can use gaze, verbal references ("the object on my right"), and deictic gestures. After defining a set of cooperation cases between modalities, such as equivalence, complementarity, redundancy, they then used an annotated multimodal corpus of human behavior to learn the rate at which humans use these cooperation strategies, and used this knowledge to specify the multimodal behavior of the LEA agent. Effectively, it is an effect of inter-modality influence that is modelled in their work.

An example of distinctive behavior modelling between characters is the model of posture generation adapted with culture parameters proposed by Lipi *et al.* [19]. Basing themselves on the Hofstede model of socio-cultural characteristics, they learn the parameters of a Bayesian network using a corpus of videos of interactions between Japanese and German speakers from which posture information is extracted, such as the mean spatial extent, or the number of mirroring occurences. This Bayesian network is then integrated into a posture decision module that can select postures that are appropriate for a culturally embodied agent.

Finally, several approaches exist for behavior generation from affective information. Closest to our goal of modelling the influence of interpersonal stance is the work of Ballin *et al.* [20]. They represent interpersonal attitude with the*affiliation* and *status* dimensions, and map appropriate postures on this dimensional space. Lee *et al.* [21] predict occurences of speaker head nods using Hidden Markov Models with linguistic and affective features of utterances. Their compare their approach with rule-based head nod generation and find the behaviors produced by machine learning are perceived more natural. In [22], Li and Mao build a model for eye movement generation from emotional information.

III. MULTIMODAL CORPUS PRESENTATION

This section presents the corpus of videos we use to analyze recruiters' multimodal behaviors and our annotation scheme.

A. Corpus description

As part of the TARDIS project, a study was conducted with practitioners and youngsters from the Misson Locale Val d'Oise Est (a national association organizing job interview coaching for youngsters in search for a job). The study consisted in creating a situation of job interviews between 5 practitioners and 9 youngsters. The setting was the same in all videos. The recruiter and the youngster sat on each side of a table. A single camera embracing the whole scene recorded the dyad from the side (see Fig. 1). We gathered a corpus of 9 videos of job interview lasting approximately 20 minutes each. We had to discard 4 videos as the recruiter was not visible due to bad position of the camera.

B. Multimodal coding scheme

Numerous coding schemes exist to annotate non-verbal corpora. For instance, Kipp [23] propose a coding scheme for the annotation of gesture phases. A widely used system for facial expressions is the Facial Action Coding System [24]. A very exhaustive coding scheme is the MUMIN multimodal coding scheme [25]. It is used for the analysis of turn-taking and multimodal sequencing analysis.

A main issue when choosing a multimodal coding scheme is the granularity level [15]. In our case, we focus on the upper body behaviors of the recruiters. The interviews were recorded with one camera recording the recruiter-candidate dyad from the side and at a rather long distance (around 3 meters). From such a distance, it is not possible to see fine grained behaviors so we choose to stay at a high level of description. We use the MUMIN multimodal coding scheme [25] and adapt it by removing any types of annotations we cannot extract from the videos (i.e. subtle facial expressions, or absent expressions). We use Praat [26] for the annotation of the audio stream and the Elan annotation tool [27] for the visual annotations.

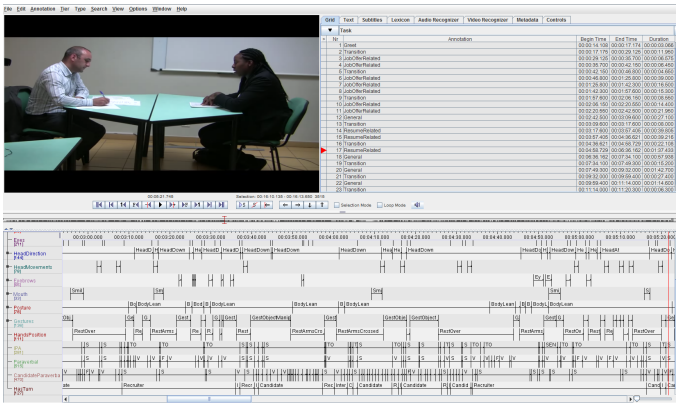

Fig. 1. A screenshot of one annotated video in the Elan annotation environment.

1) Facial behavior: Table 1 lists the annotation tags defined for the facial behavior of the recruiter. Gaze and head behaviors are included, as they are related to interpersonal stance [2], [3]. Because of the camera-dyad distance, we do not try to annotate very complex facial expressions (e.g. action units for facial muscle movements, as in [24]), however we include smiles and eyebrow movements (raised and frown).

TABLE I. FACIAL BEHAVIOR ANNOTATION TAGS

Modality	Expression	Tag	Optional
Gaze	Looking at interlocutor	GazeAt	
	Looking at object (e.g. resume)	GazeObject	
	Looking upwards	GazeUp	
	Looking downwards	GazeDown	
	Looking sideways	GazeSide	
Head Direction	Head directed at interlocutor	HeadAt	
	Head directed upwards	HeadUp	Int[1]
	Head directed downwards	HeadDown	Int[1]
	Head directed sideways	HeadSide	Int[1]
	Head tilted to the side	HeadTilt	Int[1]
Head Movement	Nod	HeadNod	Int[1], Rep[2]
	Shake	HeadShake	Int[1], Rep[2]
Eyebrow	Eyebrow raised	EyebrowUp	Int[1]
	Eyebrow frowned	EyebrowDown	Int[1]
Mouth	Smile	Smile	Int[1]

2) Body behavior: Table 2 lists the annotation tags defined for the body behavior of the recruiter. We consider posture and gestures, two important social cues [3], [4]. For gestures, we include object manipulations and adaptor gestures: individuals fidgeting with objects, e.g. a pen, or scratching, are perceived as nervous [4]. We also include several hands rest positions.

TABLE II. BODY BEHAVIOR ANNOTATION TAGS

Modality	Expression	Tag	Optional
Posture	Sitting straight	BodyStraight	
	Leaning towards the table	BodyLean	Int[1]
	Reclining back in the chair	BodyRecline	Int[1]
Gesture	Communicative gestures	GestureComm	Int[1], Spa[3]
	Object manipulation	GestureObjManip	
	Adaptor gestures (e.g. scratching)	GestureAdaptor	BPart[4]
Hand Position	Hands resting on the table	RestOver	
	Hands resting under the table	RestUnder	
	Arms crossed	RestArmsCrossed	
	Hands together	RestHandsTogether	

3) Vocal annotations: Table 3 lists the annotation tags defined for the vocal behavior. The para-verbal is used to differentiate when a participant is speaking, is silent, laughing, or performing a filled pause (e.g. with non-verbal vocalizations such as "err" or "hmm"). We also annotate whether an utterance of the recruiter is task-oriented or socio-emotional [28].

TABLE III. VERBAL BEHAVIOR ANNOTATION TAGS

Modality	Expression	Tag
Para-verbal	Speaking	VSpeaking
	Silent	VSilent
	Laughing	VLaughing
	Filled pause	VFilledPause
Verbal	Task-oriented sentence	TO
	Socio-emotional positive	SEPos
	Socio-emotional negative	SENeg

4) Context annotations: Table 4 lists the annotation tags for the interaction context. The task tags are related to the subjects of discussion in the task-domain, here job interview topics: apart the opening and closing parts of interviews, we mainly distinguish topics related to a relevant document, such as the candidate's CV or the job offer, and related classic interview questions (e.g. "Why should I pick you for the job and not another candidate?"). We also annotate turn-taking information. Note that the interruption tag is only used when there is a clear overlap of speech when the listener tries to take the turn: when the listener produces a backchannel, we consider the speaker keeps the turn [29]. The occlusion tags are used to specify the time intervals when the recruiter is not visible in the videos, e.g. when the camera operator zooms the image on the candidate's face. These tags are included to make the data analysis easier, allowing us to ignore the temporal segments with visual occlusions.

[1] $Int = High, Normal, Low$ is an optional Intensity parameter ranging from low to high for highly emphasized expressions.

[2] $Rep = Yes, No$ is an optional Repetitions parameter used for head movements to differentiate between single occurrences of a movement or repeated continuous head movements.

[3] $Spa = Small, Normal, Large$ is an optional Spatial parameter used when a gesture is particularly wide or small in amplitude.

[4] $BPart = Face, Hair, Neck, Hands, Body, Other$ is an optional BodyPart parameter used for adaptor gestures to annotate what the person is touching: which body part, or an object.

TABLE IV. CONTEXT ANNOTATION TAGS

Type	Value	Tag
Task	Question related to the CV or the job offer	DocumentRelated
	Question on another purpose	General
	Greetings or goodbyes	GreetFarewell
	Transition	Transition
Turn-taking	Recruiter has the turn	Recruiter
	Candidate has the turn	Candidate
	Interruption of the previous turn	Interruption
	Turn abandonned by both participants	Silence
Occluions	Partial or complete occlusion	Occlusion

IV. CORPUS ANALYSIS

With this corpus of multimodal behavior annotations, we then proceed on the analysis of the specificities of the behavior of job interview recruiters. We present results extracted from the analysis of a sub-set of our corpus. Further investigations are still going on.

A. Contextual influence on behavior

We check the influence of the current task and the current speaking turn on non-verbal behavior. Influence of the context on gaze behavior has been observed in [5]: an interactant looks more at his interlocutor while listening rather than while speaking. Also, the presence of a relevant object influences gaze behavior. The relationship between gestures and speech is also well established [30].

We test the dependency between behaviors and the interaction state with Pearson's Chi-square test. The results for the dependency between behaviors and the interaction state are listed on table V. For facial expressions and head movements, there were too few occurences in some states to use the Chi-squared test, and we only report for the other behaviors.

TABLE V. CHI-SQUARED TESTS FOR DEPENDENCY BETWEEN NON-VERBAL BEHAVIOR AND INTERACTION STATE

	$\chi^2\{Vid1,2,3\}$[5]	p-value	Validity
Gaze	$\{101, 190, 104\}$	< 0.001	Yes
Head Direction	$\{169, 86, 172\}$	< 0.001	Yes
Posture	$\{51, 68, 75\}$	< 0.001	Yes
Hands position	$\{81, 50, 121\}$	< 0.001	Yes
Gestures	$\{135, 41, 65\}$	< 0.001	Yes

A dependency is found between all the tested behaviors and the interaction state.

B. Inter-character behavior differences

We investigate the number of posture shifts in our video corpus. A quick glance at the data shows there are significant differences in the posture shift behavior of different recruiters (see Table VI). Using Pearson's Chi-squared test confirms this impression: $\chi^2 = 76.531 > \chi^2(0.05, 2) = 5.99, p < 0.001$

We also analyze if there is a significant difference between the different recruiters' expressive parameters of facial expressions, head movements and gestures. A dependency observed for head movements parameters (i.e. Intensity and Repetitions) and gesture parameters (i.e. Intensity and SpatialExtent), but not for facial expression parameters (i.e. Intensity).

TABLE VI. POSTURE SHIFTS

	Vid1	Vid2	Vid3
Posture shifts	78	32	13
Total unoccluded time	841s	865s	1369s
Posture shifts/sec	0.093	0.037	0.009

TABLE VII. CHI-SQUARED TESTS FOR DEPENDENCY BETWEEN EXPRESSIVE PARAMETERS AND RECRUITER

	χ^2	p-value	Validity
Eyebrows Int[1]	4.7	0.32	No
Mouth Int[1]	1.3	0.87	No
GestComm Int[1]	96.3	< 0.001	Yes
GestComm Spa[3]	97.6	< 0.001	Yes
HeadMvmt Int[1]	15.6	0.04	Yes
HeadMvmt Rep [2]	27.2	< 0.001	Yes

C. Inter-modality influences

We look at evidence for postural influence on communicative gesture parameters. As the practitionner in the third video very rarely changes posture, we only investigate the first and second videos. Statistical significance is only found between the spatial parameter and posture: $\chi^2\{Vid1, Vid2\} = \{11, 12\} > \chi^2(0.05, 4) = 9.48$

V. AGENT SPECIFICATION USING THE CORPUS

Drawing on the conclusions from the corpus analysis, we propose a methodology to extract specification parameters from the multimodal corpus. We present an adaptation of the Semaine agent architecture [31] that can be configured with these parameters. The Semaine architecture allows real-time interaction between human users and virtual agents. Its architecture is very flexible and modular and is easily adaptable to our needs.

A. Semaine SAL architecture adaptation

The details of the Semaine agent architecture can be found in [31]. As we find a dependency between all the tested behaviors and the interaction state, we argue that there is a need for an interaction manager module that keeps track of the current task and turn of speech, and subsequent modules that process behavior must adapt with respect to the interaction state. Therefore, we adapt this architecture by adding an Interaction Manager module.

The main function of the Interaction Manager is to update the interaction state, which consists of a turn-taking variable, which can take the values $\{Speaking, Listening\}$, and a task variable, which can take the values $\{Document, General\}$. The turn-taking variable is known using the previous Semaine Dialog Manager turn-taking mechanism [31]. The task variable is known in the scenario definition, provided by an external module of the TARDIS project (not shown in Fig. 2).

B. Specification of behavior parameters

1) Contextual influence on behavior: The role of the Action Selection module in the Semaine architecture is to filter backchannels opportunities triggered by the Listener Intent

[5]The results are ordered by the video indeices: Vid1, then Vid2, and Vid3.

[6]GazeUp, GazeSide and GazeDown annotations are regrouped into a GazeAway category, thus there are $(3-1)*(4-1) = 6$ degrees of freedom.

[7]On videos 1 and 2, there were no RestUnder annotations. On video 2, there are no RestArmsCrossed annotations.

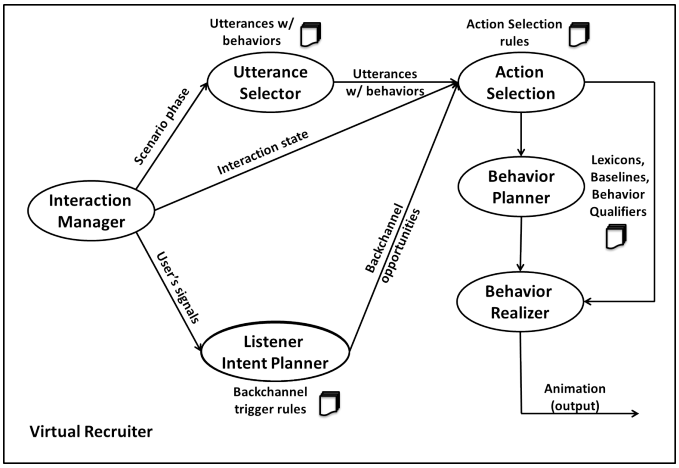

Fig. 2. Adaptation of the Sensitive Artificial Listener architecture for a virtual recruiter. Based on Fig. 1 of [12]

Planner and to select their type. We propose to use the Action Selection module as the processing module for the influence of the interaction state on behavior. As an example, we present how hands rest positions selection can happen in this module, and how it can be specified.

The Utterance Selector module selects the next utterances of the virtual recruiter. These utterances are pre-defined according to the job interview scenario definitions, and are enriched with gesture occurrences. The time segments where hands are resting are the intervals where no gestures happen. However, the hands rest positions are not instantiated in the dialog, even though they depend on the recruiter and the interaction state. We propose to specify probabilities for hands rest position selection for every recruiter depending on the interaction state. To compute them, we simply use the percentage of hands rest positions type relatively to all the hands rest occurrences for each interaction state. Between 2 gesture tags, the Action Selection module, by generating a number on an uniform distribution between 0 and 1, can then select an appropriate hands rest position according to the interaction state.

TABLE VIII. HANDS REST POSITIONS PROBABILITIES

	Document Speaking	Document Listening	General Speaking	General Listening
Under[5]	0.00 /0.00/ 0.09	0.00 /0.00/ 0.27	0.00 /0.0/ 0.32	0.00 / 0 / 1
Over[5]	0.28 /0.05/ 0.31	0.54 /0.07/ 0.16	0.02 /0.0/ 0.23	0.09 / 0 / 0
Arms-Crossed[5]	0.59 /0.61/ 0.00	0.33 /0.93/ 0.00	0.41 /0.7/ 0.00	0.65 / 0 / 0
Hands-Together[5]	0.13 /0.34/ 0.60	0.13 /0.00/ 0.57	0.57 /0.3/ 0.45	0.26 / 1 / 0

2) Inter-character behavior differences: In our corpus analysis, we found evidence of inter-character behavior differences for posture shift behavior and for expressivity gesture and head movement parameters. We use the case of the posture shift as an example.

The Semaine agents only display face expressions and can only move their head: in order to recreate the recruiters' behavior, we need to add a posture shift mechanism and to provide a methodology for specifying its parameters.

In Cassell *et al.* [7], the link between discourse structure and posture shifts is investigated on a corpus of monologues and dialogues, and they find that most posture shifts occur when there is a conversation topic change, or at the boundaries of discourse segments. We build upon those findings by adding a posture shift triggering mechanism driven by interaction state changes. Much as what the Listener Intent Planner does with its backchannels trigger mechanism, the Interaction Manager sends a message to the Action Selection module after each interaction state change. The Action Selection module then filters these messages according to selection rules specified with the multimodal corpus.

For every recruiter, we define the probability that a posture occurs with a task change, and with the start or the end of a "Speaking" turn. For these three cases, we first compute the time interval between every interaction change and the closest posture shift. Task segments and turn segments can have very different lengths, and we do not know under which time threshold a posture shift and an interaction state change can be related. As a rule of thumb, we compute the threshold as the quarter of the mean length for task segments and turn segments.

Finally, we set the probability for our selection rules for task change (resp. end of turn or start of turn) as the percentage of task changes (resp. end of turn or start of turn) occuring before this threshold.

TABLE IX. POSTURE SHIFTS AND SELECTION RULES PROBABILITIES

	Vid1	Vid2	Vid3
Task change shift probability	0.78	0.5	0.31
Start of turn shift probability	0.39	0.013	0.039
End of turn shift probability	0.19	0.09	0.065

3) Inter-modality influences: Though originally defined to model the effect of affective states and communicative intentions on behavior, we reuse the concept of behavior qualifiers of the Behavior Planner module [11] to specify the effect of posture on gestures spatial extent. As its values range from 0 to 1, a lower value indicating lower amplitudes, we propose to use the following formula:

$$Spatial = 1 * p(Large) + 0.5 * p(Medium) + 0 * p(Small)$$

With this formula, the Spatial parameter extracted from a recruiter's behavioral data will reflect their global tendency to use gestures with small or wide amplitude.

TABLE X. SPATIAL PARAMETER AND POSTURAL STATE

	BodyLean	BodyStraight	BodyRecline
Large[5]	54% / 47%	23% / 0 %	18% / 0 %
Medium[5]	41% / 45%	52% / 29%	62% / 58%
Small[5]	5 % / 8 %	25% / 71%	20% / 42%
Spatial Extent[5]	0.74 / 0.69	0.49 / 0.15	0.49 / 0.29

In this section, we presented methodologies for the extraction of parameters from our corpus, in order to specify an adapted version of the Semaine agent architecture.

VI. CONCLUSION

This paper presents a preliminary step for endowing virtual recruiters with the interpersonal stances expression: we investigated the presence of different behavioral influence types, namely monomodal behavior variations between recruiters (i.e. inter-recruiters variation), inter-modalities behavior influences,

and contextual influences on behavior, and provided methodologies for their specification.

We first described a coding scheme to encode multimodal behaviors and contextual information. We used the multimodal annotations to analyse which monomodal behavior variations, which inter-modalities behavior influences, and which contextual influences on behavior were significant. Then, after having adapted the Semaine agent architecture for contextual influences, we proposed methodologies for extracting specification parameters from the multimodal corpus.

In the future, we intend to explore the link between interpersonal stance and non-verbal behavior, using our corpus of annotated videos and asking users to rate the stance of the recruiters. This method will allow to specify a whole range of virtual recruiter behaviors, instead of replicating the recruiters annotated in the video, by setting higher order parameters thar are easy to relate to, such as dominance, or friendliness.

We also intend to look more into the interactional nature of the interviews by annotating the interviewees' behavior: this way, we will be able to analyze quantitatively the non-verbal reactions of an interactant with regard to the other interactant's behavior, e.g. explore mimicry or mirroring effects.

Acknowledgment

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

References

[1] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, pp. 695–729, 2005.

[2] A. Mignault and A. Chaudhuri, "The many faces of a neutral face: Head tilt and perception of dominance and emotion," *Journal of Nonverbal Behavior*, vol. 27, no. 2, pp. 111–132, 2003.

[3] J. K. Burgoon, D. B. Buller, J. L. Hale, and M. A. de Turck, "Relational Messages Associated with Nonverbal Behaviors," *Human Communication Research*, vol. 10, no. 3, pp. 351–378, 1984.

[4] D. R. Carney, J. A. Hall, and L. S. LeBeau, "Beliefs about the nonverbal expression of social power," *Journal of Nonverbal Behavior*, vol. 29, no. 2, pp. 105–123, 2005.

[5] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[6] J. Rickel and W. L. Johnson, "Embodied conversational agents." Cambridge, MA, USA: MIT Press, 2000, ch. Task-oriented collaboration with embodied agents in virtual worlds, pp. 95–122.

[7] J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich, "Non-verbal cues for discourse structure," in *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, 2001, pp. 106–115.

[8] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 477–486.

[9] J. Lee and S. Marsella, "Nonverbal behavior generator for embodied conversational agents," in *Proceedings of the 6th International Conference on Intelligent Virtual Agents*. Springer, 2006, pp. 243–255.

[10] W. Breitfuss, H. Prendinger, and M. Ishizuka, "Automated generation of non-verbal behavior for virtual embodied characters," in *Proceedings of the 9th international conference on Multimodal interfaces*, ser. ICMI '07. New York, NY, USA: ACM, 2007, pp. 319–322.

[11] M. Mancini and C. Pelachaud, "Dynamic behavior qualifiers for conversational agents," in *Proceedings of the 7th international conference on Intelligent Virtual Agents*, ser. IVA '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 112–124.

[12] E. Bevacqua, E. De Sevin, J. Hyniewska, Sylwia, and C. Pelachaud, "A listener model: introducing personality traits," *Journal on Multimodal User Interfaces, special issue Interacting ECAs*, p. 12, 2012.

[13] I. Poggi, *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*, ser. Körper, Zeichen, Kultur. Joachim Weidler Weidler Buchverlag Berlin, 2007.

[14] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction*. Cengage Learning, 2010.

[15] M. Rehm and E. André, "From annotated multimodal corpora to simulated human-like behaviors," in *Modeling communication*, I. Wachsmuth and G. Knoblich, Eds. Berlin, Heidelberg: Springer, 2008, pp. 1–17.

[16] M. Kipp, "Creativity meets automation: Combining nonverbal action authoring with rules and machine learning," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, Eds. Berlin, Heidelberg: Springer, 2006, vol. 4133, pp. 230–242.

[17] M. Foster and J. Oberlander, "Corpus-based generation of head and eyebrow motion for an embodied conversational agent," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 305–323, 2007.

[18] B. Stéphanie, A. Sarkis, R. Christophe, and J.-C. Martin, "Towards experimental specification and evaluation of lifelike multimodal behavior," in *Workshop "Embodied conversational agents - let's specify and compare them!"*, 2002, pp. 42–48.

[19] A. Akhter Lipi, Y. Nakano, and M. Rehm, "A parameter-based model for generating culturally adaptive nonverbal behaviors in embodied conversational agents," in *Proceedings of the 5th International on Conference Universal Access in Human-Computer Interaction.*, ser. UAHCI '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 631–640.

[20] D. Ballin, M. Gillies, and B. Crabtree, "A framework for interpersonal attitude and non-verbal communication in improvisational visual media production," in *1st European Conference on Visual Media Production*, 2004.

[21] J. Lee and S. C. Marsella, "Predicting speaker head nods and the effects of affective information," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 552–562, Oct. 2010.

[22] Z. Li and X. Mao, "Emotional eye movement generation based on geneva emotion wheel for virtual agents," *Journal of Visual Languages and Computing*, vol. 23, no. 5, pp. 299 – 310, 2012.

[23] M. Kipp, M. Neff, and I. Albrecht, "An annotation scheme for conversational gestures: How to economically capture timing and form," *Journal on Language Resources and Evaluation - Special Issue on Multimodal Corpora*, vol. 41, no. 3-4, pp. 325–339, 2007.

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

[25] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 273–287, 2007.

[26] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[27] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of Language Resources and Evaluation Conference (LREC*, 2006.

[28] R. Bales, *A Set of Categories for the Analysis of Small Group Interaction.Channels of Communication in Small Groups*. Bobbs-Merrill, 1950.

[29] S. Duncan and D. W. Fiske, *Interaction Structure and Strategy*. Cambridge University Press, 1985.

[30] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, ser. Psychology/cognitive science. University of Chicago Press, 1996.

[31] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. F. Valstar, and M. Wöllmer, "Building autonomous sensitive artificial listeners," *T. Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.