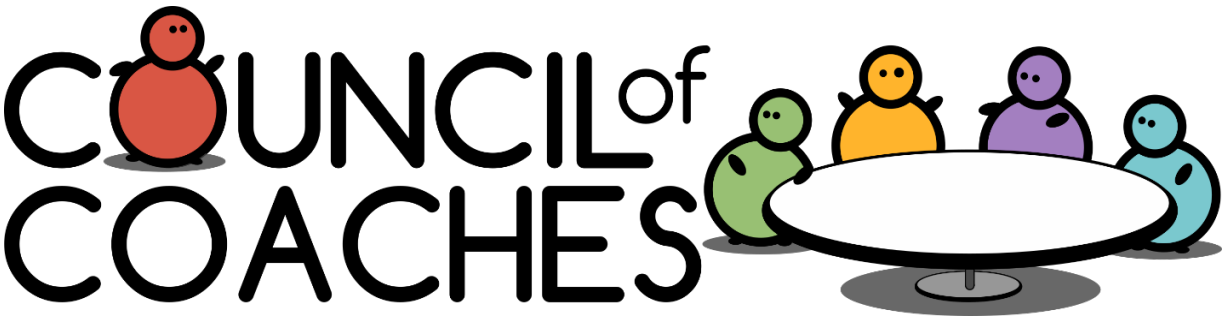

D6.4: First virtual coach design and model

Dissemination level: Public
Document type: Report
Version: 2.0.0
Date: November 23, 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Document Details

Project Number	769553
Project title	Council of Coaches
Title of deliverable	First Virtual Coach Design and Model
Due date of deliverable	February 28th 2019
Work package	WP6
Author(s)	Donatella Simonetti (SU), Fajrian Yunus (SU), Reshmashree Bangalore Kantharaju (SU), Catherine Pelachaud (SU), Harm op den Akker (RRD), Silke ter Stal (RRD)
Reviewer(s)	Merijn Bruijnes (CMC) (Original Version) Harm op den Akker (RRD), Jorien van Loon (CMC) (This Version)
Approved by	Coordinator
Dissemination level	PU: Public
Document type	Report
Total number of pages	47

Partners

- University of Twente – Centre for Monitoring and Coaching (CMC)
- Roessingh Research and Development (RRD)
- Danish Board of Technology Foundation (DBT)
- Sorbonne University (SU)
- University of Dundee (UDun)
- Universitat Politècnica de València, Grupa SABIEN (UPV)
- Innovation Sprint (iSPRINT)

Abstract

In this deliverable we present the work done so far on the development of the virtual agent's model for the Council of Coaches system. We present the result of our literature review, our proposed method, our implementation details, and our human experimentation activity. We present the state of our work as of the end of February 2019 and discuss our next steps.

Corrections

v2.0.0 This version of the deliverable was updated as a response to the project's first periodic review in April of 2019. The following changes were made:

- Improved the description of the "Evaluation study 1: First Impressions of ECAs" (see Section 5): added more details on the study method, results and discussion.
- Improved the description of the "Evaluation study 2: On the effects of agent's gender, role and focus on user's persuasion" (see Section 6): added more details on the study protocol, creation of stimuli, results and discussion.
- Added additional references, and updated the reference formatting to match the deliverable template.
- This document now uses the new, updated deliverable template format.

Table of Contents

List of figures

List of tables

Symbols, abbreviations and acronyms

ASAP	Articulated Social Agent Platform
CMC	Centre for Monitoring and Coaching
COUCH	Council of Coaches
D	Deliverable
DBT	Danish Board of Technology Foundation
EC	European Commission
ECA	Embodied Conversational Agent
GUI	Graphical User Interface
ISPRINT	Innovation Sprint
M	Month
MS	Milestone
RRD	Roessingh Research and Development
SSI	Social Signal Interpretation
SU	Sorbonne University
UDun	University of Dundee
UPV	Universitat Politècnica de València
UT	University of Twente
WP	Work Package
XML	Extensible Markup Language

1 Introduction

In this deliverable we report our development of group model, communicative gestures, gaze behaviour, and multi-agent environment. The group model development consists of three parts, namely group dynamics and cohesion, turn taking, and multi-party models. The communicative gesture development focuses on the application of machine learning to generate gestures on the fly. The gaze behaviour development is to enable the virtual agent to compute the appropriate gaze on the fly according to the saliencies. The multi-agent environment development is the software implementation to enable our virtual agent environment to have more than one agent.

In this deliverable the following studies have been conducted:

Table 1: Overview of studies reported on in this deliverable.

Study	Method	Setting	*N*	Participants <54	Participants >55	Participants with health conditions (DM-II, CP)
First impression of ECA's	Online questionnaire	Online	115	49	66	-
First impression of ECA's	Focus group	Lab Setting	13	0	13	-
On the effects of agent's gender, role and focus on user's persuasion	3 steps: Text, movie, questionnaire	Online	209	178	31	-
		Totals	**337**	**227**	**110**	**-**

2 Objectives

The objective of this deliverable is to present the work that is done on the development of a virtual coach model, that is a virtual agent able to coach humans on a variety of topics regarding their health. In summary, the aims of this deliverable are:

- To provide a description of a group conversation model.
    - Conversational capacities in dyads or in groups.
    - Theoretical models on group dynamics (mainly Cohesion)
- To provide a description of dialoguing capacities of the virtual agents.
    - Communicative gestures of the agents
    - Gaze, multi-agents
    - Example of virtual coach baseline and back-channel
- Conclusions from the evaluation studies that will be employed in the conversational model.

3 Group Model

In this section we provide an overview of the existing models that simulate group conversational dynamics and elaborate on the envisioned model for the project that focuses on group cohesion. Group cohesion is prominent when the main goal of the group is decision making or problem solving. Cohesion describes the tendency of group members' shared bond/attraction that drives the members to stay together and to want to work together (Casey-Campbell & Martens, 2009). It is a group phenomenon that emerges over time in teams (Santoro, Dixon, Chang, & Kozlowski, 2015) and a key variable for effective team performance (Beal, Cohen, Burke, & McLendon, 2003). Several existing works in literature have associated group cohesion with group performance, team satisfaction and adherence. As the goal of the project is to create a council of virtual coaches with various domain expertise and individual goals, it is important to develop agents that will be able to handle the differences in individual goals and to overcome these differences to decide and achieve the group goal. Hence, we believe cohesion is an important phenomenon for our multiparty conversation model. Further, literature on cohesion detection has shown a strong correlation with turn taking behaviours. Therefore, we will also be focusing on developing a turn-taking model with emphasis on non-verbal behaviours, i.e., gaze and feedback.

3.1 Group Dynamics and Cohesion

Everyday communication commonly takes place in groups. Group conversation is a prominent form of human communication, because often a group of humans make decisions or get ideas through information exchanges among each other (e.g. meeting, conference, council, or party). There are several works in sociology and psychology that study the various aspects of group dynamics, i.e., the action, process and changes that occur within the group (Forsyth, 2014). While research questions concerning human behaviour in groups are manifold, we focus on group cohesion since the group model involves decision making and/or problem solving.

Cohesive group in general can be defined as a group that sticks together and is accompanied by feelings of solidarity, harmony and commitment (Mudrack, 1989). Many definitions of cohesion have been presented in specific contexts such as sports team (Carron & Chelladurai, The dynamics of group cohesion in sport, 1981), group psychotherapy (Braaten, 1991). One of the earliest definitions of cohesion was proposed by Festinger et. al. as the total field of forces that act on members to remain in the group (Festinger, Schachter, & Back, Social pressures in informal groups; a study of human factors in housing, 1950). However, it was unclear as to which forces were important and to be measured. Festinger later redefined cohesion as "a resultant of forces" referring to the effects of cohesion (Festinger, Informal social communication, 1950). Several other researchers provided definitions that included "attractiveness to the group" (Back, 1951) or "commitment to the group" (Piper, Marrache, Lacroix, Richardsen, & Jones, 1983) or "commitment of members to group task" (Goodman, Ravlin, & Schminke, 1987). These authors perceived cohesion as a unidimensional construct. Later, Carron et. al., defined cohesion as "a dynamic process that is reflected in tendency of group to stick together and remain united in pursuit of its goals and objectives" (Carron, Cohesiveness in Sport Groups: Interpretations and Considerations, 1982) they looked at it as a multi-dimensional construct.

A multidimensional model was also proposed by Carron et. al., based on two dimensions, group-individual and task-social (Carron, Widmeyer, & Brawley, The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire, 1985). The group-individual distinction recognizes that cohesion results from both a member's desire to remain part of the group as a unit (group integration, GI) and from a member's personal attraction toward being a group member (interpersonal attraction to the group, ATG). The task-social distinction reflects the perceived task and social aspects of the group. Social cohesion can be defined as the interpersonal attraction among members and task cohesion can be defined as the degree to which group members work together to achieve common goals and objectives. In total, four dimensions, i.e., ATG-task, ATG-social, GI-task and GI-social were recognised. This conceptual model was proposed to measure cohesion in sport teams and is still widely accepted and used. However, some researchers presented that the empirical tests of the model outside of the sporting context yielded mixed results. Braaten proposed a five-factors model for group cohesion in group psychotherapy: attraction and bonding, support and caring, listening and

empathy, support and caring, self-disclosure and feedback, process performance and goal attainment (Braaten, 1991). Another model was proposed by Carless and De Paola (Carless & De Paola, 2000) which is a three-factor model with task cohesion, social cohesion and attraction to group. An observation of the existing models and definitions identify two constructs of cohesion the most i.e., attraction to the group or interpersonal attraction (analogous with social cohesion) and commitment to the task (analogous with task cohesion). In this task since the aim is to model a multiparty conversation between a group of virtual agents and a human user, where the agents are already a part of a council we will be focusing only on the interpersonal relationships of the members and their commitment to achieving the group goals effectively.

Hence for this work we consider the task and social dimensions where, we employ the following definitions:

- **Task Cohesion:** Attraction to group because of the shared commitment to group tasks.
- **Social Cohesion:** Attraction to group because of positive relationship with members.

3.2 Turn Taking

Turn-taking is the most naturally occurring phenomena in group conversations (Sacks, Schegloff, & Jefferson, 1978). People, in general, take turns when they involve in a conversation and these turns mostly begin and end smoothly, with short lapses of time between the turns (Cassell, Torres, & Prevost, Turn taking vs discourse structure: How best to model multimodal conversation machine conversations, 1999). Several kinds of non-verbal signals are displayed during a conversation to indicate the beginning or end of a turn. Turn yielding cues are linearly correlated with turn taking attempts, i.e. greater the number of turn yielding cues, the greater the chance that turn change will occur (Duncan, Some signals and rules for taking speaking turns in conversations, 1972) (Gravano & Hirschberg, 2011). An idealised case is, a speaker displays turn yielding signals, a listener responds to these signals and takes turn and the speaker gives the turn (Duncan, Some signals and rules for taking speaking turns in conversations, 1972). However, this is not always the case since there are overlaps, interruptions and silences (Schegloff, 2000). In this section we explain briefly about the different kinds of signals displayed and the non-verbal gestures associated with each.

Duncan (Duncan, Some signals and rules for taking speaking turns in conversations, 1972) recognizes three kinds of signal, turn yielding signals by the speaker, attempt suppressing signals by the speaker and backchannel signals by the listener. Schegloff (Schegloff, 2000) recognised different configurations of speech overlap such as Terminal Overlaps where one speaker starts to talk close to the finishing of a turn, Conditional access to the turn where a speaker of a not possibly completed turn-in-progress yields to another, or even invites another to speak in his turn's space, and choral utterance produced as convergent and consensual and not competitive (greetings) and not serially. Head movements like nod, turn, point, shake and orientation, shoulder movement like shrugs, facial expressions, hand gestures and movement, posture and posture shifts are some of the most commonly studied non-verbal cues for turn-taking (Duncan, Some signals and rules for taking speaking turns in conversations, 1972). Eye gaze is one of the most commonly studied cues in turn-taking. Kendon (Kendon, Conducting interaction: Patterns of behavior in focused encounters, 1990) states, a person tends to look away at the beginning of a long utterance, and tends to look at the interlocutor as the end of the utterance approaches in order to offer the floor. Similarly, when the listener wants the floor, s/he may look slightly up at the speaker. One may request a response from a listener by looking at the listener, and suppress the listener's response by looking away (Duncan, On the structure of speaker-auditor interaction during speaking turns, 1974). Eyebrow movements marks a new speaking turn and rapid raising/lowering of the eyebrows act as a cue to demand attention while speaking or indicate the listener wants to claim a turn (Guaïtella, Santi, Lagrue, & Cavé, 2009). Posture shifts mark syntactic boundaries inside speaking turns and aid in regulating turn taking (Hadar, Steiner, Grant, & Rose, 1984). Speakers most likely produce a posture shift when a new discourse segment is initiated at the same time as starting a turn (Cassell, Nakano, Bickmore, Sidner, & Rich, 2001). Termination of hand gesticulation or relaxation of hand position can indicate the end of a speaking turn (Duncan, Some signals and rules for taking speaking turns in conversations, 1972) and gestures like raising hand, upraised index finger act as turn requesting

signal. In the following section we present the multi-party conversational models that incorporate some of these signals for smooth turn transition.

3.3 Multiparty Models

There are several models in the literature that are enabled to handle turn-taking and interruptions in dyadic conversations (Cassell, Bickmore, Campbell, Hannes, & Yan, 2000) (DeVault, Mell, & Gratch, 2015) (Leßmann, Kranstedt, & Wachsmuth, 2004) and models that use different conversational settings e.g. presentation (interactive or non-interactive), multi-agent presentation or interaction with users (Reithinger, et al., 2006) (Rist, et al., 2004). In this section, we describe the various multi-agent models that assign turns and handle interruptions in multi-party conversations.

Padilha and Carletta (Padilha & Carletta, 2002) presented a basic model for simulation of agents engaged in group conversation that involved turn-taking and the associated non-verbal behaviours. The agents were modelled to be independent and defined by a set of attributes i.e., talkativeness, transparency, confidence, interactivity, verbosity. The agents made probabilistic decision about the display of behaviour such as speaking and listening, feedback, head nods, gestures, posture shifts, and gaze. In real life, groups are not always static and they may fragment into subgroups or merge into a larger group e. g., an agent can join an already existing conversation or two agents that are already a part of the conversation can start a new conversation of their own and form a different group. In (Jan & Traum, Dialog simulation for background characters, 2005) an algorithm that simulates behaviours and allows dynamic changes to conversational group structure is presented. The decisions are made independently by an individual agent i.e., when a character receives a message it can react immediately or update the internal state and make decision at a scheduled time. (Jan & Traum, Dynamic movement and positioning of embodied agents in multiparty conversations, 2007) presents an algorithm for simulating the dynamic movements, orientation and positioning observable in multi-party group conversations and is a continuation of the work presented in (Jan & Traum, Dialog simulation for background characters, 2005). The movement and positioning components allow agents to monitor “forces” that make it more desirable to move to one place or another while remaining engaged in conversation. This force does not directly cause any movement but provides a motivation to move.

Table 2: An overview of the existing multiparty models.

Model	Group size	Kind	Roles	States / Actions / Parameters	Factors	Handles	Non-Verbal Behaviours
(Padilha & Carletta, 2002)	>3	Agents	Speaker Addressee Side Participant	Talkativeness Transparency Confidence Interactivity Verbosity	Context	Pre-TRP cues Autonomous agents	Facial, gestures, posture, Gaze, Feedback
(Traum & Rickel, 2002)	>3	>4 Agents	Speaker Addressee Side Participant Overhearer	Take Request Release Hold Assign	Discourse, Context	Multiple floors Make or break contact	Gaze, Gestures, Proxemics
(Jan & Traum, Dialog simulation for background characters, 2005)	6	Agents	Speaker Addressee Side Participant Overhearer	Hold Release Take	Context	Multiple floors Make or break contact Pre-TRP cues	Nods, Gestures, Posture Shifts, Gaze Feedback (Pos/Neg)
(Jan & Traum, Dynamic movement	6	Agents	Speaker Addressee	Close to Speaker	Context	Multiple floors	Gestures, Postures, Gaze, Proxemics

and positioning of embodied agents in multiparty conversations, 2007)			Side Participant Overhearer	Away from Noise Proximity to Agents Circular Formation		Dynamic groups	
(Bohus & Horvitz, 2010)	>3	1 Agent 2-3 participants	Speaker Addressee Side Participant Overhearer Eavesdropper	Hold Release Take Null	Discourse, Context	Multiple floors Determines source of voice Prompts user to take turn	Gaze, Gestures, Facial
(Ravenet, Cafaro, Biancardi, Ochs, & Pelachaud, 2015)	>3	>4 Agents	Speaker Addressee Side participant	Own-the-speech Compete-for-the-speech End-of-speech Interrupted Addressed-listener Unaddressed-listener Want-to-speak	Context	Interruptions Interpersonal attitudes	Gaze, Gestures, Proxemics
(Thórisson, Gislason, Jonsdottir, & Thórisson, 2010)	12	Agents	Speaker Addressee Side Participant	(I/Other) Have Turn (I/Other) Give Turn (I/Other) Accept Turn (I/Other) Want Turn	Context	Turn signals Multi-party dialogue	Gesture Facial expression Head movements Gaze

In (Thórisson, Gislason, Jonsdottir, & Thórisson, 2010), the Ymir Turn Taking Model (YTTM) which is a computational agent-oriented model was proposed. The model has been implemented with up to 12 agents in a virtual world participating in real-time cooperative dialogue. Although it supports multi-party conversation it does not consider the expression of attitudes and dynamic group formation. In (Ravenet, Cafaro, Biancardi, Ochs, & Pelachaud, 2015) the authors presented a computational model for affective real-time turn-taking that allowed an agent to express interpersonal attitudes in a group. The multi-agent model was developed where multiple virtual agents could converse and display non-verbal behaviour, including turn-taking. Although, this model supports multi-party conversation and expression of attitudes, it does not consider the content of speech or different interruption strategies.

Ada and Grace are virtual museum guides with differing opinions and behaviours that are designed to provide useful information (Swartout, et al., 2010). The two agents interact with the user and answer the questions. The interaction between the agents is limited to sharing the responses and do not interact with each other when no one is talking with them. Gunslinger is an interactive-entertainment application where a single participant can interact verbally and non-verbally with multiple virtual characters (Hartholt, Gratch, & Weiss, 2009). A statistical text classification algorithm selects the character's responses based on the user's utterance, a visual recognition module detects and tracks the motion of the user in the physical space over time and the visual understanding module infers higher level behaviours e. g., facing a character. In Mission Rehearsal Exercise (MRE) (Traum & Rickel, 2002), Traum et. al, proposed a model for multi-party interactions in a 3D virtual environment. This model supports multi-floor dialogue. Agents can interact with a human user and as well as with each other and respond. The model consists of several layers i.e., Contact, Attention, Conversation (Participants, Turn, Initiative, Grounding, Topic, Rhetorical), Obligations, Negotiation. Each layer has an information state, dialogue acts and required signals and recognition rules. The agents can make contact by going close (eye or ear shot) and break contact by walking away. In (Bohus & Horvitz, 2010), a computational model is presented for multi-party turn-taking. It tracks the conversational dynamics to make turn decisions and

for rendering decisions about turns into an appropriate set of low-level behaviours like, coordinated gaze, gesture and speech. Barange et.al (Barange, Saunier, & Pauchet, 2017), propose a multi-party coordination model that allows several virtual and human agents to dialogue and reason to perform a collaborative task.

3.4 Future Steps

In the previous sections we presented the existing literature on cohesion turn-taking and existing multi-party models. Existing literature on cohesion detection has found correlation with turn-taking strategies, but do not provide details on specific non-verbal behaviour cues i.e., gaze, feedback signals, gestures. So, the next step is to understand the correlation of non-verbal cues with social/task cohesion. In order to do this, we are annotating non-verbal cues and cohesion on the multimodal corpus developed for the project by University of Dundee. It is a multimodal corpus of simulated consultations between a patient portrayed by an actor, and at least two medical professionals with different areas of expertise recorded by the University of Dundee for the Council of Coaches project will be used. The corpus consists of audio-visual recordings of seven sessions, involving four different medical professionals and two different actors (playing multiple personas) with a total duration of 164 minutes.

We will mainly make use of the MUMIN annotation scheme (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007) developed for the annotation of multimodal communicative behaviours to annotate various non-verbal behaviours displayed during interactions. Three communicative functions namely, feedback, turn management and sequencing are used in the MUMIN scheme. Feedback provides information about the interactions through signals e.g., facial expressions. Turn management regulates the interaction flow e.g., turn gain, turn hold. Sequencing deals with the organisation of a dialogue in meaningful sequences. The main focus will be on gaze, feedback signal and turn-taking. Additionally, we will be annotating the level of cohesion (task and social) between the participants using a set of questionnaires that will be inspired from the works in sociology (Carron, Widmeyer, & Brawley, The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire, 1985) (Carron & Brawley, Cohesion: Conceptual and measurement issues, 2000) (Braaten, 1991) and computer science (Hung & Gatica-Perez, 2010).

Utilizing the annotations, the data will be analysed to detect relevant non-verbal cues related to cohesion. Further we will distinguish the data based on social/task cohesion dimension. We will also observe the patterns of non-verbal cues (e.g. gaze, feedback) related to turn taking and frequency of non-verbal cues between segments with low and high cohesion. Based on these findings, a computational model will be developed to simulate a cohesive group of virtual agents that promote behaviour change.

4 Dialoguing capacities

4.1 Communicative gestures

Non-verbal communication involves facial expression, body posture, gaze, and gestures. Human naturally gesticulates while speaking (Iverson & Goldin-Meadow, 1998). Gesture helps the locutor to form what s/he want to convey and also helps the listener to comprehend the speech (Driskell & Radtke, 2003).

Because of how natural gesticulating is, for a virtual agent to be realistic, it must be able to generate appropriate gestures to accompany the speech. If the virtual agent is capable to interact with human, then the virtual agent should also be able to compute the appropriate gesture on the fly.

Gesture itself is known to be related to speech, although the precise relationship is complex and is still being studied. Our aim is to apply machine learning techniques to capture the relationship between gesture and speech. Speech, as we assume for this work, consists of two components: the prosody and the content.

Several taxonomies of gestures have been proposed (McNeill, 1992). We rely on McNeill's taxonomy (McNeill, 1992) that considers four types of gestures: metaphorical gesture, deictic gesture, iconic gesture, and beat gesture; metaphorical gesture is to convey an abstract concept, deictic gesture is to point at an object or a location, iconic gesture is to describe a concrete object by its physical properties, and beat gesture marks the rhythm (McNeill, 1992).

Gesture can be decomposed into several phases, namely preparation, pre-stroke-hold, stroke, post-stroke-hold, hold and retraction (McNeill, 1992). The stroke phase is obligatory while the preparation, the hold and the retraction phases are optional. Multiple consecutive gestures chain the gesture phases together. If gesture X precedes gesture Y, there might not be any retraction before gesture Y's stroke. Most of the time, gesture stroke itself happens at the same time as the speech's pitch accent or slightly before the pitch accent (Loehr, 2012).

Our aim is to endow the virtual agent to compute on the fly the gestures it should display with its speech. The text of the speech is provided by the dialog manager. The text is rendered into speech by the speech synthesizer. There are two main goals to reach for our aim. We need to know when a gesture should be triggered and we need to know the form of the gestures. At first, we tackle the first goal, namely model the link between speech prosody and gesture production.

Recurrent neural network is a machine learning technique to learn sequence. A variant of recurrent neural network has been used to generate gesture according to the speech. In the work of Hasegawa et al (Hasegawa, Kaneko, Shirakawa, Sakuta, & Sumi, 2018), the speech is represented as Mel-Frequency Cepstral Coefficients (MFCC), an abstract representation which has been used in speech recognition domain. We propose an attention-based recurrent neural network to observe the relationship between speech prosody and gesture. Speech prosody itself consists of many concrete features, such as frequency, slope, and loudness. Traditionally, psycholinguists use this prosody notion to analyse voice. We would like to study this relationship by using machine learning.

4.1.1 Related Work

There are two classes of gesture generators, namely rule based one and machine-learning based one. The rule based one works by explicitly specifying the rules governing gesture generation based on the prosody or the speech content. The machine learning model works by learning the relationship between gesture and the prosody or the speech content.

There are several rule-based generators. BEAT (Behavior Expression Animation Toolkit) (Cassell, Vilhjálmsson, & Bickmore, BEAT: The behavior expression animation toolkit, 2004) is one of the earliest models. NVBG (Non-Verbal Behavior Generator) (Lee & Marsella, 2006) uses the speech content and the agent's emotional state as its inputs. Cerebella (Lhommet & Marsella, Gesture with meaning, 2013) is an "upgrade" of the Non-Verbal Behavior Generator. Cerebella uses the speech content and prosody as the inputs. The rules are "expanded" by WordNet ontology. Rule based generators obviously need the rules, however the rules governing the relationship between gesture and speech is complex.

Machine learning based generators learn the rules from the data. Bergmann proposed a method to generate iconic gestures by using Bayesian decision network (Bergmann & Kopp, 2009). There are methods to generate metaphorical gestures based on the image schemas of the speech content (Lhommet & Marsella, Metaphoric gestures: towards grounded mental spaces., 2014) (Ravenet, Clavel, & Pelachaud, Automatic Nonverbal Behavior Generation from Image Schemas, 2018). There are also methods to generate beat gestures by using conditional random fields (Levine, Krähenbühl, Thrun, & Koltun, 2010) (Chiu & Marsella, 2014).

There is a more recent technique which uses bidirectional Long Short-Term Memory (LSTM) to generate the gesture motion based on the Mel-Frequency Cepstral Coefficients (MFCC) input (Hasegawa, Kaneko, Shirakawa, Sakuta, & Sumi, 2018). LSTM is a variant of the standard recurrent neural network. MFCC is used because it has been successfully used for speech recognition, and thus the authors expect MFCC to store richer prosodic expression. There is another recent technique which uses Conditional Random Field to generate the body motion based on the features of the written language (e.g. part of speech, bag of words, etc.) (Ishii, Katayama, Higashinaka, & Tomita, 2018). There is also an attempt of using transfer learning to transfer motion-to-motion neural network model to speech-to-motion neural network model (Ferstl & McDonnell, 2018).

Attention model is an extension of the standard recurrent neural network (Bahdanau, Cho, & Bengio, Neural machine translation by jointly learning to align and translate, 2014). The attention model is based on encoder-decoder architecture. The distinguishing feature of the attention model is a matrix called attention matrix, which is between the encoder and the decoder. The matrix is essentially a weight matrix which gives the weight of encoder at time t_1 on the decoder at time t_2. Attention model has two main advantages over the standard recurrent neural network. The first one is that it can deal with the case where the input and the output have different lengths, which standard recurrent neural network cannot do. The second one is that it shows the reasoning of how much certain input elements affect certain output elements. In fact, the attention matrix represents the weights of certain input elements on certain output elements. Attention model has also been used in speech processing domain (Bahdanau, Chorowski, Serdyuk, Brakel, & Bengio, 2016) (Chorowski, Bahdanau, Serdyuk, Cho, & Bengio, 2015).

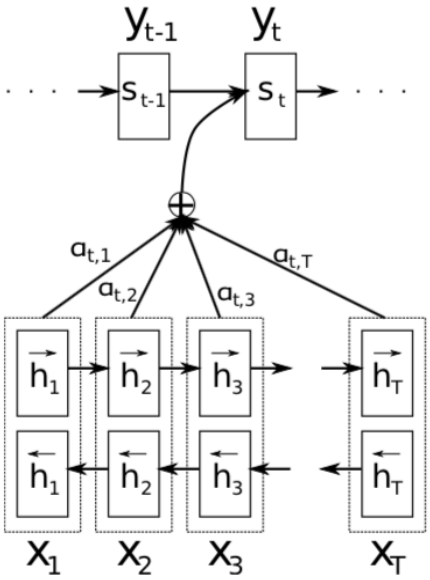

Figure 1: An illustration of attention model from (Bahdanau, Cho, & Bengio, Neural machine translation by jointly learning to align and translate, 2014). The *h* are the encoders, the *s* are decoders, the *α* are the attention matrix.

4.1.2 Proposed Method

Hasegawa used Long-Short Term Memory to generate gesture based on the voice input (Hasegawa, Kaneko, Shirakawa, Sakuta, & Sumi, 2018). The voice input is represented as Mel-Frequency Cepstral Coefficients (MFCC). MFCC representation has successfully been used in speech recognition. However, MFCC is an abstract measure. Linguists use the concept of prosody to analyse voice. Prosody itself

consists of various concrete features such as intensity and slope. Previous studies have shown a tight link between the production of gestures and prosody features. Moreover, Long-Short Term Memory is a black box model; given a certain input, it will yield a certain output, but the reasoning process is not obvious. We propose to use attention model to make the reasoning process understandable by human. We also propose to use prosody features as the input.

We start by simplifying the gesture classification to only beat gesture and other kinds of gesture. The rationale of this is because beat gestures is tightly link to speech rhythm. We also take into account only the stroke phase because this phase is known to be related to the speech's pitch accent.

To study the temporal relationship between gesture and prosody, we use recurrent neural network with attention model. Unlike the standard recurrent neural network, the attention model allows us to observe the relationship between the prosody and the gesture.

We have two datasets, the prosody dataset and the gesture dataset. We extract both of them from the Gest-IS corpus (Saint-Amand, 2018). The corpus consists of 10 dialogues between two persons. The total duration of the video is about 57 minutes. The audio, gesture phase, gesture types, and chunk boundaries are provided in the corpus. The corpus also provides the videos, classification annotations on whether the gesture is communicative or non-communicative, and the speech transcription.

We only consider the chunks when the party is speaking. After that, for each chunk, we extract the prosody by using OpenSmile (Eyben, Wöllmer, & Schuller, 2010). One measurement is taken every 10 milliseconds. We also extract the presence of gesture strokes at each time point of the prosody measurements. From the gesture data, we mark at each time point whether there is beat gesture stroke or there is other-gesture stroke, or there is no gesture stroke. Each chunk, with the corresponding prosody and gesture data, is considered as one sample.

The neural network model we use requires all samples to have the same length/duration. However, the chunks have variable lengths/duration. Therefore, we zero-pad the prosodies at the end so that all samples have the same length. For the gesture markers in those padded sections, we mark them as not having any gesture stroke.

The samples are fed into the attention-based recurrent neural network. The encoder is a bidirectional Long Short-Term Memory (LSTM) while the decoder is a unidirectional LSTM. Attention model allows the input and the output to have different length. However, in our particular case, we infer the gesture stroke class for each time we measure the prosody, and therefore the input sequence and the output sequence have the same length.

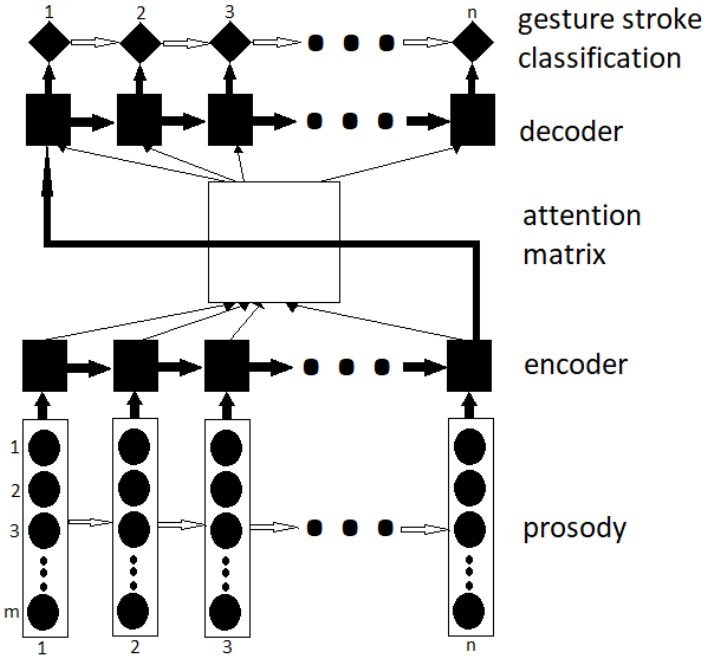

Figure 2: An illustration of the proposed method with *m* prosody features and n time steps.

We measure the performance of our classifier by calculating its F_1 score. Our classifier has already achieved the average F_1 score of 0.425. Comparison with other existing works is difficult as they rely on other corpora which are not available. Therefore, we conducted a preliminary study by comparing the performance of the neural network against the performance of random result generator with the prior knowledge of the class distribution. The random result generator yields an average F_1 score of 0.332 with the standard deviation of 0.003. Therefore, the classifier has outperformed random result generator.

4.1.3 Future Work

We are continuing improving the classification performance. We are working to extend the attention model so that we can also observe the effect of an individual prosody feature on the gesture stroke classification. We want to know if, for example, a certain prosody feature (e.g. loudness) has a very large impact on the gesture class for the next n milliseconds while certain other prosody features (e.g. slope) has only a little impact.

The overall objective of this work is performing automatic gesture generation. At the current phase, we are solving the question on when to perform gestures. The next phase of the work will be predicting the shape of the gesture. Our model will be integrated in the Council of Coaches platform taking in input the speech to be said by the agent from Flipper and computing the corresponding gestures.

4.2 Gaze behaviour

In this section we are going to explain the gaze behaviour presented in the Greta platform to gaze at the user in real-time and the functional workflow followed to obtain this behaviour.

The agent can gaze at the user in real-time if his/her head position is continuously provided. The SSIGazeToSignal module takes as input the position and orientation of the user's head and generates GazeSignals with the attribute target="user". When the SSIGazeToSignal module is added to the configuration, a node (id="user") is created in the environment. Once the module starts to receive the info about the user's head position and orientation, those are stored in the user node.

The GazeSignal generated by the SSIGazeToSignal module is sent to the BehaviorRealizer that can access to the environment. So the BehaviorRealizer knows the user's head position and computes the rotation angle to look at the user. The gaze behaviour of the agent is then translated into animation

parameters BAP (if influence > eyes)/FAP using the algorithm described in D6.3. The animation parameters are sent to the respectively FAP/BAP performers and finally to the MPEG4 player.

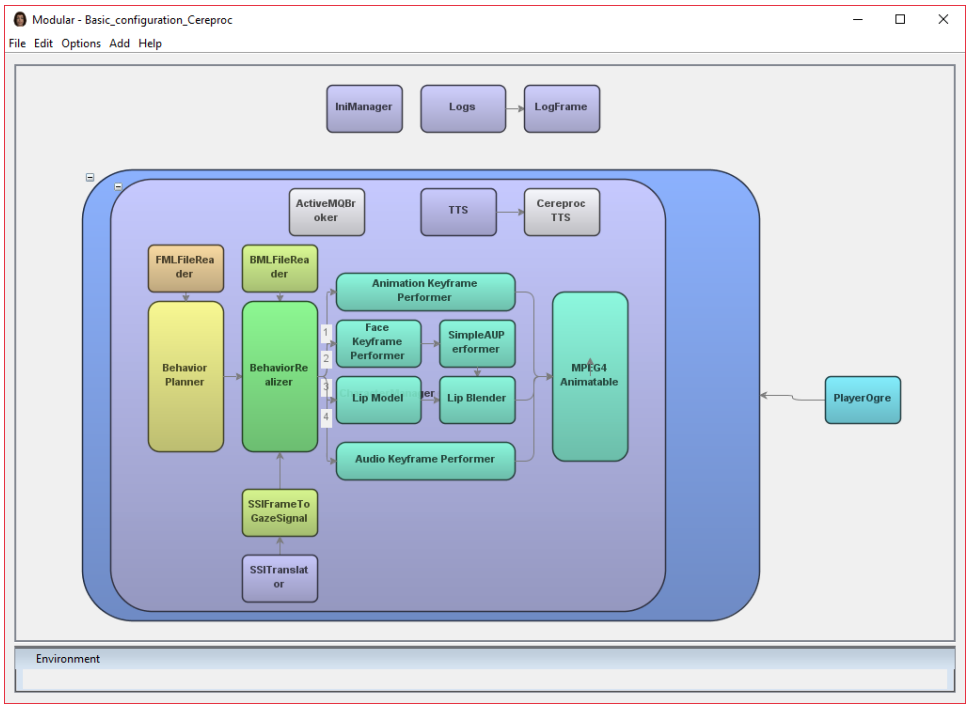

Figure 3: Example of basic configuration including the modules (SSIGazeToSignal and SSITranslator) necessary to gaze at the user.

The user's head information can be provided by a simple camera or a Kinect. The connection between camera/Kinect and Greta platform is made via EyesWeb/SSI and ActiveMQ. EyesWeb/SSI take the info about the user, put them in the right format (same xml file used for the back-channel, Figure 6) and send it to Greta via ActiveMQ.

With the camera the head positions are given with respect to the position of the camera itself. We should consider the difference in positions between the camera and the eyes of the agent. To make the coordinate system of camera match with the coordinate system of the agent, a GUI can be used, that will appear after clicking on SSIGazeToSignal module. The interface allows one to set the position of the camera and its orientation with respect to the screen where the agent is visible (see Figure 4).

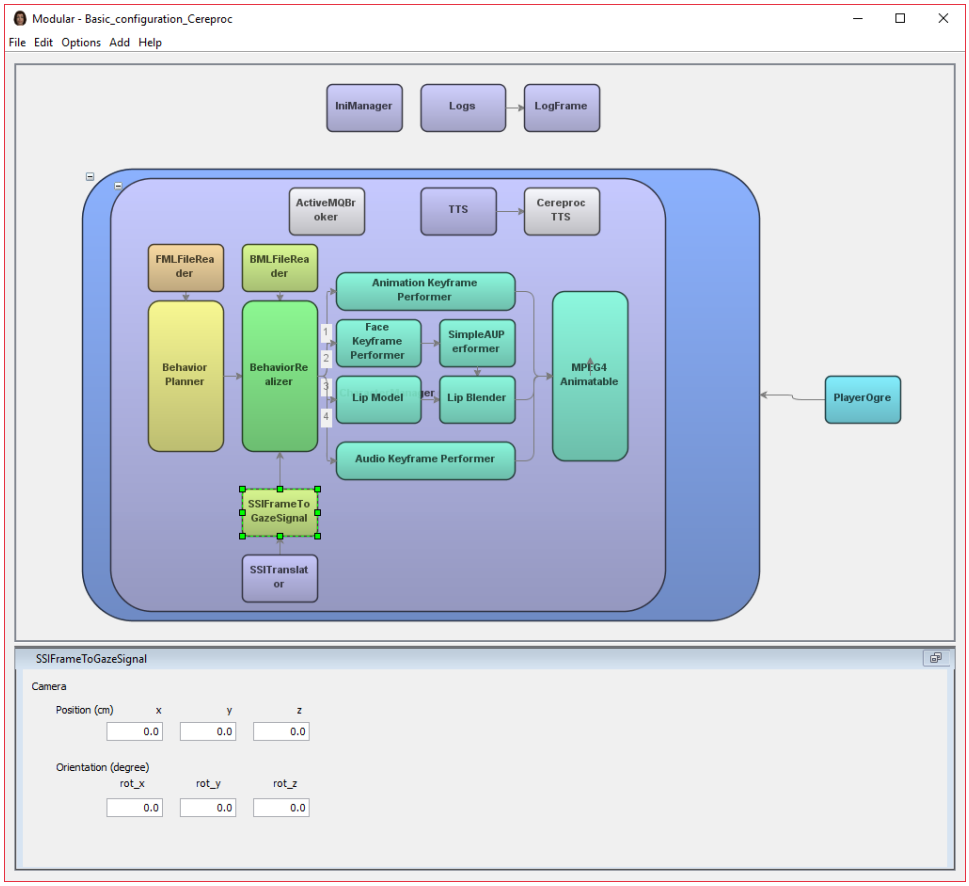

Figure 4: Example of basic configuration including modules to gaze at the user and interface to set the camera position.

4.2.1 Multi-agents gaze behaviour by Flipper for ASAP

We created a social saliency gaze model for the agents in the Council of Coaches system. The model is implemented in Flipper templates. The ASAP controlled agents will gaze in a social way at the speaker during the dialogue. The model will update the gaze targets of the agents. Targets can become more or less salient. When one of the agents starts speaking for example, it becomes more salient so the other agents will gaze at this agent. The BML gazeshift tag is used to update the gaze targets of the agents. The head of the agents is used as gaze target in this model. After a gazeshift the saliency of the agent belonging to the previous gaze target will become less salient to prevent gaze behaviour that returns too quickly to that agent. In this model we keep track of which agent is speaking, how many times the agent was speaking during the conversation, and where each agent is gazing at. Based on the prediction of the speech, who is speaking, when does it start, and when does it end, the saliency of the agents is updated. First steps in integrating the GRETA agents were made by listening to the BML feedback stream from the GRETA platform.

4.2.2 Back-channels

In this section we introduce first the concept of back-channel and then we describe how we integrate the back-channels model developed by Bevacqua (Bevacqua, Heylen, Pelachaud, & Tellier, 2007) into the GRETA platform.

During a conversation, all participants, whether they are speakers or addressees, are active. The addressees produce communicative intentions in order to provide feedbacks on the recognition of communicative behaviour of the current speaker.

The term back-channel term can be used to identify the feedbacks exchanged during a conversation between speaker and addressee to express intentions of perception, attention, interest, understanding, attitude or acceptance. Therefore, the participants can receive information on the successfulness of the communication.

The channels used to provide these feedbacks can be voice, head, face, gaze, posture and gesture.

To produce the signals that the agent provides to the speaker to specify its communicative intention, the listener's mental state and the behaviour tendencies is considered in the back-channel model. With behaviour tendencies, or baseline, we mean it is the agent's specific way to produce the non-verbal behaviour signals. Instead the agent's *mental state* is how the ECA reacts towards the interaction, that is how the agent reacts to the user's speech (if it agrees / refuses / understands what is being said).

The ECA should be able to decide when to trigger a back-channel and select which communicative intentions it wants to transmit through that signal.

The probability that user behaviour provokes a backchannel signal from the agent depends on the user's estimated level of interest. When the interest level decreases the user might want to stop the conversation, consequently the agent provides low frequency of backchannels.

When a backchannel is triggered, the agent's mental state decides which communicative intentions the agent has to convey.

4.2.3 Back-channels in Greta platform

In the Figure 5, it is shown a basic configuration to which is added modules needed to have the agent do back-channels, namely: a ListenerIntentPlanner, SSIXMLToFrameTranslator and SSIFrameToSignalTranslator modules. The last two modules allow recognizing user's verbal and nonverbal behaviours and sending them to the ListenerIntentPlanner module.

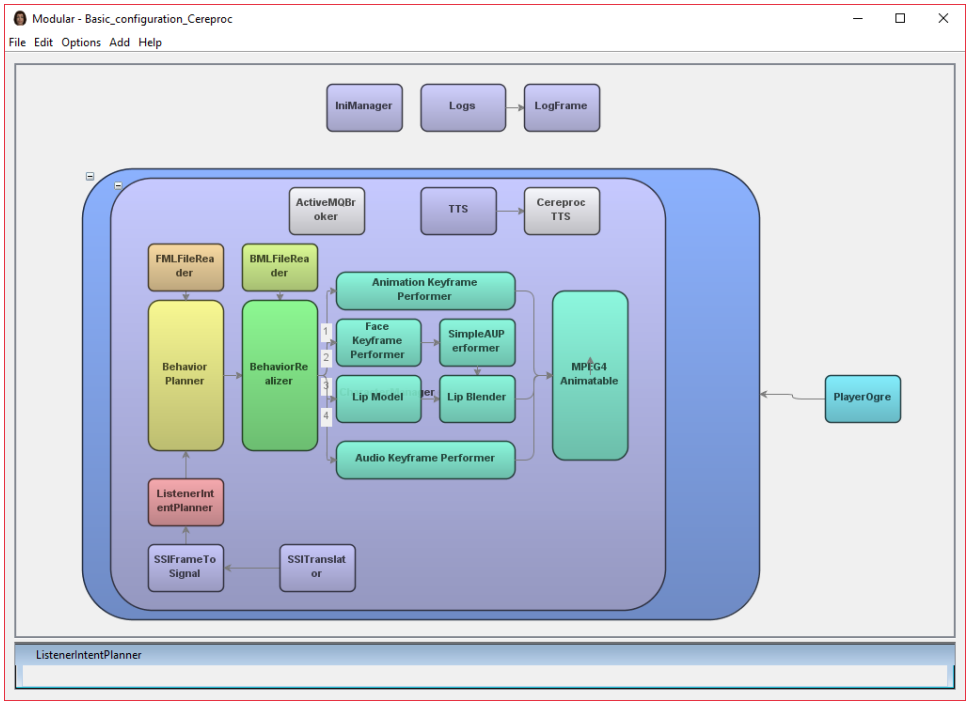

Figure 5: Example of basic configuration including the modules (ListenerIntentPlanner, SSIFrameToSignal, SSITranslator) necessary to use the back-channel model.

The SSIXMLToFrameTranslator receives an xml message (Figure 6) from SSI/EyesWeb with detected user's information. Audio-visual signals are captured from user's behaviours (via camera and microphone). In particular acoustics cues linked to prosody are gathered. Head and gesture positions are also stored.

This information contained in xml message are the inputs that the Greta platform needs to trigger the back-channels.

The file can be created via SSI or EyesWeb and then sent to GRETA via ActiveMQ. The SSIXMLToFrameTranslator receives the xml (the receiving frequency can be set in SSI or EyesWeb) and translate it in frames, according to the type of information provided, i.e. headframe, prosodyframe or

bodyframe. The frames are sent to SSIFrameToSignal module in order to be translated in signals (headSignal, audioSignal, AUSignal, etc.). The Signal produced are defined UserSignals, i.e. translation in Greta platform language of what the user just performed. The UserSignals are the key to trigger the back-channels.

The ListenerIntentPlanner receive the UserSignals, check if all of these are in any of the rule (example of rules in Figure 7) and in positive case, according to the rule, can trigger signals to be perform by the agent.

As shown in Figure 7, each rule is divided in two parts: UserSignals and back-channels. In the first part there are all the signals perform by the user necessary to trigger the back-channels. The back-channels linked to those usersignals are listed in the second part. Here the back-channels can be a mimicry_signals or response_reactive. For both we can specify the probability the agent can select this back-channel type. For the mimicry_signal we have also the attributes name and modality that specify respectively the name of signal (the same name that we found in the gesture or facial expression libraries) and the modality (i.e. head, gaze, torso, face).

```
<?xml version="1.0" encoding="UTF-8"?>
<ssi>
    <prosody>
        <voice activity="1">
            <speech></speech>
            <laughter></laughter>
        </voice>
        <praat>
            <feature name="Pitch median (Hz)"></feature>
            <feature name="Pitch mean (Hz)"></feature>
            <feature name="Pitch sd (Hz)"></feature>
            <feature name="Pitch min (Hz)"></feature>
            <feature name="Pitch max (Hz)"></feature>
            <feature name="Pulses number"></feature>
            <feature name="Pulses per sec (pulses/sec)"></feature>
            <feature name="Periods number"></feature>
            <feature name="Period mean (sec)"></feature>
            <feature name="Period sd (sec)"></feature>
            <feature name="Fraction locally unvoiced frames (%)"></feature>
            <feature name="Voice breaks number"></feature>
            <feature name="Voice breaks degree (%)"></feature>
            <feature name="Jitter local (%)"></feature>
            <feature name="Jitter local abs (sec)"></feature>
            <feature name="Jitter rap (%)"></feature>
            <feature name="Jitter ppq5 (%)"></feature>
            <feature name="Jitter ddp (%)"></feature>
            <feature name="Shimmer local (%)"></feature>
            <feature name="Shimmer local (dB)"></feature>
            <feature name="Shimmer apq3 (%)"></feature>
            <feature name="Shimmer apq5 (%)"></feature>
            <feature name="Shimmer apq11 (%)"></feature>
            <feature name="Shimmer dda (%)"></feature>
            <feature name="Harmonicity mean autocor"></feature>
            <feature name="Harmonicity mean noise-to-harmonics ratio"></feature>
            <feature name="Harmonicity mean harmonics-to-noise ratio (dB)"></feature>
            <feature name="Speechrate duration (sec)"></feature>
            <feature name="Speechrate voiced count"></feature>
            <feature name="Speechrate (syllables/sec)"></feature>
            <feature name="Intensity minimum (dB)"></feature>
            <feature name="Intensity maximum (dB)"></feature>
            <feature name="Intensity median (dB)"></feature>
            <feature name="Intensity average (dB)"></feature>
        </praat>
        <opensmile>
            <feature name="Pitch"></feature>
            <feature name="PitchDirection"></feature>
            <feature name="Energy"></feature>
            <GenevaMinimalFeatureSet>
                <feature name="F0semitoneFrom55Hz_sma3nz_amean"></feature>
                <feature name="F0semitoneFrom55Hz_sma3nz_stddevNorm"></feature>
                <feature name="F0semitoneFrom55Hz_sma3nz_percentile20"></feature>
                <feature name="F0semitoneFrom55Hz_sma3nz_percentile50"></feature>
                <feature name="F0semitoneFrom55Hz_sma3nz_percentile80"></feature>
                <feature name="F0semitoneFrom55Hz_sma3nz_pctlrange0-2"></feature>
                <feature name="StddevUnvoicedSegmentLength"></feature>
            </GenevaMinimalFeatureSet>
        </opensmile>
        <keyword>hello</keyword>
    </prosody>
    <head>
        <headposition>
            <xpos>15.244831</xpos>
            <ypos>0.000000</ypos>
        </headposition>
        <headorientation>
            <pitch>302.401489</pitch>
            <roll>234.510056</roll>
            <yaw>0.000000</yaw>
            <headfocus>1.000000</headfocus>
            <headtilt>0.000000</headtilt>
        </headorientation>
        <headnod>0.000000</headnod>
        <headshake>0.000000</headshake>
        <smile>-91.000000</smile>
    </head>
    <body>
        <leanposture>0.000000</leanposture>
        <openness>0.000000</openness>
        <overallactivity>0.000000</overallactivity>
        <energy>0.000000</energy>
        <gesture name="ArmsOpen"></gesture>
        <gesture name="ArmsCrossed"></gesture>
        <gesture name="LeftHandHeadTouch"></gesture>
        <gesture name="RightHandHeadTouch"></gesture>
        <gesture name="LeanFront"></gesture>
        <gesture name="LeanBack"></gesture>
    </body>
</ssi>
```

Figure 6: Example of XML received by the SSITranslator with user information.

The ListenerIntentPlanner computes the backchannel signals that the agent provides while listening. This module implements three types of backchannels:

1. Reactive: derive from a first process of perception of the speaker's speech and they show contact and perception;
2. Response: are generated by a more aware evaluation that comprehends memory and cognitive process;
3. Mimicry: derive from the imitation of the speaker's behaviour

Responsive/Reactive back-channels are link to agent's mental state to decide which communicative functions the agent should convey.

The mimicry module determines which signals would mimic the agent. So far, we are considering solely speaker's head movement (nod, shake) and some facial-expressions(smile) in the signals to mimic.

```
<rule name="trigger-head_nod">

    <usersignals>
        <usersignal id="1" name="nod" modality="head"/>
    </usersignals>

    <backchannels probability="0.1" priority="2">
        <mimicry probability="0.1">
            <mimicry_signal name="head=Nod_Middle" modality="head"/>
        </mimicry>

        <response_reactive probability="0.1"/>
    </backchannels>
</rule>


<rule name="trigger-head_shake">

    <usersignals>
        <usersignal id="1" name="shake" modality="head"/>
    </usersignals>

    <backchannels probability="0.1" priority="2">
        <mimicry probability="0.1">
            <mimicry_signal name="head=Shake_Middle" modality="head"/>
        </mimicry>

        <response_reactive probability="0.1"/>
    </backchannels>
</rule>
```

Figure 7: Example of rules takes into account in order to trigger the back-channels.

The agent's mental state in the GRETA platform is an xml file where can be specified the communicative intention. All of them are listed in the xml file example of Figure 8 below.

```
<?xml version="1.0"?>

<profile xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:noNamespaceSchemaLocation="../../Common/Data/xml/profiles.xsd">

    <parameter-set name="listener_function">
        <Parameter name="disagreement" value="0.0"/>
        <Parameter name="agreement" value="0.0"/>
        <Parameter name="acceptance" value="0.0"/>
        <Parameter name="refusal" value="0.0"/>
        <Parameter name="belief" value="0.0"/>
        <Parameter name="disbelief" value="0.0"/>
        <Parameter name="liking" value="0.0"/>
        <Parameter name="disliking" value="0.0"/>
        <Parameter name="interest" value="0.0"/>
        <Parameter name="no_interest" value="0.0"/>
        <Parameter name="understanding" value="0.0"/>
        <Parameter name="no_understanding" value="0.0"/>

        <Parameter name="anger" value="0.0"/>
        <Parameter name="sadness" value="0.0"/>
        <Parameter name="amusement" value="0.0"/>
        <Parameter name="happiness" value="0.0"/>
        <Parameter name="contempt" value="0.0"/>
        <Parameter name="low-anticipation" value="0.0"/>
        <Parameter name="high-anticipation" value="0.0"/>
        <Parameter name="low-solidarity" value="0.0"/>
        <Parameter name="high-solidarity" value="0.0"/>
        <Parameter name="low-antagonism" value="0.0"/>
        <Parameter name="high-antagonism" value="0.0"/>
    </parameter-set>

    <!-- seems to be not used -->
    <parameter-set name="interest_level">
        <Parameter name="interest" value="0.5"/>
    </parameter-set>


    <!-- seems to be not used -->
    <parameter-set name="mimicry_level">
        <Parameter name="mimicry" value="0.0"/>
    </parameter-set>


</profile>
```

Figure 8: Example of agent's mental state file.

According to the agent's persona the type of back-channels can change. For example, if an agent is happy and positive, it tends to communicate mainly through facial expression and tends to provide Back-channel signals that are the expression of positive communicative intentions, such as liking, acceptance and interest. Instead, an agent that is gloomy and sad tends to produce back-channels signals mostly with the head and tends to convey negative communicative intentions, in particular disbelief, refusal and no understanding. Or an agent that is pragmatic and sensitive tends to perform slow movements mainly on the head and face modalities and conveys positive communicative intentions, in particular agreement, belief and understanding.

4.2.4 Baseline

In this section we present the concept of baseline and it is explained its parameters.

The Baseline of an agent is defined as a set of fixed parameters (Expressivity Parameters) that represent the agent's general, underlying behaviour tendencies. It is a static global parameter.

The expressivity parameters allow you to adjust how a gesture sequence is performed in Greta to alter its style. The parameters operate at two levels. There are parameters that are applied to all motion that is generated. These are called baseline. There are also parameters that are applied only to a particular intention (convey a concept, appear sad, etc.). These are the dynamicline. The final appearance of a gesture will be the result of applying the sum of the baseline and dynamicline expressivity parameters to it.

Here are the definitions of the expressivity parameters from (Mancini & Pelachaud, 2009). Most parameters are defined on [-1,1]:

1. Overall Activity - **OAC**: amount of activity (e.g., passive/static versus animated/engaged). This parameter influences the number of single behaviours occurring during the communication. For example, as this parameter increases, the number of head movements, facial expressions, gestures and so on, increases. Its value is a floating point number ranging from 0 to 1 where a value of zero corresponds to no activity, and a value of one corresponds to maximum activity. So, if the value is less than 1.0, your gesture may not be executed.
2. Spatial Extent - **SPC**: amplitude of movements (e.g., expanded versus contracted). This parameter determines the amplitude of, for example, head rotations and gestures. The attribute, like all the following, is a real number defined in the interval [−1,1]. A value of zero corresponds to a neutral behaviour, that is, the behaviour of the agent without any expressivity control; in such a case, the agent performs nonverbal signals with the amplitude that was defined by the system designer. A value of −1 corresponds to the reproduction of very small and contracted movements, while value of 1 corresponds to very wide and large movements.
3. Temporal Extent - **TMP**: duration of movements (e.g., quick versus sustained actions). This parameter modifies the speed of execution of movements. They are slow if the value of the parameter is negative, or fast when the parameter is positive. The effects of the TMP parameter on the calculation of the agent's movements is different depending on the involved modalities.
4. Fluidity - **FLD**: smoothness and continuity of movement (e.g., smooth, graceful versus sudden, jerky). Higher values allow smooth and continuous execution of movements while lower values create discontinuity in the movements.
5. Power - **PWR**: dynamic properties of the movement (e.g., weak/relaxed versus strong/tense). Higher (resp. lower) values increase (resp. decrease) the acceleration of the head or limbs rotation, making the overall movement look more (resp. less) powerful. Increasing this parameter also produces movement overshooting.
6. Repetitivity - **REP**: this parameter permits the generation of rhythmic repetitions of the same rotation/expression/gesture. For example, a head nod with a high amount of repetition becomes a sequence consisting of very fast and small nods.

```
<?xml version="1.0"?>
<profile xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="../../Common/Data/xml/profiles.xsd">
    <parameter-set name="face">
        <Parameter name="preference" value="0.9"/>
        <Parameter name="OAC" value="0.95"/>
        <Parameter name="SPC" value="0.0" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.05" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="-0.05" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0.1" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="-0.1" min="-1.0" max="1.0"/>
    </parameter-set>
    <parameter-set name="gesture">
        <Parameter name="preference" value="0.9"/>
        <Parameter name="OAC" value="0.95"/>
        <Parameter name="SPC" value="0.00" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.05" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="-0.5" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="-0.5" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="-0.1" min="-1.0" max="1.0"/>
    </parameter-set>
    <parameter-set name="gaze">
        <Parameter name="preference" value="0.9"/>
        <Parameter name="OAC" value="0.95"/>
        <Parameter name="SPC" value="0.0" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.05" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="-0.05" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0.1" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="-0.1" min="-1.0" max="1.0"/>
    </parameter-set>
    <parameter-set name="torso">
        <Parameter name="preference" value="0.9"/>
        <Parameter name="OAC" value="0.95"/>
        <Parameter name="SPC" value="0.00" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.05" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="-0.05" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0.1" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="-0.1" min="-1.0" max="1.0"/>
    </parameter-set>
    <parameter-set name="head">
        <Parameter name="preference" value="0.9"/>
        <Parameter name="OAC" value="0.95"/>
        <Parameter name="SPC" value="0.0" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.05" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="-0.05" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0.1" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="-0.1" min="-1.0" max="1.0"/>
    </parameter-set>
    <parameter-set name="speech">
        <Parameter name="preference" value="1"/>
        <Parameter name="OAC" value="1"/>
        <Parameter name="SPC" value="0" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="0" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="0" min="-1.0" max="1.0"/>
    </parameter-set>

    <parameter-set name="pointing">
        <Parameter name="preference" value="1"/>
        <Parameter name="OAC" value="1"/>
        <Parameter name="SPC" value="0.5" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.5" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="0" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="0" min="-1.0" max="1.0"/>
    </parameter-set>

    <parameter-set name="shoulder">
        <Parameter name="preference" value="0.9"/>
        <Parameter name="OAC" value="0.95"/>
        <Parameter name="SPC" value="0.00" min="-1.0" max="1.0"/>
        <Parameter name="TMP" value="0.05" min="-1.0" max="1.0"/>
        <Parameter name="FLD" value="-0.05" min="-1.0" max="1.0"/>
        <Parameter name="PWR" value="0.1" min="-1.0" max="1.0"/>
        <Parameter name="REP" value="-0.1" min="-1.0" max="1.0"/>
    </parameter-set>
</profile>
```

Figure 9: An example of a baseline file.

4.2.5 Mapping Agents' personality

In this section, we present an example of the personality traits for two of the seven virtual coaches developed and the mapping of the agent's personalities to create a baseline for behaviour expressivity and interaction sensibility. Table 2 presents the personality traits and in Table 3 we present the mapping based on the personality traits.

Table 3: Personality traits for two virtual coaches.

Traits	Social Coach: Emma	Physical Coach: Olivia
Openness	Medium	High
Conscientiousness	High	Medium
Extraversion	High	Medium
Agreeableness	High	Medium
Neuroticism	Low	Low

Table 4: An example of specified parameters for baseline and backchannel.

		Social Coach: Emma	Physical Coach: Olivia
Baseline	SPC	0.8	0.7
	TMP	0.5	0.5

	PWR	0.9	0.4
	OAC	1.0	0.6
	FLD	0.3	0.7
	REP	0.0	0.0
Back-channel	disagreement	0.0	0.2
	agreement	1.0	0.8
	acceptance	0.5	0.6
	refusal	0.5	0.4
	belief	0.7	0.7
	disbelief	0.3	0.3
	liking	0.9	0.7
	disliking	0.1	0.3
	interest	0.8	0.8
	no_interest	0.2	0.2
	understanding	0.7	0.8
	no_understanding	0.3	0.2

4.3 Multi-agents

From what was written in the previous deliverable, 6.3, the way to build a basic configuration with one or more agents in the same environment has changed. As before, to add any modules related to the agent the Environment module must be placed. It is the parent module for the CharacterManager module that is, in its turn, the parent module of all the other modules used to create a character.

A warning has been introduced in order to help the user to add the module in the right order and so a message box will appear when a module is added in the wrong way.

The basic configuration now is displayed like in the Figure 10. The Environment is the blue module in the background, parent module of CharacterManager (CM) module, in purple. The CM is the parent module of all the other modules that has inside.

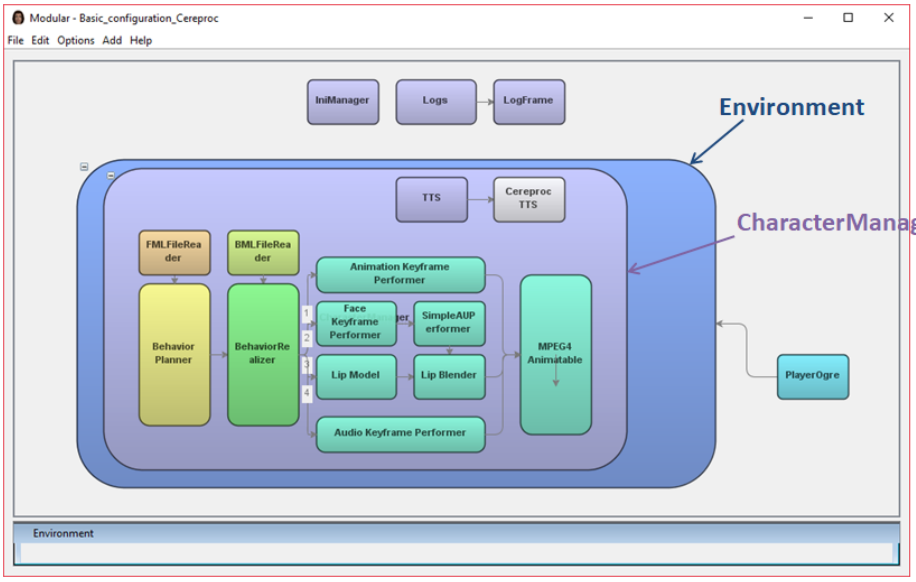

Figure 10: Example of new basic configuration.

5 Evaluation study 1: First Impressions of ECAs

This first evaluation study looks at the effect of age, gender and role on users' first impressions of embodied conversational agents in eHealth. The study has been submitted for publication as follows:

> (ter Stal, Tabak, op den Akker, Beinema, & Hermens, 2019 (Submitted)) ter Stal, S., Tabak, M., op den Akker, H., Beinema, B., & Hermens, H. (2019 (Submitted)). Who Do You Prefer? The Effect of Age, Gender and Role on Users' First Impressions of Embodied. International Journal of Human Computer Interaction.

5.1 Objectives

The aim of this research is to investigate which design features establish a positive first impression of an agent in the eHealth domain. We are specifically interested in how the agent's gender, role and age affect the first impressions.

5.2 Methods

Two different methods were used:

1) **Questionnaire**

First, we investigated the impressions of a general population using an online questionnaire.

2) **Focus groups**

Second, we investigated impressions of an elderly population by means of focus groups.

5.2.1 The Agent Designs

The agents differed on three features: the agent's age (young or old), the agent's gender (male or female) and the agent's role (expert or peer). Combinations of all variations were tested, leading to a set of eight agent permutations. The individual agent designs are shown in Figure 11.

Figure 11: The agents subjected to testing, differing in gender, age and role.

5.2.2 The Participants

Questionnaire

Respondents to the questionnaire should be fluent in the Dutch or the English language. No other inclusion or exclusion criteria were set. We recruited the respondents via a Dutch panel of adults that indicated they were interested in participating in research on eHealth and through snowball sampling via social media and personal connections. The questionnaire was accessible via a public link of the survey program Qualtrics and available for two months, in July and August 2018.

In total, 155 people participated in the online questionnaire, of which 115 people filled out the complete questionnaire. The age of the population that filled out the complete questionnaire ranged from 17 to

87 years (M = 51.36 years, SD = 20.71, 22 unknown) and 69 were female and 67 were male (19 unknown). In total, 66 of the participants of this study fall within the target population for Council of Coaches (aged 55 and over), while 49 participants where younger than 55.

Focus group

Participants in the focus groups should be aged 55 years or above and fluent in the Dutch language and were recruited via the same panel of adults. The focus groups were performed in July 2018.

Thirteen people (N = 13) participated in the focus groups. The age of the participants ranged from 58 to 81 (M = 71.23 years, SD = 6.82). Five males and eight females participated. In the first focus group, four males and four females (N = 8, M = 73.13 years, ages ranged from 58 to 81 years) participated. In the second focus group, one male and four females participated (N = 5, M = 68.20 years, ages ranged from 61 to 75 years).

5.3 Measurements

Questionnaire

The following data were collected via the online questionnaire:

- Characteristics of respondents (age, gender, education, housing status, technology literacy, health literacy and stage of change in nutrition and physical activity).
- Preferred agent design at first glance.
- For each agent design: likeliness of following the agent's advice.
- For each agent design: ratings of importance of five agent characteristics: friendliness, trustworthiness, involvement, expertise and authority.

Focus group

The following data were collected via the focus groups:

- Ratings of importance for a set of twenty predefined agent characteristics (hair color, skin color, clothing, gender, age, voice, language usage, humor, intelligence, reliability, cultural background, political preferences, posture, role, shape, friendliness, expertise, authority, involvement and hobbies)
- Explanation of the rating of the importance of all agent characteristics.

5.4 Procedure

Questionnaire

The questionnaire consisted of three parts. The first part consisted of questions on the characteristics of the respondent. In the second part, the eight agent designs were shown to the respondent simultaneously. Then, the respondent selected one of the designs as his or her preferred design at first glance. The position of the various agents on the screen was randomized to avoid any bias. In addition, the respondent had the opportunity to state the rationale behind his or her preference in a text box. In the final part of the questionnaire, each agent was shown individually. For each agent, the respondent rated the likeliness of following the advice of the agent on a 7-point Likert scale. Also, he or she was asked to rate the five characteristics of the agent on a 7-point Likert scale. The order in which the individual agent designs were shown was randomized.

Focus Groups

The participants individually performed a card-sorting task. Participants received a set of cards with twenty key characteristics of agents and a few blank cards. In addition, they received a sheet with the title: "My ideal coach is. . . " with two columns below, labelled "important" (left) and "less important" (right). Participants were asked to place the cards of characteristics they perceived as important at the left and the cards of characteristics they believed were less important at the right. After the card-sorting

task, each participant explained which characteristics he or she believed were important and why. Participants were then encouraged to respond to each other in a general discussion.

5.5 Results

5.5.1 Preference Agent Designs at First Glance

Figure 12 shows the frequencies of the preference for the agent designs at first glance. Overall, these results indicate a high preference for the young female agents and a low preference for the old male agents.

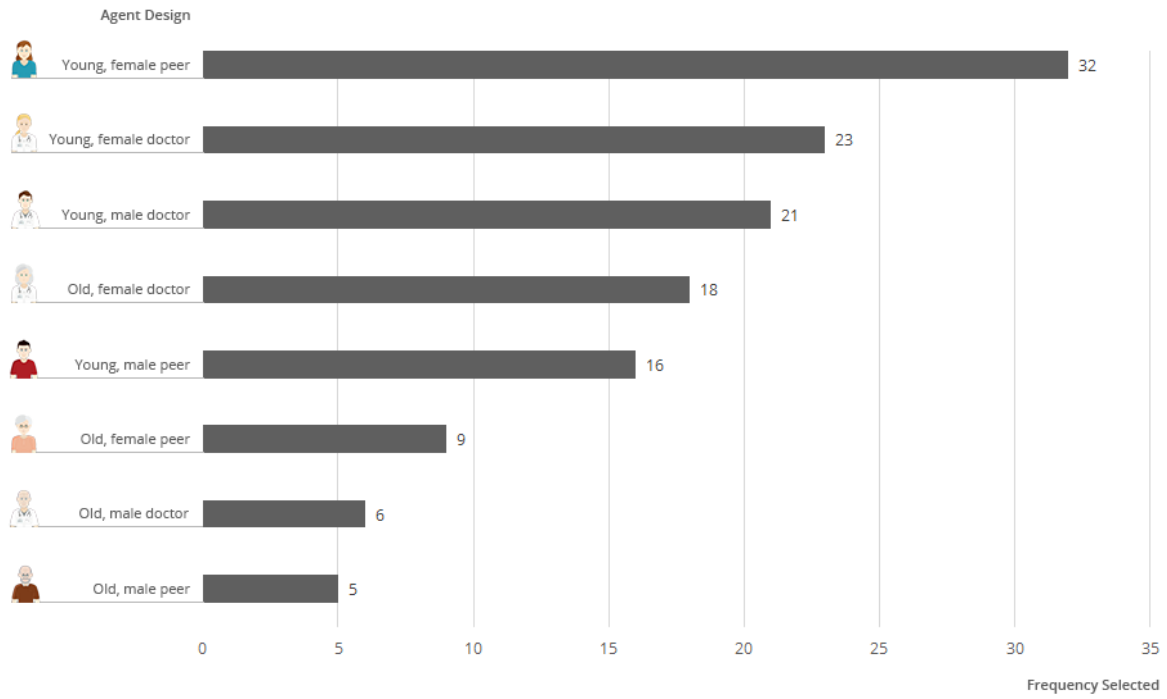

Figure 12: Frequencies of the agent designs preferred at first glance.

5.5.2 Comparison Perceived Characteristics Agents Designs

For each agent characteristic (friendliness, expertise, reliability, authority and involvement) and the likeliness of following advice, the mean ratings were compared for both agent feature categories (i.e. young agents vs old agents, female agents vs male agents and peer agents vs expert agents). Table 5 shows the results.

Table 5: Results of the paired-samples t-tests (N = 117) comparing the mean ratings of the two categories for the agent features (age, gender and role) for each of the agent characteristics. The range of the mean ratings is from 1 (strongly disagree) to 7 (strongly agree). The relations for which the p-values are shown in bold are statistically significant.

	Agent Age			**Agent Gender**			**Agent Role**		
	Young, M(SD)	*Old, M(SD)*	*p*	*Female, M(SD)*	*Male, M(SD)*	*p*	*Peers, M(SD)*	*Experts, M(SD)*	*p*
Likeliness of following advice	4.78(1.127)	4.35(1.150)	**<0.001**	4.64(1.150)	4.49(1.163)	0.063	3.69(1.353)	5.16(1.076)	**<0.001**
Friendliness	5.50(0.880)	4.84(1.038)	**<0.001**	5.28(0.936)	5.06(0.882)	**<0.001**	5.20(1.047)	5.14(0.873)	0.464
Expertise	4.79(0.960)	4.62(0.960)	**0.038**	4.63(0.891)	4.77(0.898)	**0.005**	4.04(1.167)	5.37(0.974)	**<0.001**
Reliability	4.97(0.938)	4.83(0.935)	**0.046**	4.93(0.919)	4.88(0.881)	0.354	4.55(1.061)	5.25(1.061)	**<0.001**
Authority	3.59(1.155)	4.19(0.935)	**<0.001**	3.82(0.911)	3.96(0.956)	**0.019**	3.44(1.003)	4.34(1.067)	**<0.001**
Involvement	4.79(1.099)	4.50(1.019)	**0.001**	4.74(1.023)	4.54(0.980)	**<0.001**	4.57(1.131)	4.72(0.992)	0.077

These results show that:

- Young agents are rated higher on friendliness, expertise, reliability and involvement, whereas old agents are seen as more authoritative.
- Female agents are seen as more friendly and involved, whereas males are perceived as more authoritative and having more expertise.
- Expert agents are seen as more reliable and authoritative and having more expertise than peer agents.
- Advice of young and expert agents is more likely to be followed.

5.5.3 Relation Characteristics Respondent and Features Preferred Agent

We tested the relation of the age, gender and health literacy of the respondents with the three agent features – age, gender and role – respectively.

5.5.3.1 Relation Age Preferred Agent and Age Respondent

Table 6 shows that there was a significant difference between the mean age of the respondents' that preferred an image of a young agent (M = 47.52 years, SD = 21.925) and the mean age of the respondents' that preferred an image of an old agent (M = 64.46 years, SD = 13.312). Older respondents are more likely to select an image of an old agent than younger respondents are.

Table 6: Results of the independent-samples t-test testing the relation between the age of the respondent and the age of the agent design preferred by the respondent (N = 123): the relation is statistically significant.

	Age Preferred Agent		
	Young, M(SD)	*Old, M(SD)*	*p*
Age Respondent	47.52(21.925)	64.46(13.312)	**<0.001**

5.5.3.2 Relation Gender Preferred Agent and Gender Respondent

Table 7 shows that there was a significant relation between the gender of the respondents (male or female) and the gender of the selected agent design. Female respondents are more likely to select an image of a female agent.

Table 7: Results of the Chi-square tests testing the relation between the gender of the respondent and gender of the agent design preferred by the respondent (N = 126): the relation is statistically significant.

	Gender Preferred Agent		
	Female, N(%)	*Male, N(%)*	*p*
Female	48(73.8%)	17(26.2%)	**0.008**
Male	31(50.8%)	30(49.2%)	

5.5.3.3 Relation Role Preferred Agent and Health Literacy Respondent

Table 8 shows that no significant effect of the role of health literacy of the respondent on the role of the preferred agent was found, neither for the general and elderly population.

Table 8: Results of the Fisher's exact tests testing the relations between the health literacy of the respondent and the role of the agent design preferred by the respondent (N = 126): the relation is not statistically significant.

	Role Preferred Agent		
	Peer, N(%)	*Expert, N(%)*	*p*
Low literate	3(42.9%)	4(57.1%)	1.000
Moderate or high literate	56(47.1%)	63(52.9%)	

5.5.4 Attitude Elderly Population towards Agent Characteristics

In the focus groups, elderly provided their opinion on agent features important for an agent in the health domain. Figure 13 shows the frequencies of the ratings of importance for agent characteristics of the card sorting task in the focus groups.

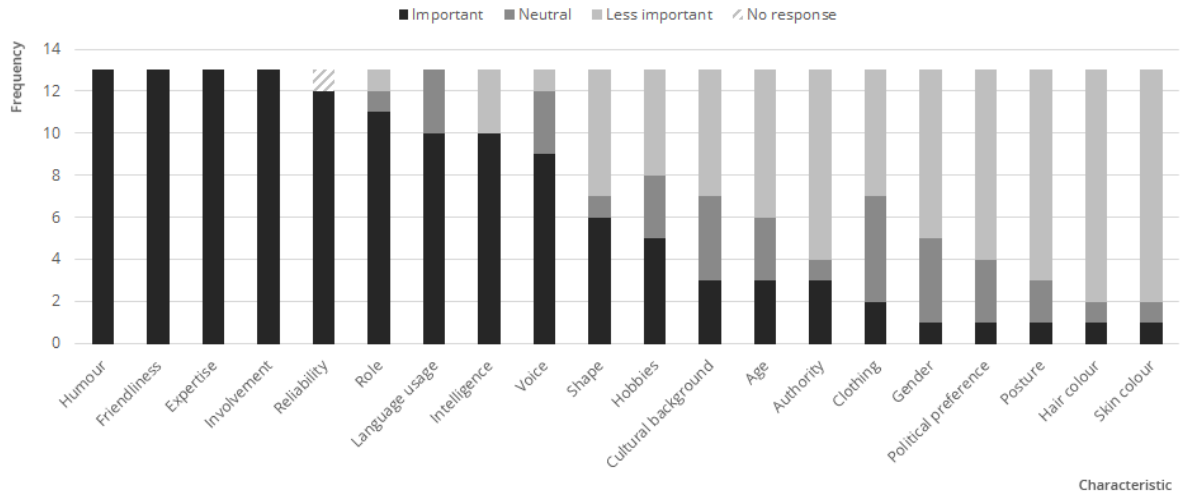

Figure 13: Results of card sorting task in focus groups. Participants rated each characteristic as either important, neutral or less important.

Design Features – Figure 13 shows that, overall, the design features age and gender subjected to test in the online questionnaire were not considered important. On the contrary, the agent's role was considered important by almost all participants.

Agent Characteristics – Figure 13 shows that almost all agent characteristics used in the online questionnaire – friendliness, expertise, reliability and involvement – were considered important. Authority was considered less important

5.6 Discussion

Although this research focused on first impressions of static agent images – images representing agents, but without interaction between the agent and the user –, the results of our research might be translated to embodied conversational agents, or just conversational agents, such as chat bots. What can be learned from this research is that:

- The agent design features – age, gender and role – affect the perception of the characteristics of a static agent image at first glance. Therefore, considering these design features when developing an (embodied) conversational agent can be beneficial.
- In addition, particular user characteristics, such as age and gender, affect the first impressions we have of the characteristics of a static agent image; personalization of agent designs matters. Adapting the agent design to the user could optimize the user's first impression and result in a positive start of the interaction with an (embodied) conversational agent.

6 Evaluation Study 2: On the effects of agent's gender, role and focus on user's persuasion

6.1 Objectives

The evaluation study is focused on understanding the effects of gender and role (authoritative and peer) on user's persuasion in single agent (dyadic) and multiple agent setting (vicarious: the process where the aim is to persuade the audience rather than the person with whom a proponent is directly engaged in discussion; and user-directed: where the aim is to persuade the person with whom a proponent is directly engaged in discussion) was conducted. Participants were presented with a persuasive message by one or several virtual agents, and a standard questionnaire was used to measure perceived interpersonal attitude, credibility and persuasion.

6.2 Methods

The participant began the study by filling in the demographics data i.e., age, gender and education level followed by accepting the consent form. The study is divided into three main steps, (1) Pre-questionnaire, (2) Answering questionnaire after watching a video clip with persuasive dialogue (collected 3 times i.e., one per given film genre), (3) Post-questionnaire. The pre-questionnaire is designed to measure the extent to which the participant is persuadable.

The users are first presented with a short textual description of three films of a given genre and asked to rate the likeliness of watching the films respectively. Once the ratings are provided, the user is assigned randomly to one of the 12 conditions and presented with a persuasive video clip about the film. Since we want to measure the persuasion in user, we opted to show the clip corresponding to the film that received lowest rating by the user. The clip generally is 60s - 90s long, consisting of virtual characters presenting opinion and information about the film. The participants were again asked to rate the likeliness of watching the film again followed by questionnaire to measure attitude, perceived credibility, and perceived persuasiveness. This step is repeated again for the other two film genres. The condition does not differ between the genre of films and remains the same throughout the experiment. Finally, a post-questionnaire is used to measure persuasiveness, trust in the agents, overall satisfaction and intention to continue using the system.

6.2.1 Stimuli

6.2.1.1 Discussion topic:

We selected films as our discussion topic. To avoid any biases that may occur if our participants had previously watched any of the films, we created our own film descriptions in three different genres: Comedy, Crime and History. The description of each of the 9 films followed a similar structure, while we kept the film titles and the language of the film descriptions as neutral as possible.

6.2.1.2 Agents:

We designed four characters that differed in gender and status. We manipulated the visual appearance, non-verbal behaviours and linguistic style to fit the roles of an authoritative and a peer agent. The design of the appearance of the agents were largely based on literature that studied the effects of gender and status of virtual agents in motivating and learning environments (Baylor & Kim , 2004). The authoritative agent was designed to fit the role of a film critique and to be perceived as an expert in films. Research shows that expertise in humans requires several years of deliberate practice in a domain (Ericsson, Krampe, & Tesch-Römer, 1993). Hence, we modelled the agent to appear aged in late-forties and dressed formally in a professional manner. The peer agent was designed to fit the role of a film-enthusiast who enjoys watching films and appeared as a student in the early-twenties and dressed casually. The appearances of the agents were designed using the Autodesk Character Generator software (see for example Figure 14 below).

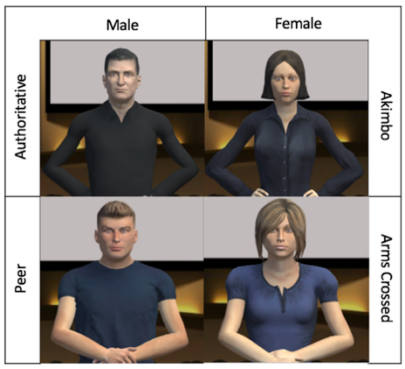

Figure 14: Agent appearance.

We define the peer agent as warm, friendly and the authoritative agent as competent and dominant. The chosen behaviours for the agents were based on literature on persuasion in humans (Kendon, Gesture: Visible action as utterance, 2004) (Poggi & Vincze, 2009). Table 9 provides an overview of the variables we manipulated. We chose to characterize agents' behaviours depending on their status only, either authoritative or peer. We did not differentiate in behaviours computation for the age or gender variables. We made use of the virtual agent platform Greta (Pecune, Cafaro, Chollet, Philippe, & Pelachaud, 2014) with Unity3D for generating the animations of the virtual agent and Cereproc TTS voice synthesizer to generate the audio for the virtual agents.

Table 9: Overview of the distinctive characteristics for authoritative and peer agents.

Parameters	Authoritative	Peer
Appearance	Aged Formal Clothing	Young Casual Clothing
Facial Expressions	Frown Corner lip down	Eyebrow raise Open smile
Frequency of Smile	Low	High
Rest Position	Akimbo	Arms Crossed
Gestures	Ring Deictic finger pointing	Beat open hand

6.2.1.3 Dialogues:

Our dialogues utilised Aristotle's three modes of persuasion i.e. ethos: appeal to authority, pathos: appeal to emotion, and logos: appeal to logic. Each dialogue began with an introduction, before six arguments were made in front of the user. Two arguments were based on ethos, two on logos and two on pathos. Each dialogue then ended with a wrap-up section to conclude it. In the multiple agent scenarios using direct persuasion, the agents were both directing their arguments at the user with each using one ethos, pathos and logos argument. To distinguish between peer and authority approaches we incorporated ethos into the authority agent, used more formal language and vocabulary as well as utilised more credible sources. For the multi-agent dialogues using vicarious persuasion, one agent tried to persuade the other. In each case the persuading agent was the only one making an argument, while the other agent was listening and appearing as if they were being persuaded.

6.2.2 Experiment

6.2.2.1 Design

The experiment is based on 2 x 2 x 3 design. The variables include agent gender (male vs. female), status (authoritative vs. peer) and persuasion type (multiple agent user-directed vs. multiple agent vicarious vs. single agent). Since we are also studying the effects of gender, in multiple agent condition, a male agent and a female agent were present and only the status is altered. In vicarious persuasion, the status of both speaker and addressee agent will always be the same and only the gender is altered. Table 10 provides the overview of the twelve conditions used in the study.

Table 10: Overview of the twelve randomly allocated experimental conditions; F: Female, M: Male, A: Authoritative, P: Peer.

Focus	Composition	Condition	Participants
Single Agent	F A	C1	18
	M A	C2	18
	F P	C3	18
	M P	C4	15
Multiple Agent (vicarious)	F A persuades M A	C5	18
	M A persuades F A	C6	17
	F P persuades M P	C7	18
	M P persuades F P	C8	16
Multiple Agent (user-directed)	F A + M A	C9	18
	F A + M P	C10	18
	F P + M A	C11	15
	F P + M P	C12	16

6.2.2.2 Procedure

The participant began the study by filling in the demographics data i.e. age, gender and education level followed by accepting the consent form. The study is divided into three main steps, (1) Pre-questionnaire, (2) Answering questionnaire after watching a video clip with persuasive dialogue (collected 3 times i.e. one per given film genre), (3) Post-questionnaire. The pre-questionnaire is designed to measure the extent to which the participant is persuadable. Along with this the participant also provides information about overall openness and comfort towards technology and interest in films. A short introductory clip was designed using a virtual agent who presented the study. The age of the agent was in its 30s and its appearance was smart casual. This was done in order to familiarize the participants with the animations of the virtual agents to avoid collecting responses based on the first impression generated.

The users are first presented with a short textual description of three films of a given genre and asked to rate the likeliness of watching the films respectively. Once the ratings are provided, the user is assigned randomly to one of the 12 conditions specified above and presented with a persuasive video clip about the film. Since we want to measure the persuasion in user, we opted to show the clip corresponding to the film that received lowest rating by the user. The clip generally is 60s - 90s long, consisting of virtual characters presenting opinion and information about the film. The participants were again asked to rate the likeliness of watching the film again followed by questionnaire to measure attitude, perceived credibility, and perceived persuasiveness. This step is repeated again for the other two film genres. The condition does not differ between the genre of films and remains the same throughout the experiment.

Finally, a post-questionnaire is used to measure persuasiveness, trust in the agents, overall satisfaction and intention to continue using the system.

6.2.3 Participants

For this study we collected responses in two stages. Initially 282 participants were recruited from Crowdflower. A total of 156 responses were removed from the collected data due to inconsistencies and non-naivety as several participants did not adhere to the instructions and responded multiple times and we considered the responses to be not genuine. We also collected 79 responses by contacting respondents through mailing lists.

6.3 Results

In total, we had 209 participants where 55% were male (n = 113) and 45% were female (n = 92). 46% of the participants were between the age range of 21-30 years, 22% between 31-40, and 15% between 41-50 and 14% above 50 years old. The participants came from different cultural backgrounds with the three most prominent groups from, North America (37%), Europe (27%), and Asia (20%).

The likeliness score of watching a film before and after the persuasive clip, is an indication that agents were successful in persuading the users to reconsider their decision about wanting to watch a film. 153 participants reconsidered their rating at least once and increased it. Participants who were grouped under `easily persuadable' (n = 150) reported significantly higher persuasion from both authoritative and peer agents and the change in scores indicate the same. Agent's role did not have any effect on participants grouped under `not easily persuadable' (n = 55), however, the vicarious setting was more effective in persuading them.

Authoritative agents were reported to be more credible regardless of the gender of the agent and participants reported higher level of trust in the information provided by them, indicating that the authoritative agents were perceived as competent (Fiske, Cuddy, & Glick, 2007). This is in line with (Baylor & Kim , 2004), where expert-like agents were perceived to be more credible. Additionally, they were also reported to be more persuasive than peer agents. In (Holzwarth, Janiszewski, & Neumann, 2006), the expertise of the agent influenced the perceptions of credibility, and credibility mediated the influence of the agent's expertise on persuasion.

While status played an important role, agent's gender did not have any significant effect. In previous studies (Baylor & Kim , 2004) (Guadagno, Blascovich, Bailenson, & Mccall, 2007) gender had a role in persuasion. However, in our study, gender was simply differentiated by the appearance of the agent and there was no other difference at the behaviour level or at the interaction level which can explain why there was no significant effect of gender.

Table 11: Mean value of change in likeliness score of watching a film (before & after the persuasive clip) and the self-reported persuasiveness for the three conditions.

Condition	Change in score	Self-report
Single Agent	0.285	2.882
Multiple Agent (Vicarious)	**0.604**	2.969
Multiple Agent (User Directed)	0.413	**3.028**

The main finding of this evaluation study is that, a multiple agent setting was more effective than a single agent. The persuasiveness questionnaire revealed that participants reported being more influenced by the user-directed multiple agent setting. However, we measured the mean change in rating, for each condition and this revealed that vicarious setting was more effective in persuading the user to change their score than user-directed setting cf. Table 11. In particular, authoritative agents were more effective in vicarious setting than single agent setting. Since the difference between the three settings was not statistically significant, we suggest that there is a strong tendency in the result that needs to be further verified with more participants.

Additionally, the agents in the multiple setting (user-directed) were considered to be more credible than a single agent and also users reported that they would consult the agents again and would recommend it to friends. This setting was also more helpful and users reported high satisfaction. 39% of the users in single agent setting preferred to have multiple agents with different perspective while only 16% preferred to have one agent condition. From the above results it is quite evident that multiple agent condition is indeed more effective, in particular, when vicarious persuasion is used.

Our results on effects of settings are in line with human studies from social cognitive science. In (Meier, 2012) it is argued that verbal persuasion by a single person is less efficient than vicarious experience on self-efficacy and behavioural change. Studies on interactive narrative systems report also that users are more influenced and engaged when experiencing vicarious social relationships and emotional responses than when experiencing events from their own direct environment (Slater, 2002). Moreover, these studies underline how the effect of persuasion depends on the level of identification of the users with the interaction content.

In our study, participants who reported being `easily persuadable' did report high persuasion. Further, the association between perceived credibility and perceived persuasion was observed using Spearman's rank correlation coefficient cf. Figure 13. We observe that perceived credibility positively affects the perceived persuasion (rs = 0:92, p < 0:01). This tendency has been studied in detail in (Burgoon, Birk, & Pfau, 1990) (Lehto, Oinas-Kukkonen, & Drozd, 2012) (Pornpitakpan, 2004), where credibility is linked with persuasion. We can conclude that a credible agent can be effective for promoting behaviour change in a multiple agent setting.

Figure 15: Mean rating of perceived credibility (p = 0:0003) C1-C4: Single agent (m = 2:894); C5-C8: Vicarious (m = 2:936); C9-C12: Multiple agent (m = 2:966); and Perceived Persuasion (p = 0:00002) over 12 conditions. C1-C4: Single agent (m = 2:882); C5-C8: Vicarious (m = 2:969); C9-C12: Multiple agent (m = 3:028); over 12 conditions.

6.4 Discussion

In this section we presented the results of two evaluation studies that were conducted in order to understand the user's perception of the agents and their perceived level of persuasion. From the results of the first study we can observe that at first glance users tend to prefer young and expert agents and are most likely to follow their advice. It also shows that the preferred agents have a lower rating on authority. In the second study reported we see that authoritative agents in general were more persuasive. Since the first study involved first impression of the participants and the second study focused on presenting persuasive arguments, the results cannot be comparable. However, this brings up interesting opportunities to understand the differences in the two studies for example, the context in which the virtual agents would provide advice and also the type of interactions. Therefore, we plan to conduct evaluation studies throughout the project to better understand the perception of agents by the user.

7 Conclusion

In this document, we report our progress on the development of the virtual agent's model. The model consists of group cohesion, turn-taking behaviour, gesticulation, gaze behaviour, and back-channel performance. We also report the result of our human experimentation on the impact of virtual agent's profile on user's perception. Lastly, we have also enhanced GRETA platform to allow users to have multiple agents in the same instance.

We have done literature review on group cohesion and turn taking behaviour. We have found prior work, from both computer science and social science. Our next steps will be studying the relationship between certain non-verbal behaviour cues with turn-taking behaviour and group cohesion.

On gesticulation, we have done literature review and we are trying to learn gesture stroke timing by using attention model. Our next steps will be improving the attention model performance, studying the impact of certain prosodic features on the gesture stroke timing, and predicting the shape of the gesture.

On gaze behaviour, we have done the implementation which enables the agent to continuously gaze at the user's eyes. This feature has also been integrated with ASAP.

On back-channel performance, we have done the implementation which enables the agent to detect the user's behaviour and perform back-channel appropriately. The production of back-channel is governed by a set of rules stored in GRETA.

On the human experimentation, we tested user's perception on different virtual agents. The differences between the agents are on gender, role, and age. We found that those factors matter on the user's perception. However, the user's perception is also affected by their age, gender, and education.

8 Bibliography

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation, 41*(3-4), 273-287.

Back, K. (1951). Influence through social communication. *The Journal of Abnormal and Social Psychology, 46*(1).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ICLR 2015*.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4945-4949). IEEE.

Barange, M., Saunier, J., & Pauchet, A. (2017). Multiparty Interactions for Coordination in a Mixed Human-Agent Teamwork. *International Conference on Intelligent Virtual Agents* (pp. 29-42). Springer.

Baylor, A., & Kim , Y. (2004). . Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. *International Conference on Intelligent Tutoring Systems.*, 592–603.

Beal, D., Cohen, R., Burke, M., & McLendon, C. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology, 88*(6).

Bergmann, K., & Kopp, S. (2009). GNetIc-Using bayesian decision networkds for iconic gesture generation. *International Workshop on Intelligent Virtual Agents* (pp. 76-89). Springer.

Bevacqua, E., Heylen, D., Pelachaud, C., & Tellier, M. (2007). Facial feedback signals for ECAs. *AISB, 7*, 328-334.

Bohus, D., & Horvitz, E. (2010). *Computational Models for Multiparty Turn Taking.* Microsoft Research Technical Report MSR-TR 2010-115.

Braaten, L. (1991). Group cohesion: A new multidimensional model. *Group, 15*(1), 39-55.

Burgoon, J., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human communication research, 17*(1), 140-169.

Carless, S., & De Paola, C. (2000). The measurement of cohesion in work teams. *Small group research, 31*(1), 71-88.

Carron, A. (1982). Cohesiveness in Sport Groups: Interpretations and Considerations. *Journal of Sport psychology, 31*(1), 123-138.

Carron, A., & Brawley, L. (2000). Cohesion: Conceptual and measurement issues. *Small group research, 31*(1), 89-106.

Carron, A., & Chelladurai, P. (1981). The dynamics of group cohesion in sport. *Journal of Sport Psychology, 3*(2), 123-139.

Carron, A., Widmeyer, W., & Brawley, L. (1985). The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. *Journal of sport psychology, 7*(3), 244-266.

Casey-Campbell, M., & Martens, M. (2009). Sticking it all together: A critical assessment of the group cohesion-performance literature. *International Journal of Management Reviews, 11*(2), 223-246.

Cassell, J., Bickmore, T., Campbell, L., Hannes, V., & Yan, H. (2000). Conversation as a system framework: Designing embodied conversational agents. *Embodied Conversational Agents*.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., & Rich, C. (2001). Non-verbal cues for discourse structure. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 114-123). Association for Computational Linguistics.

Cassell, J., Torres, O., & Prevost, S. (1999). Turn taking vs discourse structure: How best to model multimodal conversation machine conversations. 143-154.

Cassell, J., Vilhjálmsson, H., & Bickmore, T. (2004). BEAT: The behavior expression animation toolkit. *Life-Like Characters*, 163-185.

Chiu, C., & Marsella, S. (2014). Gesture generation with low-dimensional embeddings. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 781-788). International Foundation for Autonomous Agents and Multiagent Systems.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems*, 577-585.

DeVault, D., Mell, J., & Gratch, J. (2015). Toward natural turn-taking in a virtual human negotiation agent. *2015 AAAI Spring Symposium Series*.

Driskell, J., & Radtke, P. (2003). The effect of gesture on speech production and comprehension. *Human Factors, 45*(3), 445-454.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology, 23*(2).

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in society, 3*(2), 161-180.

Ericsson, K., Krampe, R., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review, 100*(3).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1459-1462). ACM.

Ferstl, Y., & McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 93-98). ACM.

Festinger, L. (1950). Informal social communication. *Psychological review, 57*(5).

Festinger, L., Schachter, S., & Back, K. (1950). *Social pressures in informal groups; a study of human factors in housing.* American Psychological Association.

Fiske, S., Cuddy, A., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences, 11*(2), 77-83.

Forsyth, D. (2014). *Group dynamics.* Wasworth Cengage Learning.

Goodman, P., Ravlin, E., & Schminke, M. (1987). Understanding groups in organizations. *Research in organizational behaviour*.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language, 25*(3), 601-634.

Guadagno, R., Blascovich, J., Bailenson, J., & Mccall, C. (2007). Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology 10, 1*, 1–22.

Guaïtella, I., Santi, S., Lagrue, B., & Cavé, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. *Language and speech, 52*(2-3), 207-222.

Hadar, U., Steiner, T., Grant, E., & Rose, F. (1984). The timing of shifts of head postures during conversation. *Human Movement Science, 3*(3), 237-245.

Hartholt, A., Gratch, J., & Weiss, L. (2009). At the virtual frontier: Introducing Gunslinger, a multi-character, mixed-reality, story-driven experience. *Proceedings of the International Workshop on Intelligent Virtual Agents* (pp. 500-501). Springer.

Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., & Sumi, K. (2018). Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 79-86). ACM.

Holzwarth, M., Janiszewski, C., & Neumann, M. (2006). The influence of avatars on online consumer shopping behavior. *Journal of Marketing, 70*(4), 19-36.

Hung, H., & Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia, 12*(6), 563-575.

Ishii, R., Katayama, T., Higashinaka, R., & Tomita, J. (2018). Generating Body Motions using Spoken Language in Dialogue. *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 87-92). ACM.

Iverson, J., & Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature, 396*(6708).

Jan, D., & Traum, D. (2005). Dialog simulation for background characters. *International Workshop on Intelligent Virtual Agents* (pp. 65-74). Springer.

Jan, D., & Traum, D. (2007). Dynamic movement and positioning of embodied agents in multiparty conversations. *Proceedings of the Workshop on Embodied Language Processing* (pp. 59-66). Association for Computational Linguistics.

Kendon, A. (1990). Conducting interaction: Patterns of behavior in focused encounters. *7*.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge University Press.

Leßmann, N., Kranstedt, A., & Wachsmuth, I. (2004). Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max. *Proceedings of the Workshop on Embodied Conversational Agents: Balanced Perception and Action.*

Lee, J., & Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. *Proceedings of the International Workshop on Intelligent Virtual Agents* (pp. 243-255). Springer.

Lehto, T., Oinas-Kukkonen, H., & Drozd, F. (2012). Factors affecting perceived persuasiveness of a behavior change support system.

Levine, S., Krähenbühl, P., Thrun, S., & Koltun, V. (2010). Gesture controllers. *ACM Transactions on Graphics (TOG), 29*(4).

Lhommet, M., & Marsella, S. (2013). Gesture with meaning. *Proceedings of the International Workshop on Intelligent Virtual Agents* (pp. 303-312). Springer.

Lhommet, M., & Marsella, S. (2014). Metaphoric gestures: towards grounded mental spaces. *Proceedings of the International Conference on Intelligent Virtual Agents* (pp. 264-274). Springer.

Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71-89.

Mancini, M., & Pelachaud, C. (2009). Generating distinctive behavior for embodied conversational agents. *Journal on Multimodal User Interfaces, 3*(4), 249-261.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago Press.

Meier, S. (2012). *Language and narratives in counseling and psychotherapy.* Springer Publishing Company.

Mudrack, P. (1989). Defining group cohesiveness: A legacy of confusion? *Small group behavior, 20*(1), 37-49.

Padilha, E., & Carletta, J. (2002). A simulation of small group discussion. *Proceedings of EDILOG*, (pp. 117-124).

Pecune, F., Cafaro, A., Chollet, M., Philippe, P., & Pelachaud, C. (2014). Suggestions for extending SAIBA with the VIB Platform. *Proceedings of the Workshop on Architectures and Standards for Intelligent Virtual Agents at IVA*, (pp. 16-20).

Piper, W., Marrache, M., Lacroix, R., Richardsen, A., & Jones, B. (1983). Cohesion as a basic bond in groups. *Human Relations, 36*(2), 93-108.

Poggi, I., & Vincze, L. (2009). Gesture, gaze and persuasive strategies in political discourse. *Multimedia corpora*, 73-92.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied social Psychology, 34*(2), 243-281.

Ravenet, B., Cafaro, A., Biancardi, B., Ochs, M., & Pelachaud, C. (2015). Conversational behavior reflecting interpersonal attitudes in small group interactions. *Proceedings of the International Conference on Intelligent Virtual Agents* (pp. 375-388). Springer.

Ravenet, B., Clavel, C., & Pelachaud, C. (2018). Automatic Nonverbal Behavior Generation from Image Schemas. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 1667-1674). International Foundation for Autonomous Agents and Multiagent Systems.

Reithinger, N., Gebhard, P., Löckelt, M., Ndiaye, A., Pfleger, N., & Klesen, M. (2006). VirtualHuman: dialogic and affective interaction with virtual characters. *Proceedings of the 8th international conference on multimodal interfaces* (pp. 51-58). ACM.

Rist, T., André, E., Baldes, S., Gebhard, P., Klesen, M., Klpp, M., & Schmitt, M. (2004). A review of the development of embodied presentation agents and their application fields. *Life-Like Characters*, 377-404.

Sacks, H., Schegloff, E., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. *Studies in the organization of conversational interaction*, 7-55.

Saint-Amand, K. (2018). Gest-IS: Multi-lingual Corpus of Gesture and Information Structure. *Unpublished Report*.

Santoro, J., Dixon, A., Chang, C., & Kozlowski, S. (2015). Measuring and monitoring the dynamics of team cohesion: methods, emerging tools, and advanced technologies. *Team cohesion: Advances in psychological theory, methods and practice*, 115-145.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in society, 29*(1), 1-63.

Slater, M. (2002). *Entertainment education and the persuasive impact of narratives.* Lawrence Erlbaum Associates Publishers.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., & Lane, C. (2010). Ada and Grace: Towards realistic and engaging virtual museum guides. *Proceedings of the International conference on Intelligent Virtual Agents* (pp. 286-300). Springer.

ter Stal, S., Tabak, M., op den Akker, H., Beinema, B., & Hermens, H. (2019 (Submitted)). Who Do You Prefer? The Effect of Age, Gender and Role on Users' First Impressions of Embodied. *International Journal of Human Computer Interaction*.

Thórisson, K., Gislason, O., Jonsdottir, G., & Thórisson, H. (2010). A multiparty multimodal architecture for realtime turntaking. *Proceedings of the International conference on intelligent virtual agents* (pp. 350-356). Springer.

Traum, D., & Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: Part 2* (pp. 766-773). ACM.

Acknowledgements

The Council of Coaches project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement #769553. This result only reflects the author's view and the EU is not responsible for any use that may be made of the information it contains.

Headings and titles in this document, as well as the Council of Coaches logo use the Comfortaa font, designed by Johan Aakerlund and Cyreal and licensed under the Open Font License[1].

Additional text in this document uses the Roboto font, designed by Christian Robertson and licensed under the Apache License, Version 2.0[2].

The Council of Coaches logo and Blobmen graphics were *drawn freely* in Inkscape, licensed under the GNU General Public License[3].

[1] Open Font License: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=OFL_web
[2] Apache License, Version 2.0: http://www.apache.org/licenses/LICENSE-2.0
[3] Inkscape License Information: https://inkscape.org/about/license/