Chapter 4

Advanced Content and Interface Personalization through Conversational Behavior and Affective Embodied Conversational Agents

Matej Rojc, Zdravko Kačič and Izidor Mlakar

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75599

Abstract

Conversation is becoming one of the key interaction modes in HMI. As a result, the conversational agents (CAs) have become an important tool in various everyday scenarios. From Apple and Microsoft to Amazon, Google, and Facebook, all have adapted their own variations of CAs. The CAs range from chatbots and 2D, carton-like implementations of talking heads to fully articulated embodied conversational agents performing interaction in various concepts. Recent studies in the field of face-to-face conversation show that the most natural way to implement interaction is through synchronized verbal and co-verbal signals (gestures and expressions). Namely, co-verbal behavior represents a major source of discourse cohesion. It regulates communicative relationships and may support or even replace verbal counterparts. It effectively retains semantics of the information and gives a certain degree of clarity in the discourse. In this chapter, we will represent a model of generation and realization of more natural machine-generated output.

Keywords: co-verbal behavior generation, affective embodied conversational avatars, humanoid robot behavior, multimodal interaction, unity, EVA framework

1. Introduction

One of the key challenges in the modern human-machine interaction (HMI) is the generation of more natural, more personalized, and more human-like human-machine interaction [1]. As a result, the conversational agents (CAs) are gaining interest and traction, especially due to the fact that most of the user devices are already capable to support multimedia and the concept of conversational agents (CAs). Apple, Microsoft, Amazon, Google, Facebook,

etc., already have adapted their own variations of CAs. Moreover, the newest technologies, such as the Amazon Echo and Google Home, which are positioned as supporting multiuser and highly personalized interaction in collocated environments (e.g., homes and ambient-assisted living environments), integrate virtual agents supporting both visual and auditory interactions [2–4]. Thus, exploring the conversational models and challenges around CA supported interaction represents a timely topic. The production of conversational behavior, e.g., socially shared information and attitude, incorporates much more than just speech verbal exchange. Namely, it is multimodal and multilayered, since it entails multiple verbal and nonverbal signals that are correlated in a dynamic and highly unpredictable settings. One might even say that the social interaction involves synchronized signal verbal and nonverbal channels. The verbal channels carry symbolic/semantic interpretation of message through the linguistic and paralinguistic features of interaction, while the co-verbal channels serves as an orchestrator of communication [5–7]. Thus, such an interaction facilitates full embodiment of the collocutors. It also exploits physical environment in which the interaction is positioned in [8–10]. Further, the co-verbal behavior is actually equally relevant as speech. Namely, it actively contributes to the information presentation and understanding, as well as the discourse itself. It establishes semantic coherence and regulates communicative relationships. It may support or even replace the verbal communication in order to clarify or reinforce the information provided by the verbal counterparts [11–13]. The co-verbal behavior goes well beyond an add-on or a style of information representation. For instance, spatial orientation of the face and eye gaze are key nonverbal cues that shape the footing of the conversational participants [14]. Through the co-verbal responses, the listeners may signal their interest, attention, and understanding [15]. As a result, the role of co-verbal (and nonverbal) behavior in human communication and in human-machine interaction has been increasingly scrutinized over the last few decades, within a wide range of contexts [16–21]. Embodied conversational agents (ECAs) are nowadays the most natural selection for the generation of affective and personalized agents. ECAs are those CAs that can facilitate full virtual body and the available embodiment in order to incorporate humanlike responses. The ECA technology ranges from chatbots and 2D/3D realizations in a form of talking heads [22–24] to fully articulated embodied conversational agents engaged in various concepts of HMI, including sign language [25], storytelling [26], companions [27], and virtual hosts within user interfaces, and even used as moderators of various concepts in ambient-assisted living environments [28–32].

In this chapter we present novel expressive conversational model for facilitating human-like conversations and a solution for affective and personalized human-machine interaction. The model facilitates (i) a platform for the generation of "conversational" knowledge and resources, (ii) a framework for planning and generation of (non-)co-verbal behavior, and (iii) a framework for delivery of affective and reactive co-verbal behavior through attitude, emotion, and gestures synchronized with the speech. Namely, the EVA expressive conversational model is outlined in Section 2. The main idea is to formulate various forms of co-verbal behavior (gestures) with respect to arbitrary and unannotated text and broader social and conversational context. The "conversational" knowledge and resources required are generated via annotation of spontaneous dialog and through the corpus analysis as presented in Section 3.

http://dx.doi.org/10.5772/intechopen.75599

In Chapters 4 and 5, how these resources are integrated into the two-folded approach of the automatic co-verbal behavior generation is then described. The presented approach involves (a) the problem of behavior formulation (intent and behavior planning) and (b) the problem of behavior realization (animation via ECA). Finally, we conclude with synthesis of affective co-verbal behavior within interfaces and final remarks.

2. EVA conversational model for expressive human-machine interaction

In order to cope also with the complexity in multiparty conversations, and in order to apply the knowledge to various concepts in human-machine interaction in a form of conversational behavior, we have envisaged and deployed an advanced EVA conversational model, which is used (a) to study the nature of natural behavior of human-collocutors; (b) to create conversational knowledge in form of linguistic, paralinguistic verbal, and nonverbal features; (c) and to test theories and to apply knowledge in various conversational settings as part of situation understanding or as a part of output generation processes. The presented EVA conversational model is outlined in **Figure 1**. As can be seen, it consists of the following three cooperative frameworks/platforms: *conversational analysis platform*, *EVA framework*, and *EVA realization framework*.

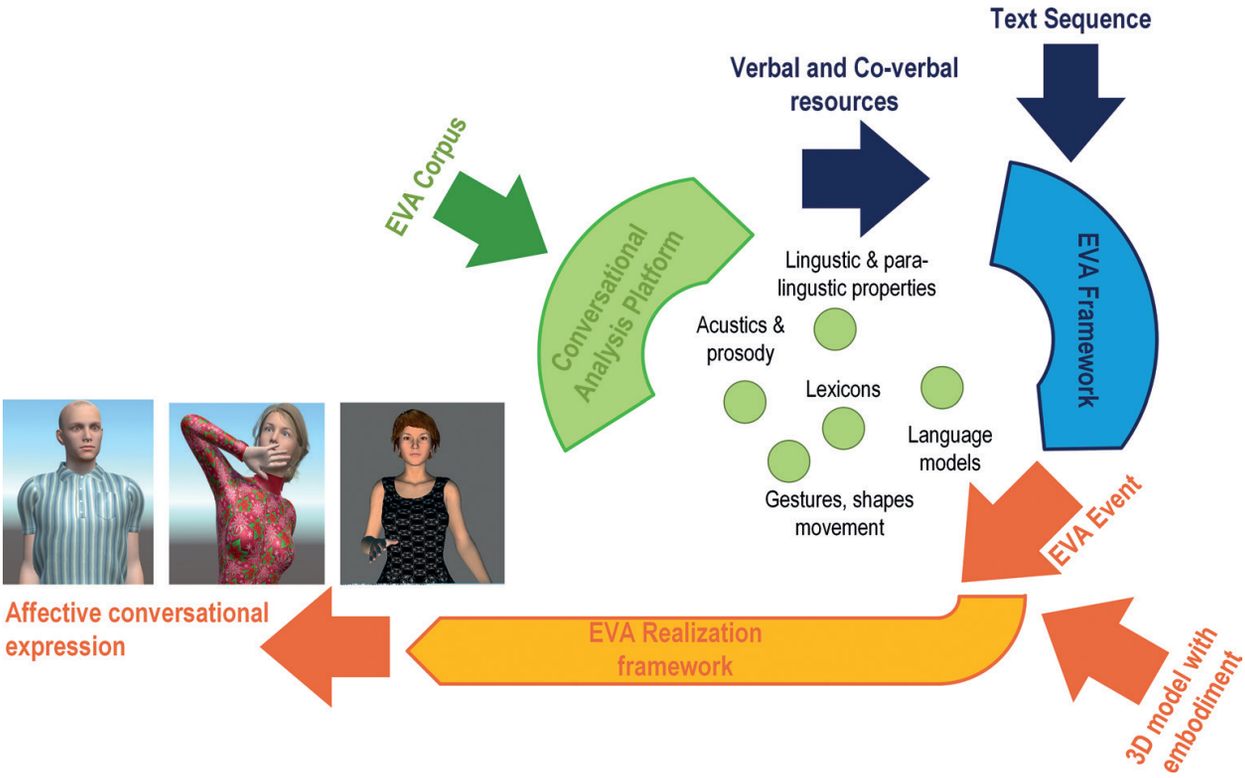

Figure 1. EVA conversational model for generation and exploitation of expressive human-like machine responses.

The model builds on the notion that verbal to co-verbal alignment and synchronization are driving forces behind affective and social interaction. Thus, the *conversational analysis platform* is used for analyzing how linguistic and paralinguistic features interplay with embodiments during complex, spontaneous, and multiparty interactions. Its processing is based on multimodal corpus, named EVA corpus [33], the EVA annotation scheme developed to describe the complex relations of co-verbal behavior (proposed in [34]), and *EVA framework* capable (a) to capture various contexts in the "data" and (b) to provide the basis to analytically investigate into various multidimensional correlations among co-verbal behavior features.

Communication is in its fundaments a multimodal process. In order to describe the nature of face-to-face interaction, we have chosen to build upon the concept of "multimodality in interaction" over linguistic basis [2]. We extended this concept with a cognitive dimension consisted of various concepts, such as emotions, sentiment, and communicative intent. We also blend it with the meta-information represented as a set of paralinguistic features, such as facial expressions and gestures, prosody, pitch, dialog function and role, etc., [17, 35]. As outlined in **Figure 1**, the annotated EVA corpus is used to build resources required for the planning and generation of conversational behavior, both verbal part (e.g., text-to-speech (TTS) synthesis) and (non-)co-verbal part (conversational behavior synthesis). These are lexicons, language models, semiotic grammar of communicative intent, lexicon of conversational shapes, gestures and movements, acoustic and prosodic properties, and other linguistic and paralinguistic features (e.g., word/syllable segmentation, sentence type, sentiment, etc.) that are used within behavior generation rules and machine-learned behavior generation models. The resources generated within the *conversation analysis platform* are fed to the *EVA framework*. The main idea of the proprietary *EVA framework* proposed in [36] is to evoke a social response in human-machine interaction through affective synthetic response generated on arbitrary and unannotated texts. Thus, the EVA behavior generation model within the *EVA framework* is data-driven and also driven by the text-to-speech (TTS) synthesis engine. The model is modular and merged with the TTS engine's architecture, into the first omni-comprehensive TTS system's engine as proposed in [37]. The output of the EVA framework represents the complete co-verbal behavior described by using the proprietary procedural description language, named EVA-Script, and the synthesized speech, both perfectly synchronized at the phoneme level. The co-verbal behavior is described within an EVA event, where it represents a contextual link between the language, context-independent motor skills (shapes, movements, and poses that conversational agent can display), and the context-dependent intent for which the behavior has been generated (e.g., situation, relation, communicative function, etc.). The behavior is already adapted to the nature and capabilities of the virtual entity representing it. The EVA event then represents input for the *EVA realization framework*.

The *EVA realization framework* is built on the premises that natural multimodal interaction is much more than speech accompanied with the repetitive movements of the limbs and face. Namely, natural interaction entails multiple behavior variations that are correlated in dynamic and highly unpredictable settings [6]. It also incorporates various social and interpersonal signals in order to "color" the final outcome and can dynamically adapt to various intra- and interpersonal contexts as well as various situational contexts [4, 38]. The role of this framework is to transform the co-verbal descriptions contained in EVA events into

articulated movement generated by the expressive virtual entity, e.g., to apply the EVA-Script language onto the articulated 3D model EVA in the form of animated movement. Further, the framework contains the *animation-parameters builder* and the *animation-realization engine*. Both are maintained within the *EVA behavior realizer* and implemented as a series of standalone modules. The *animation-parameters builder* is used to understand the specified behavior and to adapt it to the restriction of the targeted agent. Thus, it transforms the EVA-Script sequence into animation parameters that are then mapped to different control units of the ECA's 3D articulated model, while the *animation-realization engine* is responsible for scheduling and execution of the translated animation parameters in a form of sequences of parallel/sequential transformation of ECA-related resources (e.g., meshes textures, bones, morphed shapes, etc.). These resources actually constitute virtual agent's embodiment in a form of hand/arm gestures, posture and body configuration, head movement and gaze, facial expressions, and lip sync. Finally, the *EVA realization framework* also incorporates procedures required for efficient integration of the embodied conversational agent into various user interfaces. In the next section, we will present particular modules of the *EVA conversational model* in more detail. Firstly, we will describe the *conversational analysis platform*, which represents the basis not only for the generation of communicative behavior but also for understanding the nature of complex conversational behavior and face-to-face interaction as a whole.

3. Conversation analysis and annotation scheme

Conversation analysis represents a powerful tool for analyzing language and human co-verbal behavior in various aspects of social communication [39]. Namely, interaction through dialog is an act of conveying information, in which humans can convey information through a variety of methods, such as speaking, body language (gestures and posture) and facial expression, and even social signals [40]. Interpersonal communication can involve the transfer of information between two or more collocutors that use verbal and nonverbal methods and channels. Symbolic/semantic interpretation of message is presented through linguistic and paralinguistic features, while the co-verbal part in general serves as an orchestrator of communication [5]. The concept of the co-verbal behavior has become one of the central research paradigms and one of the important features of human-human interaction. It has been investigated from various perspectives, e.g., from anthropology and linguistics and psychosociological fields to companions, communication and multimodal interfaces, smart homes, ambient assisted living, etc. The multimodal corpora represent the results of various research efforts. They are the tools through which researchers analyze the inner workings of interaction. The knowledge generated by using such multimodal corpora and annotation schemes, therefore, represents a key resource for better understanding the complexity of the relations between verbal and co-verbal parts of human-human communication. It provides insights into understanding of several (social) signals and their interplay in the natural exchange of information. In the following section, we will represent EVA corpus, a corpus of spontaneous and multiparty face-to-face dialog. We will also outline the EVA annotation scheme designed as a tool for corpus analytics and generation of verbal and nonverbal resources [33].

3.1. The multimodal EVA corpus

Among video corpora, television (TV) interviews and theatrical plays have shown themselves to be very usable resource of spontaneous conversational behavior for the analytical observation, and annotation of co-verbal behavior and emotions, used during conversation. In general, TV discussions represent a mixture of institutional discourse, semi-institutional discourse and casual conversation. Material used in existing corpora is often subject to certain restrictions in order to reduce the conversational noise, such as time restriction, strict agenda, strict scenario and instructions to implement targeted concepts, and technical features (camera direction and focus, editing) that further influence especially communicative function of co-verbal behavior and its expressive dimensions (speech, gestures, facial displays). However, the conversational noise, if properly analyzed and incorporated, may unravel a lot of features and contexts that model the natural multimodal conversational expressions. Namely, by exploiting the casual nature and noise in the material, as we do with the EVA corpus, we can take into consideration the complete interplay of various conversation phenomena, such as dialog, emotional attitude, prosody, communicative intents, structuring of information, and the form of its representation. All these can give us a true insight into how informal communication works, what stimuli triggers conversational phenomena, and how do these impulses interact and reflect on each other. Such relations can then provide synthetic agents with the basis for the multimodal literacy, namely, the capacity to construct meaning through understanding of situation and responding to some not predefined situation. The conversational setting in the EVA corpus is totally relaxed and free and is built around a talk show that follows some script/scenario; however, the topics discussed are highly unpredictable, changeable, informal, and full of humor and emotions. Further, although sequencing exists, it is performed highly unorderly as are also the communicative functions. This guarantees a highly causal and unordered human discourse, with lots of overlapping statements and roles, vivid emotional responses, and facial expressions. The goals of the corpus and the annotation scheme are built around (semiotic) communicative intent as the driving force for the generation of co-verbal and communicative behavior. The communicative intent is a concept through which we are able to correlate the intent of the spoken information (defined, e.g., through part-of-speech (POS) tags, prosodic features, and classification of interpretation through meaning) with co-verbal behavior (gestures). Human face-to-face interactions are multimodal and go well beyond pure language and semantics. Within the EVA corpus and corpus analysis outlined in this section, we decided for the extension of semantics by applying the concept of communicative intent and other linguistic and paralinguistic features, such as dialog role and functions, attitude, sentiment and emotions, prosodic phrases, pitch, accentuation, etc., to the observed exchange of information.

3.2. The EVA annotation scheme

In order to capture and analyze conversational phenomena in EVA corpus, the video material is annotated by following the EVA annotation scheme that incorporates linguistic and paralinguistic features and interlinks them with nonverbal movement [37, 41]. The annotation process is performed separately for each speaker. The formal model of the annotation

http://dx.doi.org/10.5772/intechopen.75599

scheme is outlined in **Figure 2**. In addition to symbolic conversational correlations, the presented scheme also targets the analysis of the form of movement in high resolution. This allows us to test and evaluate also the low-level correlation between movement and prosody, communicative intent, and other linguistic and paralinguistic features. As a result, we can analyze the face-to-face interactions in greater detail. Further, through the extracted knowledge, we are able to pair features into complex stimuli used for triggering the generation of the conversational artifacts and to improve the understanding of the situation through multimodality. As can be seen in **Figure 2**, the annotation session per speaker is separated into symbolical annotation (e.g., annotation of function) and into annotation of the form (e.g., annotation of visualization). Each of the annotated features (linguistic, paralinguistic, and movement/shape related) is captured on a separate track and interlinked with spoken content and movement via a shared timeline. In this way we are able to analyze and search for various multidimensional relationships between conversational artifacts and identify and establish temporal and symbolic links between verbal and co-verbal features of complex multiparty interaction.

As outlined in **Figure 2**, the EVA annotation scheme has the capacity not only to establish links between features on the symbolic level but also to interlink the form of co-verbal movement and its manifestation (e.g., the visualization) with symbolic artifacts, such as dialog role, emotions, lemma, POS tags, sentence type, phrase breaks, prominence, sentiment, and semiotic intent. This is quite important for investigating into the multidimensional interlinks between various features. For instance, the co-verbal behavior may originate as a reflection of attitude/emotion or even be a supportive artifact in the implementation of the communicative function (e.g., feedback, turn taking, turn accepting, sequencing, etc.), while the verbal behavior primarily used for representation of information may also reflect attitude/emotion or be adjusted to serve as a part of the implementation of a communicative function. Through

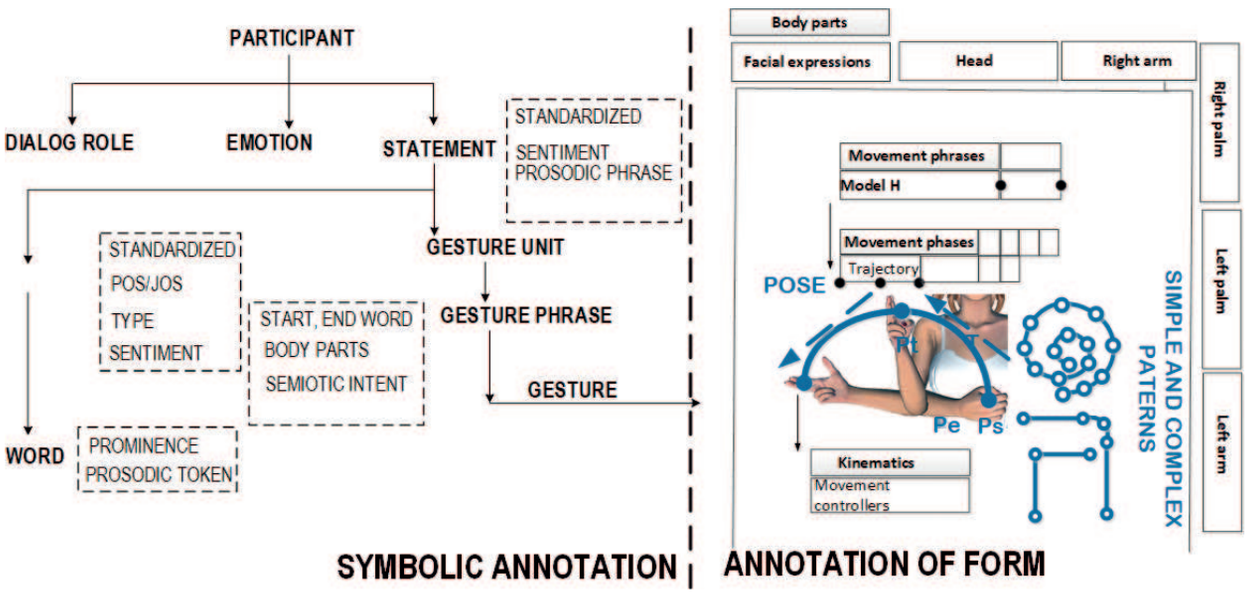

Figure 2. The formal model of the EVA annotation scheme and the topology of the annotation of the form.

the annotation scheme, all artifacts are interconnected through temporal domain and can be related among each other in numerous ways and combinations.

Overall, the symbolic annotation allows us to identify and describe in detail the nature of communicative acts performed during information exchange. The annotation of the form, on the other hand, then describes the shapes and movements generated during these symbolically defined communication acts. Thus, the concept of EVA annotation is based on the idea that symbolic relations and concepts are established on the functional/symbolic level and realized via body expressions, e.g., hand gestures (left and right arm and hands), facial expression, head movement, and gaze. During each symbolic act, the movement of each body part (hands, arms, head, and face) can be described with movement phrase, movement phases, transitions, and the articulators propagating the observed movement. The movement phrase describes the full span of movement phases, where each movement phase contains a mandatory stroke and optional preparation, hold, and retraction phases. Further, each movement phase identifies a pose at the beginning *Ps* and a pose at the end *Pe*, where poses are "interconnected" with a trajectory that identifies the path over which the observed body parts propagate from the start pose to the end pose. The trajectory *T* is a parametric description of propagation, which includes the partitioning of the trajectory *T* into movement primes (simple patterns), such as linear and arc, each defined through the intermediate poses.

4. Advanced co-verbal behavior generation by using EVA framework

The EVA behavior generation model proposed in [36] is used to convert general unannotated texts into co-verbal behavior description automatically. The model integrates several nonverbal elements that are associated (correlated) with the verbal behavior (speech). Therefore, general texts can be presented as multimodal output, consisting of spoken communication channel as well as synchronized visual communication channel. The EVA behavior generation model performs synchronization of the verbal and nonverbal elements that is necessary in order to achieve desired naturalness, in the domain of meaning (intent) and in the temporal domain. Further, the EVA model generates the co-verbal behavior descriptions and the verbal behavior simultaneously. The EVA model distinguishes between the behavior generation and behavior realization step. **Figure 3** outlines the expressive conversational behavior generation module, which consists of the following three concurrent processes: intent classification, behavior planning, and speech synthesis. The *speech synthesis process* converts general text into speech signal and also represents a source of linguistic and prosodic features that are used for planning and synchronizing the nonverbal behavior. Further, the *intent classification process* identifies the nature of the spoken content through pattern matching incorporating linguistic and prosodic features, where the intent of the input text is defined in the form of classification of linguistic expressions into semiotic classes. The result is a set of possible interpretations of the input text. Further, the *behavior planning process* involves filtering of several interpretations, the pose/gesture selection process based on target cost calculation mechanism, and the temporal synchronization step based on prosodic and acoustic features obtained during synthesizing the speech signal. As outlined in **Figure 3**, the EVA behavior generation model

http://dx.doi.org/10.5772/intechopen.75599

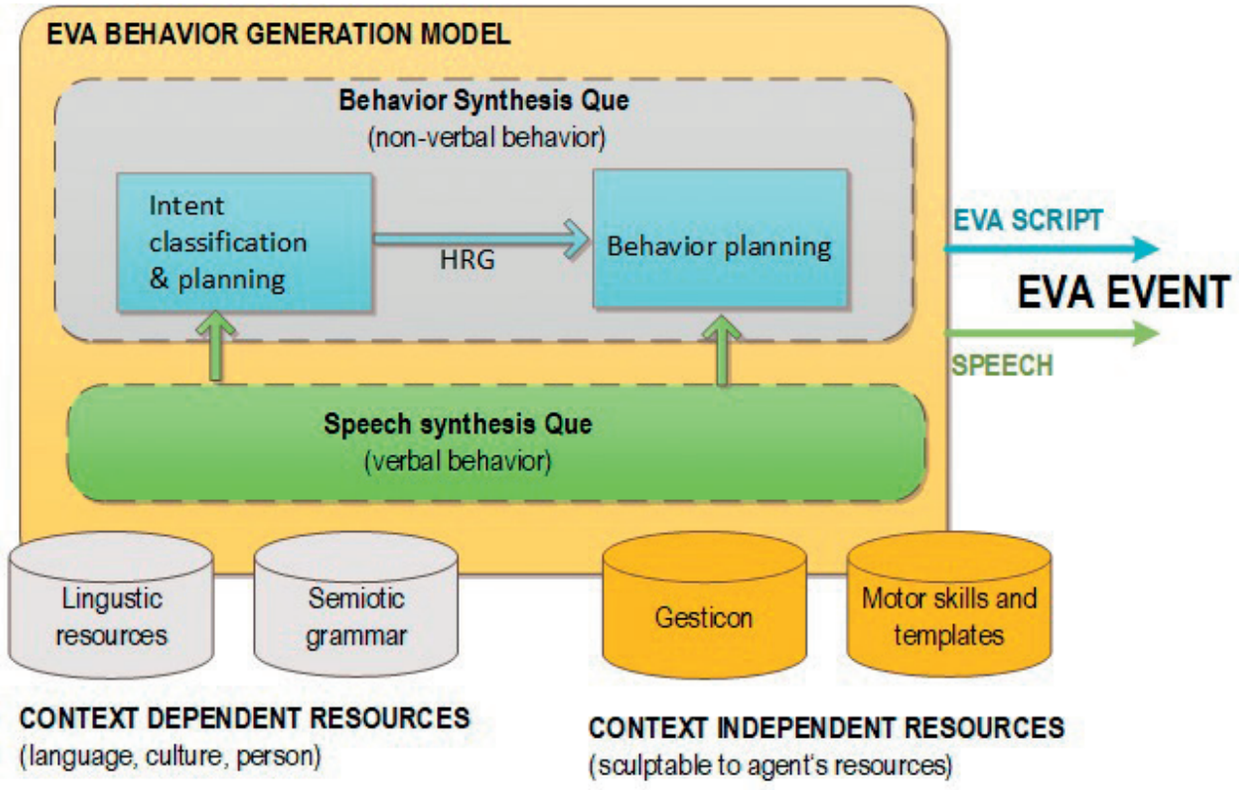

Figure 3. The architecture of the EVA behavior generation model for the generation of expressive co-verbal behavior.

facilitates language–/context-dependent and language–/context-independent resources. The language–/context-dependent resources are linguistic resources and the nonlinguistic semiotic grammar. The linguistic resources include lexicons, language models, rules, and corpuses, while the nonlinguistic grammar includes sets of semiotic rules for pairing language with communicative intent. The language–/context-independent resources are *Gesticon* and the repository of *motor skills*. *Gesticon* couples the context-independent *motor skills* with the semiotic nature of the intent. Namely, we associate a given semiotic pattern (communicative intent) with a unique set of features describing the manifestation of the shape/pose. The semiotic pattern incorporates a semiotic class/subclass, the movement phase within which the pose manifestation is observed, and the POS tag of the word represented as the nucleus of meaning. The unique set of features describing the manifestation of shape/pose incorporates body-pose identification, representing a pair of initial and final pose, a general trajectory of hand movement, semantic word relation, and minimal and maximal temporal values within which the gesture was observed to be carried out, and the number of occurrences of the given gesture that was observed in the EVA corpus. The *semiotic grammar* and *Gesticon* are created and populated with patterns and associated with unique sets based on the analysis and annotation discussed in Section 3.

The conceptual EVA behavior generation model has been actualized in the form of the EVA engine outlined in **Figure 4**. The EVA engine converts a general text into the speech signal accompanied by humanlike synchronized gesticulation and lip movement. The EVA engine

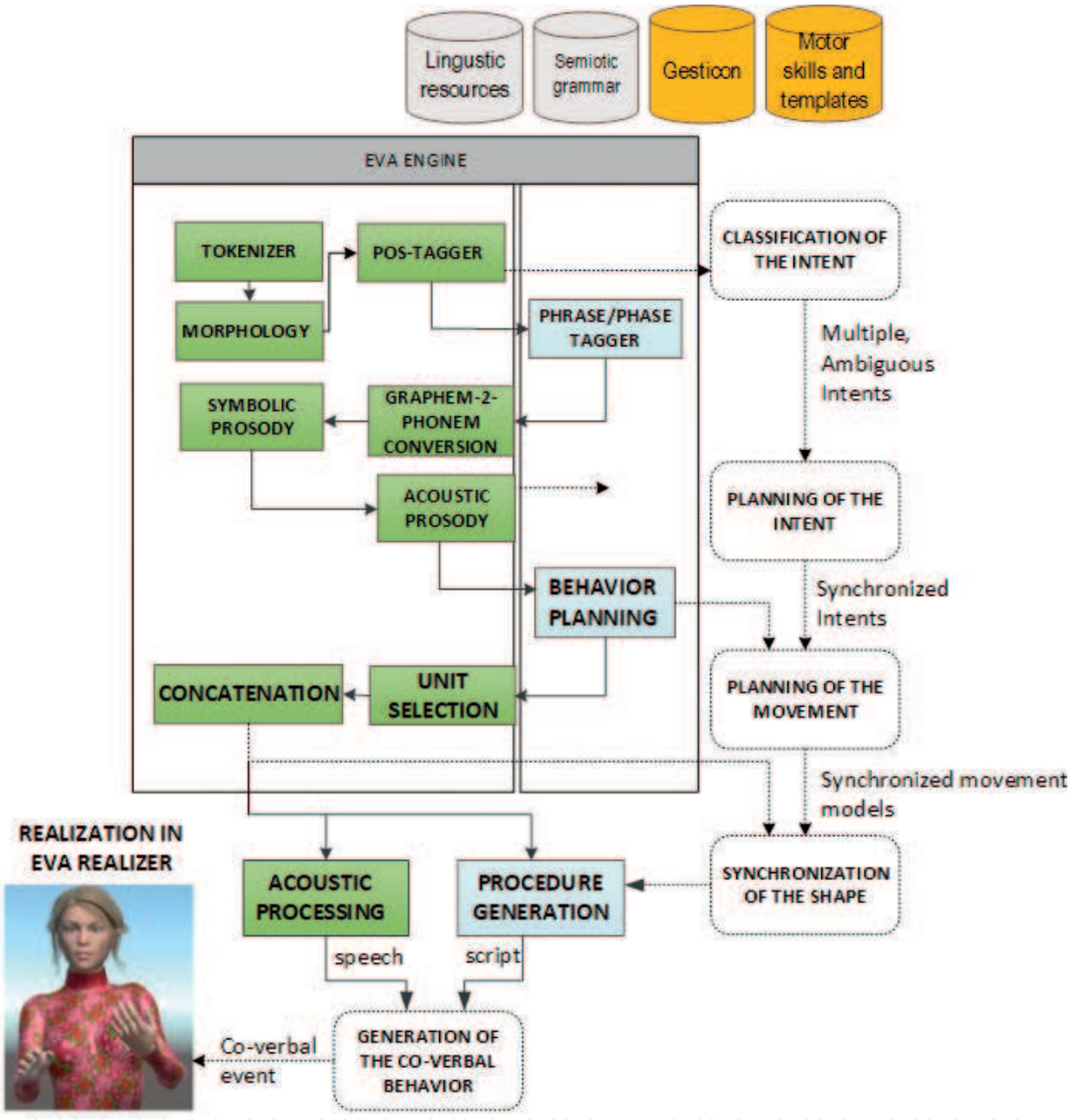

Figure 4. EVA engine: implementation of the EVA behavior generation model.

is composed of (a) processing steps for text-to-speech synthesis as proposed in the TTS system PLATTOS [42] and of (b) processing steps for expressive co-verbal behavior generation algorithm. All steps are fused into a compact processing EVA engine. In this way, the expressive co-verbal behavior generation algorithm works with the verbal modules concurrently, by sharing data, machine-trained models, and other resources. The EVA engine takes into account the latest results of research on multimodal communication, goes beyond traditional computational speech processing techniques, and facilitates heuristic and psychological models of human interaction. The algorithm for the generation of expressive co-verbal behavior implemented within the engine in **Figure 5** generates the co-verbal behavior by considering

http://dx.doi.org/10.5772/intechopen.75599

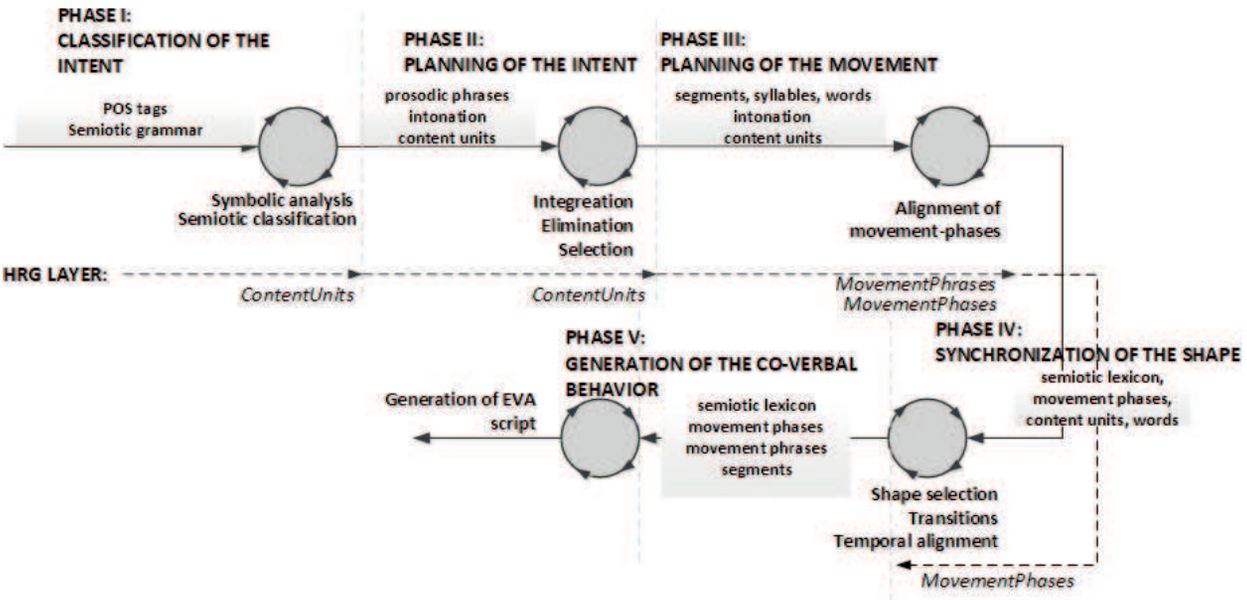

Figure 5. The algorithm for the generation of expressive co-verbal behavior.

the synchronization of intent and shape (the intent classification and planning), *the synchronization of several movements* (movement planning), and *prosodic and timing synchronization regarding speech and gestures* (synchronization of the form).

In phase I, named *classification of the intent*, the input is POS-tagged text and *semiotic grammar*. The *semiotic grammar* is used for mapping individual morphosyntactic sequence in the text onto a corresponding parametric description of the semiotic class (subclass) $Z = \{S, I, \omega_s, p_s\}$.[1] In this way, we are able to perform the classification of the intent by defining the semiotic classes and corresponding cores of meaning. The algorithm searches for the longest morphosyntactic sequence $x_{i\text{-}j}$ that can be found in the *semiotic grammar*, while the following two rules are implemented:

- *If at a specific word index the sequence* x_A *happens to be* $x_A(S) \subseteq x_B(S)$, *where both sequences belong to the same semiotic class, then sequence* x_A *must be discarded.*
- *If a sequence x at word index j is already contained in previous CU elements started at word index i and with the same semiotic intent (i < j), then it is discarded.*

In this way, for each such sequence, the content unit (*CU*) is created. The *CUs* are used to store the semiotic classification of the intent as well as those meaningful words that actually induce the shape in the stroke movement phase F_S of the gesture.

Nevertheless, sentences/statements can have multiple interpretations. Further, the classified *CU* interpretations can partly or fully overlap, therefore, introducing ambiguities and a number of inconsistencies. Thus, in phase II, named *planning of the intent*, these inconsistencies and

[1]Represents a unique set of features describing the semiotic class/subclass *S*, the index of the word(s) that represent a nucleus of meaning *I*, the distribution of the POS sequences for the specific semiotic class/subclass *ωs*, and the distribution of selecting the specific POS sequence among the semiotic classes/subclasses *ps*.

ambiguities have to be resolved by using *integration*, *elimination*, and *selection*. For this task the prosodic information (prominence, stress, prosodic phrases), as predicted by the TTS modules, is used. The prosodic information includes syllables labeled with *PA* (most prominent) and *NA* (stressed) tags that exist within minor phrases (*B2*) or major phrases (*B3*). The resolving is implanted by observing the following rules:

- *Each CU must include the most prominent syllable (PA) within a given prosodic phrase (B2 or B3), except in the case of enumeration.*
- *Each CU element must lie within the prosodic phrase (B2 or B3).*
- *Each prosodic phrase can be represented by no more than one concept of motion, i.e., with no more than one element CU.*
- *When the CU element contains the semiotic class enumeration, the CU boundaries must remain unchanged (the boundaries of prosodic phrases are not considered).*

At the end, the structure of the *intent* is uniquely defined, and the co-verbal behavior *G* can be represented as a sequence of co-verbal models *H* that are related with the *CU* as follows:

$$\hat{G} = T_m^{-1}\hat{H} = H[CU_1, t_1] \times H[CU_2, t_2] \times \ldots \times H[CU_n, t_n] \tag{1}$$

where $H[CU_i, t_i]$ describes the movement model that depends on semiotic classification and prosodic characteristics in each *CU* element. However, *internal overlapping* can still occur, when several *CU* elements contain one or more of the same words, while their boundaries lie within the same prosodic phrase. In this case we have to decide, which of the *CU* elements must be kept, since only one *CU* element is allowed within each prosodic phrase. Firstly, the algorithm removes all those *CU* elements that do not contain a word with a *PA* syllable, and if overlapping still exists, then on the remaining *CUs*, their normal distribution ω_s is considered, as calculated for the given semiotic class, when the *CU* with its maximum value is only kept. Therefore, common *CU* is created as:

$$\begin{aligned} &CU_{S(CU_m)=S(CU_m)} = f(CU_m, \ldots, CU_n) = \\ &\begin{cases} CU_m; \text{if } \omega_s(CU_m) > \omega_s(CU_n) \\ CU_n; \text{if } \omega_s(CU_m) < \omega_s(CU_n) \\ \max(len(CU_m), len(CU_n)); \text{if } \omega_s(CU_m) = \omega_s(CU_n) \end{cases} \end{aligned} \tag{2}$$

Further, in the case of *ambiguity*, there are multiple *CUs* within a single prosodic phrase but without overlapping. When these *CU*s are consecutive and classify the same semiotic class *S* (e.g., represent the same communicative intent), they are merged into a single *CU* element as follows:

$$\begin{aligned} &CU = f(CU_i, \ldots, CU_j) = \\ &\begin{cases} \text{Join } CU_k, CU_{k+1}, \ldots; \text{ if} \\ \quad CU_k.POS = CU_{k+1}.POS \wedge S(CU_k) = S(CU_{k+1}) \\ \text{Add into a set; otherwise} \end{cases} \end{aligned} \tag{3}$$

http://dx.doi.org/10.5772/intechopen.75599

The semiotic indicator *I* is defined as a set of corresponding semantic indicators contained within individual *CUs* in the sequence. Finally, *external overlapping* can occur, when some *CU* boundaries are stretched over the boundaries of the prosodic phrase. In this case the *CU* is kept only when the following rules are met:

- *The CU element includes the PA syllable, and this PA syllable lies within the boundaries of the prosodic phrase B2:* $PA \in B2 \wedge PA \in CU_i$.
- *The semiotic indicator I of the CU element lies within the boundaries of the prosodic phrase B2, i.e.,* $I \in B$.

And, only when both rules are met, the following two rules are implemented:

- *If* $CU \cap next(CU)$ *is not empty, the right boundary of the CU is set to the left boundary of the next (CU) element, in order that* $CU \cap next(CU) = \varnothing$ *is true*.
- *If* $CU \cap next(CU)$ *represents an empty set, the CU element is kept completely. Further, those words that lie outside the prosodic phrase that contains semiotic indicator I can only represent the holding phase and/or the retraction phase*.

In this case the common *CU* is created as:

$$CU = f(CU_i, CU_{i+1}) = \begin{cases} \textit{keep \& reduce; if} \\ \quad I \in B \wedge PA \in B \wedge PA \in CU_i \wedge CU_i \cap CU_{i+1} \neq \varnothing \\ \textit{keep \& extend; if} \\ \quad I \in B \wedge PA \in B \wedge PA \in CU_i \wedge CU_i \cap CU_{i+1} = \varnothing \\ \textit{remove; if} \\ \quad I \notin B \vee PA \notin B \vee PA \notin CU_i \end{cases} \tag{4}$$

In phase III, named *planning of the movement*, the movement models, which are based on *CU* units, are defined. A movement model is an animated sequence of shapes/poses that together represent a co-verbal expression. For each *H*, at least a stroke movement phase F_S has to be defined, which is aligned with the acoustic prosody information, as specified by the TTS engine. Therefore, the prosodic synchronization of movement phases is based on temporal information (regarding phoneme and pause durations). The following rule is used for the stroke movement phases F_S:

- *The stroke phase* F_S *is always performed on the PA word and is ended together with the corresponding PA syllable.*

The next step then represents the synchronization of all F_S with the gesticulation in case of *enumeration* and/or *search*, which are not directly related to the *PA* syllables, by using the following rule:

- *If the word that represents the semiotic indicator I for the specific CU does not contain the PA syllable, the NA syllable is considered in the same way instead.*

The most prominent words (with the *PA* syllable) do not necessarily represent the semiotic indicator *I* for the given *CU* element. If this is the case, the following rule is applied:

- *If the* semiotic *indicator I and the PA syllable are not represented by the same word, the stroke phase F_S must be defined by following the previous rules, but the hold movement phase F_{SH} must end with the NA syllable of the word that represents the semiotic indicator I within the given CU.*

Within the algorithm movement models, *H* is represented by movement phrase units (*MPHRs*), where each unit can contain several movement phases (*MPHs*). Further, each *MPHR* must contain at least stroke phase F_S. The syllables that occur before the stroke phase F_S are used for the preparation movement phase F_P, while the *sil* segment just before the first syllable of the F_S can be used for the hold movement phase F_{HS} (hold before stroke). And, those syllables after the F_S are used for the retraction movement phase F_R. In this way, the behavior structure is applied by the following rules:

- *The* preparation *phase F_P starts before the stroke phase F_S and lasts from the NA syllable to the beginning of the F_S.*
- *The sil segment, which can have a nonzero duration between the words with the preparation phase F_P and the stroke phase F_S, represents the so-called hold before stroke, which (if it exists) represents a ready-made idea regarding the content.*

Additionally, the created *MPHs* can be extended or merged by the following rules:

- The right boundary of movement phase (with the exception of the hold phase FH) must be a PA or NA syllable.
- The stroke phase FS and the preparation phase FP can be joined into the stroke phase FS, as this often occurs in multimodal communication (as observed in database annotations).

Nevertheless, the extensions are always limited by the boundaries of the specific *CU*. The structure of the movement models *H* is now synchronized with verbal content on the symbolic level. Further, temporal synchronization is performed by considering the durations of phonemes and pauses. The duration of individual movement phase is described by the following sum of the syllable durations that they may include:

$$t\left(MPH_i\right) = \left(\sum_{j=0}^{n-1} t\left(syl\, l_{j+k}\right)\right) \tag{5}$$

where *n* represents the number of syllables in each *MPH* unit and *k* its first syllable. F_P can be fused with the F_H phase, resulting in the following maximal duration:

$$t_{\max}\left(F_P\right) = t\left(F_P\right) + \sum_{j=k}^{n-1} t\left(z_{k-j}\right) + t\left(z_r\right) \tag{6}$$

where *n* represents the number of syllables before the F_P, *k* its first syllable, and *r* the first syllable after the F_P. Further, F_S can be fused with F_P, resulting in the maximal duration:

http://dx.doi.org/10.5772/intechopen.75599

$$t_{\max}(F_S) = t_{\max}(F_P) + t_{\min}(F_S) \tag{7}$$

Finally, F_R can be extended with subsequent syllables but only up to the last *NA* syllable:

$$t_{\max}(F_R) = t_{\min}(F_R) + \sum_{j=0}^{n-1} t(z_{k+1}) \tag{8}$$

where n represents the number of subsequent syllables, while k is the first syllable after the F_R. Temporal descriptions of movement phases define time instants when the individual shape must be fully manifested and also the time that is available for the transition between the shapes. Further, this time also restricts the set of suitable motion trajectories for the transition, as well as restricting the set of shapes.

In phase IV, named *synchronization of the shape*, the movement is then temporally aligned with the temporal features of verbal information (durations of phonemes and pauses). Based on morphosyntactic sequences, movement models, and durations of the movement phases, a lookup into *Gesticon* is carried out, in order to specify the best shapes *V* (or poses *P*) and trajectories *T* of the realization of the co-verbal behavior. Thus, a lookup for possible configurations of the embodiment within F_S phases is performed. It selects a set of probable poses *P* for each F_S. These poses are evaluated by using the *suitability functions* [43]. If there are no matched poses in the *Gesticon*, the set of most appropriate poses is selected by the CART (classification and regression tree) model, while the most appropriate pose *P* is assigned to each F_S. After defining the pose candidates on each stroke phase F_S, the poses for F_P, F_R, and F_H are also defined. At the end, the transition between the poses are also defined and aligned with given temporal and prosodic specifications. Namely, the trajectory describes the local space in which body part should move when traversing from the start to the end pose. The huge diversity of trajectories within the material demands restrictions, when describing them in the *Gesticon*. We are specifically interested in the trajectories of hands or the curve that the hand describes during the transition. The definition of the trajectory between two poses is performed by considering the temporal structure of the movement phase *MPH*, the semiotic class, the movement phase type, the morphosyntactic tags, prosodic features within the phase, and possible semantic relation. The lookup in the *Gesticon*, therefore, results in several possible trajectories. Therefore, only the most appropriate and closest to the temporal predictions on each sequence is used at the end.

In phase V, named *generation of the co-verbal behavior G*, the conversion of the defined movement models (stored within the heterogeneous relation graph (HRG)) into a procedural description can be understood as a parameterization of the animation. Each movement phrase is transformed into a symbolically, prosodically, and spatially coherent movement of an individual body part. Thus, it viably illustrates the communicative intent of the corresponding verbal segment. In order to be applied on an ECA and represented to the user, it has to be converted into a procedural description in the EVA-Script notation. Each model *H* represents simultaneous execution and is described within the block <*bgesture*>. The stroke movement phase F_S and the preparation movement phase F_P within the block <*bgesture* > represent sequences during which a change in configuration of embodiment

is actually requested. The hold movement phase F_H and the retraction movement phase F_R, however, do not require procedural description. Namely, F_H only represents a hold of the existing configuration, while F_R a retraction into a rest/neutral state. The transformation of movement model H into an EVA event (co-verbal behavior written in EVA-Script) is outlined in **Figure 6**. **Figure 6** also outlines how EVA event is applied to an ECA and then realized as a multimodal expression, which is built from the synchronized verbal and co-verbal sequences. As can be seen, the F_P is defined across the word "*bila*" (*was*), with predicted extension up to the word "*je*" (*is*). Further, the predicted duration of the F_P phase is 413 ms, while the maximum duration of the F_P is 593 ms. The F_S is defined across the *PA* word "*tako*" (*that*), with a predicted duration 300 ms and a maximal duration 893 ms. The post-stroke-hold phase F_{HS} is identified across the semiotic nucleus, the word "*velika*" (*big*), with a duration of 451 ms.

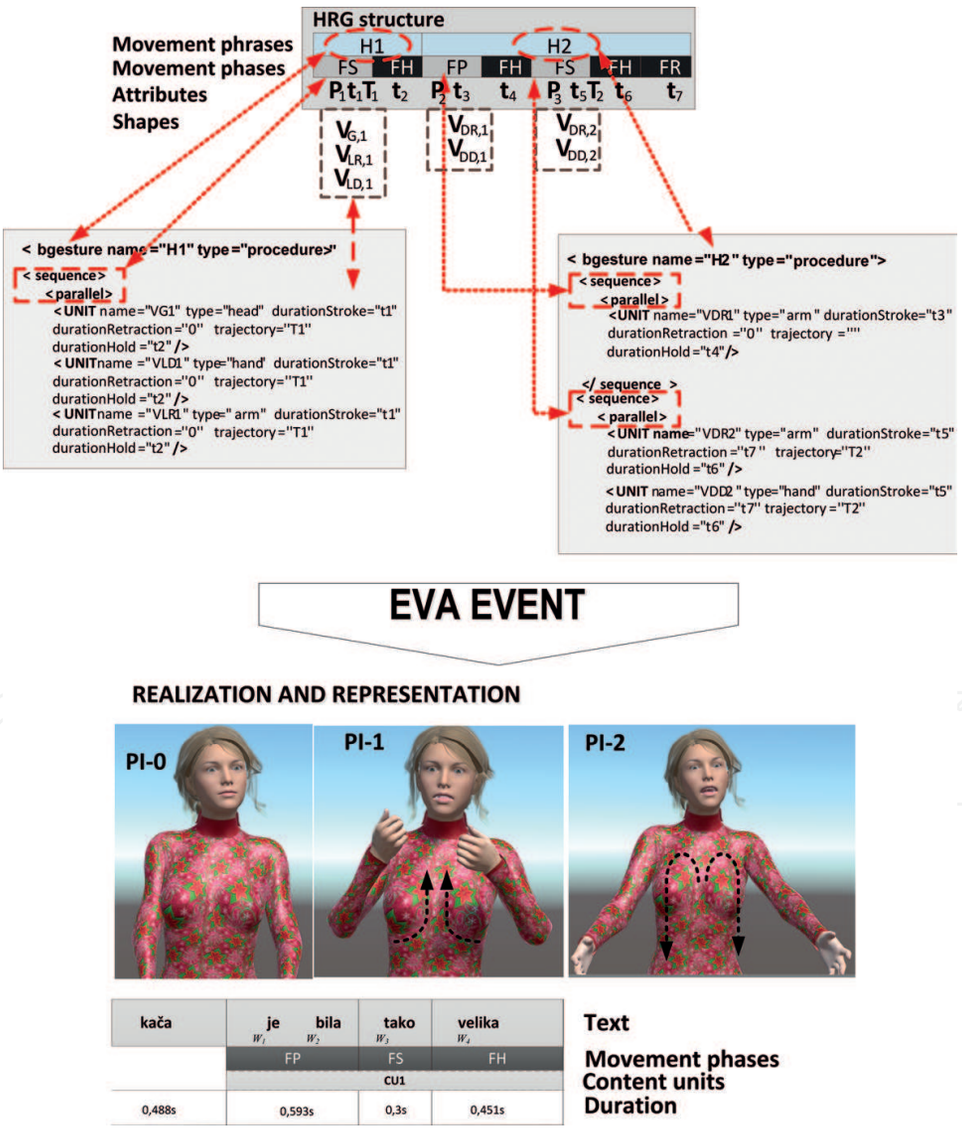

Figure 6. Realization of a sentence with the conversational agent EVA.

http://dx.doi.org/10.5772/intechopen.75599

5. Realization of expressive conversational behavior on embodied conversational agents

The proprietary *EVA realization framework* proposed in [36–37, 44] has been developed in order to be able to evoke a social response in human-machine interaction based on expressive conversational behavior generated through the previous modules of the EVA model. The framework enables machines to engage with the user on the more personal level, namely, through humanlike entity realization of multimodal interaction models, which are based on the concept of conversation. Thus, this framework integrates ECAs as virtual bodies and generates responses via their embodiment. The ECA's artificial body and articulation capabilities (embodiment) are already close to those found on real humans. From the skin, face, hands, and body posture, these virtual entities tend to look and behave as realistically as possible. ECAs also tend to imitate as many features of human face-to-face dialogs as possible and integrate them into interaction as synchronized as possible. Although the humanlike equation is mostly defined via co-verbal behavior generation model and the corpus analysis, the framework actually represents the final component through which users actually come in contact with the response. Thus, one could say that the framework brings responses to "life." Further, diversity and capacity to handle highly dynamic and interchangeable contexts of human interaction are in addition to realism of appearance, one of the key challenges of the modern ECAs. 3D tools, such as Maya, Daz3D, Blender, Panda3D, and Unity, have opened up completely new possibilities to design virtual entities, which appear almost like real-life persons. The modern behavior generators open the possibility to plan and model responses almost completely to the context of situation and collocutor. The behavior realizer, therefore, represents the bridge between the two concepts. The *EVA realization framework* also creates an environment that is capable to deliver expressive and socially colored responses in the form of facial expressions, gaze, head, and hand movement. Its architecture is outlined in **Figure 7**.

It consists of animation-parameters builder, animation-realization engines, articulated 3D models, and created 3D resources. The *animation-parameters builder* is used to understand and transform the co-verbal events into animation parameters and integrate them onto the animation execution plan. The *animation-realization engines* then realize these animation parameters through their internal renderers and display them to the user. As outlined in **Figure 7**, two animation engines are implemented, one is based on Panda 3D[2] game engine and the other based on Unity 3D game engine[3]. Each of them incorporates its own set of articulated 3D models. However, all articulated 3D models support the same movement controllers (bones and morphed shapes). Thus, any EVA event can be used by either realizer, and the result will be still practically the same. The major difference between the supported animation engines is their implementation of frame-by-frame operations. Namely, in the Panda 3D engine, frame-by-frame operations are handled internally by the renderer, while in the Unity 3D engine, the renderer only renders each frame. This means that all

[2]Panda 3D: https://www.panda3d.org/.
[3]Unity 3D: https://unity3d.com/.

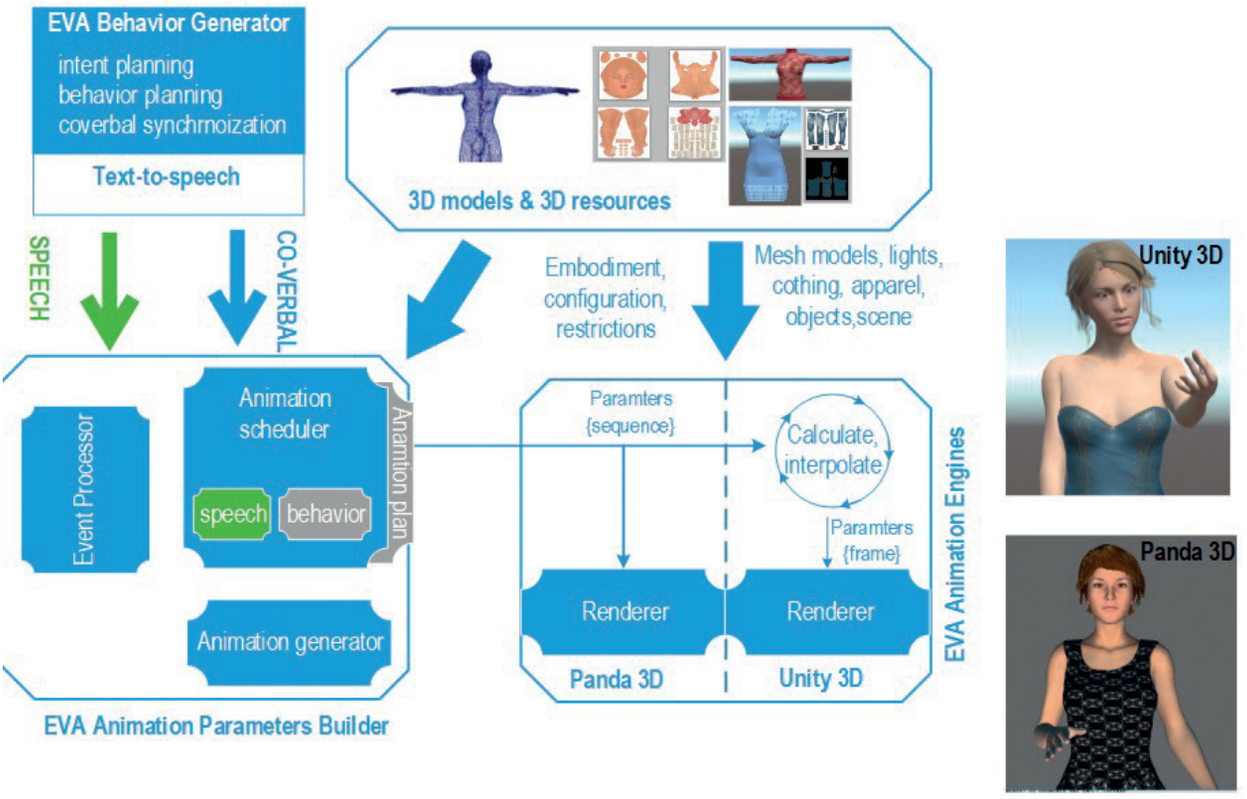

Figure 7. EVA realization framework.

calculations for the next "in-between" pose are calculated via the implanted algorithms. As a result, the Unity 3D implementation allows for controlling the scheduled animation and even animation already being executed. In order that the *EVA realization framework* realizes the generated synchronized behavior and represents it to the user, the *EVA behavior realizer* transforms conversational events into their physical representations. This is achieved by applying the co-verbal features described in the co-verbal events into the 3D resources available in the renderer. The *animation-parameters builder* translates the EVA-Script into animation parameters. This is achieved by interfacing each script's tag with the control unit or behavioral template and by extrapolating different groups of movements. Each group of movement is defined by semantic (which control units in which order), temporal (durations of stroke, hold, and retraction movement phases), and spatial features (ending position of the control unit). The main components of the *animation-parameters builder* are *event processor*, *animation generator*, and *animation scheduler*. The *event processor* intercepts and handles the conversational events. It parses event stream and checks event's type and priority. The *animation generator* then transforms the conversational behavior into animation sequences. As part of this process, the *animation generator* applies temporal and spatial constraints adjusted to the agents' articulated body. The a*nimation scheduler* then inserts the generated animation sequences into the execution plan. It handles the animation graph and feeds them to the rendering engines accordingly. Finally, after the realization of each animation sequence is completed, the *event processor* signals its status (conversational context) to the

behavior generator and optionally to the dialog handlers. Similarly, after the full realization of the behavior que, a change in conversational context event is raised, and the generation of inactive (rest) behavior is triggered.

The communication between processes within the *EVA realization framework* is implemented via event-oriented publish/subscribe model. Namely, when the *event processor* intercepts a conversational event, it firstly checks its type and priority. Afterward, it pushes it into the *animation scheduler*. When the *animation scheduler* receives the conversational event, it initiates internal interpreter in order to segment the behavior into three animation streams. The interpreter transforms the EVA-Script behavior into body part-segmented schedule of parallel/consecutively executed behavior in a form of animation streams. At the same time, the *scheduler* smoothly stops any idle behavior, destroys its handlers, and moves it to the rest pose. The overall result of the *animation-parameters builder* is, therefore, a set of animation parameters, which describe the execution of one or more animations representing the planned co-verbal act (multimodal response). The animation parameters are those features that specify how the animation engine should build its animation graph. The animation parameters are "fed" to the second component of the framework, the *EVA animation engine*. The *EVA animation engine* takes care for the transformation of animation parameters into animated sequences. The animation plan contains the co-verbal behavior que. After its que is emptied, the *animation scheduler* signals that the animation stream has been completed and will destroy animator objects, in order to release the reserved resources. After all animation segments are completed, the *animation scheduler* signals the end of the conversational event. As a result, the *event handler*, if no more co-verbal events arrive, triggers the manifestation of the idle behavior. Each animation engine transforms the animation parameters maintained in the animation plans into corresponding sequential and/or parallel movements of control points (bones or morphed shapes). Further, both animation engines provide the forward kinematics (and inverse kinematics) and animation-blending techniques, which enable for animation parameters to appear as viable behavior even on segments that have to be controlled by different gestures at the same time, e.g., simultaneously animating smile and viseme. Each gesture/expression/emotion is realized by combining different sequential/concurrent transformations of different movement controllers (embodiments) of the ECA. The EVA-Script events describe the facilitation of the movement controllers in a form of temporally defined end poses. Thus, each entry contains "next" configuration, which should be rendered over the specified temporal distribution. The in-between frames, which actually generate movement, however, are calculated and interpolated by animation engines. In case of Panda 3D engine, the render receives the required end pose and calculates the in-between frames automatically, while in the case of Unity 3D engine, the *animation handler* handles frame-by-frame operations, e.g., the render receives the next in-between configuration, which is calculated by the *animation handler* at each frame. In this way, any animation can be modified at any step, even during the execution of some step/sequence. For a smooth transition, the scheduler does not have to wait and adjust its temporal scheduling. It just has to adjust its frame-by-frame schedule and replace it with new configurations. It can actually instantiate changes instantly as they occur. It can also insert new behavior between configurations, etc. As a result, the virtual character becomes more responsive and can react to changes of the conversational, environmental, and other

contexts instantly. The agent also "remembers" what it was gesturing prior to the excited state. Additionally, it can continue with the realization of that behavior after the excited state dissipates.

6. Realization of complex co-verbal acts

When the realization framework receives some co-verbal EVA event, it transforms it into a synchronized and fluid stream of movement, performed by the following independent body parts: hands, arms, face, and head. To retain naturalness (especially regarding visual speech coarticulation) and at the same time prevent "jerky" expressions, the animation-blending techniques are used. The same animation-blending techniques are also used to realize three different types of complex emotions: emotion masking, mixed expressions, and qualified emotions. These complex emotions further intensify the expressive factor of the framework and enable the implantation of highly humanlike representation of feelings in facial region (e.g., by modulating, falsifying, and qualifying an expression with one or more elemental expressions). The modulation of expression is realized by using animation-blending techniques, based on intensification or de-intensification. Both are similar to qualification of expression and implemented through the power (e.g., stress attribute) and temporal (durationUp, durationDown, delay, and persistence attributes) components of the domain of expressivity [36–37, 44]. **Figure 8** outlines output from the *realization framework* as interpretation and realization of EVA-Script events, including several layers of complexity, and EVA-Script attributes described through the EVA-Script language. In **Figure 8**, the behavior generator defines a co-verbal act that consists of two co-verbal events. The first one resembles the end of "searching idea" event (when some idea of a solution comes to our mind), and the second one reassembles the beginning of revelation of the idea (e.g., how one starts outlining the solution to collocutors). The co-verbal behavior is described in order to be performed via full embodiment (all co-verbal artifacts), namely, by using the arms, hands, face, and head. To add an additional layer of complexity, the inner synchronization and the temporal distribution are different for each co-verbal artifact.

During the first co-verbal event (e.g., revelation), the head, face, and right hand are the dominant artifacts. Thus, they appear to "move" with most significance and power. On the other hand, the left hand moves to its targeted position slightly delayed but as fast as possible. In the second act, however, the left hand is the dominant artifact. Thus, its movement will appear as most significant, e.g., the longest duration and with most power, while the face/head and right arm/hand are moved to the position as "quietly" as possible. The overall duration of the first co-verbal act (Act 1) act is 1.567 s. During this time period, the agent has to perform a pointing gesture, by pointing to the sky and by moving its left arm to a position that is relevant for the specified pointing gesture (e.g., almost touching the torso). Additionally, the agent should express a blend of happy/surprised emotion on his/her face. As outlined in **Figure 8**, head/gaze and facial expression started to appear first (delay = 0.0 s). The two co-verbal artifacts then moved to their final configuration in 0.5 s, while the right- and left-hand movements are delayed for 0.4 s. Thus, both configurations started to form 0.1 s interval,

http://dx.doi.org/10.5772/intechopen.75599

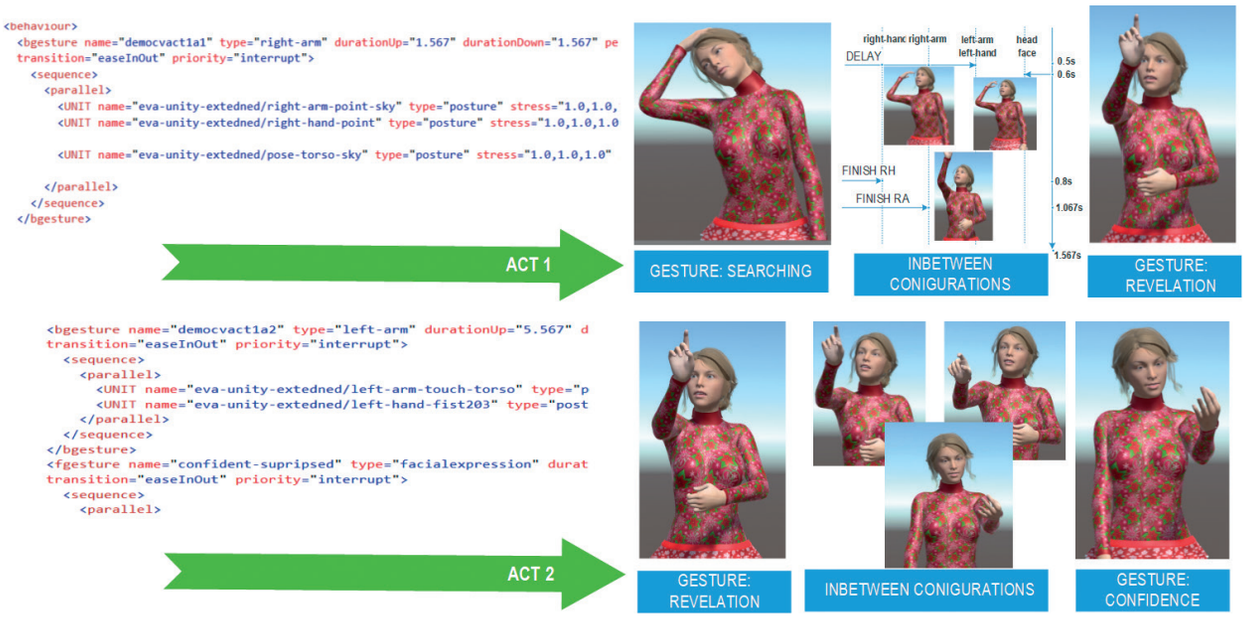

Figure 8. Realization of EVA event on ECA EVA rendered in unity-based realizer.

before the previous two co-verbal artifacts finished. The right arm finished with its animation after 0.567 s, while the right hand manifested the targeted hand shape in 0.3 s. The left hand and arm propagated to their end-configuration until the overall end of the event (e.g., for 1.167 s). Those co-verbal artifacts, which have already finished, just maintained their configuration. The second co-verbal act is targeted to last 5.567 s. During the realization of the second co-verbal act, the right arm (with hand) is regarded as less significant; therefore, it is moved to its intended rest position as slowly as possible. The left arm (with hand) is, in this situation, regarded as one of the significant co-verbal artifacts carrying some conversational meaning. The same holds true for the head and face. The left-hand movement in this case appeared as with most power in order to gain the most attention of the collocutor, and the face expressed confidence colored with excitement. Finally, by directing gaze to the collocutor, the ECA EVA prepares the conversational environment, which facilitates the full attention of the collocutor. Thus, it can start with the representation of the recently developed idea.

7. Conclusion

Natural interaction entails multiple behavior variations correlated in a dynamic and highly unpredictable setting. It also incorporates various social and interpersonal signals to "color" the final outcome. Furthermore, multimodality in interaction is not just an add-on or a style of information representation. It goes well beyond semantics and even semiotic artifacts. It significantly contributes to representation of information as well as in interpersonal and textual function of communication. In this chapter we have outlined approach to automatic synthesis of more natural humanlike responses generated based on EVA conversational model. The presented model consists of three interlinked and repetitive frameworks. The

first framework involves conversational analysis through which we analyze multiparty and spontaneous face-to-face dialogs in order to create various types of conversational resources (from rules and guidelines to complex multidimensional features). The second framework then involves an omni-comprehensive algorithm for the synthesis of affective co-verbal behavior based on the arbitrary and unannotated text. In contrast to the related research, the proposed algorithm allows for the conversational behavior to be driven simultaneously by prosody and text and modeled by various dimensions of situational, inter- and intrapersonal contexts. Finally, the predicted behavior well synchronized to its verbal counterparts has to be represented to a user in a most viable manner. Thus, the third framework in the proposed model involves implementation of co-verbal behavior realizer. In our case we have decided to fuse advantages of state-of-the-art 3D modeling tool and game engines with the latest concepts in behavior realization in order to deploy an efficient and highly responsive framework through which the generated co-verbal expressions may be represented to users via realistic and humanlike embodied conversational agents. Namely, modern behavior realizers have the capacity to support several parameters of believability of conversational behavior, such as diversity and multimodal planning, situational awareness, synthesis of verbal content, synchronization, etc. The game engines on the other hand are a powerful tool for rapid and high-quality design and rendering of virtual humanlike entities including ECAs. They enable the design and delivery of beautiful and highly realistic graphics and the efficient handling of hardware resources. To sum up, the ability to express information visually and emotionally plays a central role in human interaction and thus in defining ECA's personality, its emotional state, and can make such an agent an active participant in a conversation. However, in order to make him be perceived even more natural, the agent must be able to respond to situational triggers smoothly and almost instantly as they are perceived and by facilitating synchronized verbal and co-verbal channels. Thus, the presented model presents an important step toward generating more natural and humanlike companions and machine-generated responses.

Acknowledgements

This work is partially funded by the European Regional Development Fund and the Ministry of Education, Science and Sport of Slovenia (project SAIAL).

This work is partially funded by the European Regional Development Fund and the Republic of Slovenia (project IQHOME).

Author details

Matej Rojc*, Zdravko Kačič and Izidor Mlakar

*Address all correspondence to: matej.rojc@um.si

Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

References

[1] Luger E, Sellen A. Like having a really bad PA: The gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM. 2016. pp. 5286-5297

[2] Feyaerts K, Brône G, Oben B. Multimodality in interaction. In: Dancygier B, editor. The Cambridge Handbook of Cognitive Linguistics. Cambridge: Cambridge University Press; 2017. pp. 135-156. DOI: 10.1017/9781316339732.010

[3] Li J, Galley M, Brockett C, Spithourakis GP, Gao J, Dolan B. A Persona-Based Neural Conversation Model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany; 2016; pp. 994-1003

[4] Porcheron M, Fischer JE, McGregor M, Brown B, Luger E, Candello H, O'Hara K. Talking with conversational agents in collaborative action. In: Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM. 2017. pp. 431-436

[5] Bonsignori V, Camiciottoli BC, editors. Multimodality Across Communicative Settings, Discourse Domains and Genres. Cambridge Scholars Publishing; 2016. Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK. ISBN (10): 1-4438-1107-6. ISBN (13): 978-1-4438-1107-1

[6] Kopp S, Bergmann K. Using cognitive models to understand multimodal processes: The case for speech and gesture production. In: The Handbook of Multimodal-Multisensor Interfaces. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool; 2017. pp. 239-276

[7] McNeill D. Why We Gesture: The Surprising Role of Hand Movement in Communication. Cambridge: Cambridge University Press; 2016. ISBN-10: 1316502368. ISBN-13: 978-1316502365

[8] Davitti E, Pasquandrea S. Embodied participation: What multimodal analysis can tell us about interpreter-mediated encounters in pedagogical settings. Journal of Pragmatics. 2017;**107**:105-128

[9] Hazel S, Mortensen K. Embodying the institution—Object manipulation in developing interaction in study counselling meetings. Journal of Pragmatics. 2014;**65**:10-29

[10] Vannini P, Waskul D, editors. Body/Embodiment: Symbolic Interaction and The Sociology of the Body. New York, NY, USA: Ashgate Publishing, Ltd.; 2012. ISBN: 1409490610, 9781409490616

[11] Colletta JM, Guidetti M, Capirci O, Cristilli C, Demir OE, Kunene-Nicolas RN, Levine S. Effects of age and language on co-speech gesture production: An investigation of French, American, and Italian children's narratives. Journal of Child Language. 2015;**42**(1):122-145

[12] Esposito A, Vassallo J, Esposito AM, Bourbakis N. On the amount of semantic information conveyed by gestures. In: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI); IEEE. 2015. pp. 660-667

[13] Kendon A. Gesture: Visible Action as Utterance. Cambridge University Press; 2004. ISBN 0 521 83525 9. ISBN 0 521 54293 6

[14] Zhao R, Sinha T, Black AW, Cassell J. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In International Conference on Intelligent Virtual Agents; Springer International Publishing. 2016. pp. 218-233

[15] Pejsa T, Gleicher M, Mutlu B. Who, me? How virtual agents can shape conversational footing in virtual reality. In: International Conference on Intelligent Virtual Agents; Cham: Springer. 2017. pp. 347-359

[16] Allwood J. A framework for studying human multimodal communication. In: Coverbal Synchrony in Human-Machine Interaction. Boca Raton; London; New York: CRC Press; 2013. cop. 2014. XIV, 420 str., ilustr. ISBN 1-4665-9825-5. ISBN 978-1-4665-9825-6

[17] Bozkurt E, Yemez Y, Erzin E. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. Speech Communication. 2016;**85**:29-42

[18] Chen CL, Herbst P. The interplay among gestures, discourse, and diagrams in students' geometrical reasoning. Educational Studies in Mathematics. 2013;**83**(2):285-307

[19] Holler J, Bavelas J. In: Breckinridge Church R, Alibali MW, Kelly SD, editors. Multimodal Communication of Common Ground. Why Gesture? How the Hands Function in Speaking, Thinking and Communicating. Vol. 7. 2017. pp. 213-240

[20] Poggi I. Hands, Mind, Face and Body: A Goal and Belief View of Multimodal Communication. Berlin: Weidler; 2007. ISBN (10): 3896932632. ISBN (13): 978-3896932631

[21] Yumak Z, Magnenat-Thalmann N. Multimodal and multi-party social interactions. In: Context Aware Human-Robot and Human-Agent Interaction. Switzerland: Springer International Publishing; 2016. pp. 275-298

[22] Kuhnke F, Ostermann J. Visual speech synthesis from 3D mesh sequences driven by combined speech features. In 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE. 2017. pp. 1075-1080

[23] Peng X, Chen H, Wang L, Wang H. Evaluating a 3-D virtual talking head on pronunciation learning. International Journal of Human-Computer Studies. 2018;**109**:26-40

[24] Wang N, Ahn J, Boulic R. Evaluating the sensitivity to virtual characters facial asymmetry in emotion synthesis. Applied Artificial Intelligence. 2017;**31**(2):103-118

[25] Gibet S, Carreno-Medrano P, Marteau PF. Challenges for the animation of expressive virtual characters: The standpoint of sign language and theatrical gestures. In: Dance Notations and Robot Motion. Switzerland: Springer International Publishing; 2016. pp. 169-186

[26] Tolins J, Liu K, Neff M, Walker MA, Tree JEF. A verbal and gestural corpus of story retellings to an expressive embodied virtual character. In LREC. 2016

[27] Ochs M, Pelachaud C, Mckeown G. A user perception-based approach to create smiling embodied conversational agents. ACM Transactions on Interactive Intelligent Systems. 2017;**7**(1):33. DOI: DOI: 10.1145/2925993, article 4 (January 2017)

[28] Bellamy RK, Andrist S, Bickmore T, Churchill EF, Erickson T. Human-agent collaboration: Can an agent be a partner? In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM. 2017. pp. 1289-1294

[29] Neff M. Hand Gesture Synthesis for Conversational Characters. Handbook of Human Motion. Switzerland: Springer International Publishing; 2017. pp. 1-12. ISBN: 978-3-319-30808-1

[30] Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: A scoping review. Journal of Medical Internet Research. 2017;**19**(5):e151, pp.1-17

[31] Rojc M, Presker M, Kačič Z, Mlakar I. TTS-driven expressive embodied conversation agent EVA for UMB-SmartTV. International Journal of Computers and Communications. 2014;**8**:57-66

[32] Shaked NA. Avatars and virtual agents—Relationship interfaces for the elderly. Healthcare Technology Letters. 2017;**4**(3):83-87

[33] Mlakar I, Kačič Z, Rojc M. A corpus for investigating the multimodal nature of multi-speaker spontaneous conversations–EVA corpus. WSEAS Transactions on Information Science and Applications. 2017;**14**:213-226. ISSN 1790-0832

[34] Mlakar I, Kačič Z, Rojc M. Describing and animating complex communicative verbal and nonverbal behavior using Eva-framework. Applied Artificial Intelligence. 2014;**28**(5): 470-503

[35] Shamekhi A, Czerwinski M, Mark G, Novotny M, Bennett GA. An exploratory study toward the preferred conversational style for compatible virtual agents. In: International Conference on Intelligent Virtual Agents. 2016. pp. 40-50

[36] Rojc M, Mlakar I, Kačič Z. The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm. Engineering Applications of Artificial Intelligence. 2017;**57**:80-104

[37] Rojc M, Mlakar I. An expressive conversational-behavior generation model for advanced interaction within multimodal user interfaces. In: Computer Science, Technology and Applications. New York: Nova Science Publishers, Inc.; 2016, cop. XIV, p. 234 str. ISBN 978-1-63482-955-7. ISBN 978-1-63484-084-2

[38] Pelachaud C. Greta: An interactive expressive embodied conversational agent. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems; May 2015. pp. 5-5

[39] Mondada L. New challenges for conversation analysis: The situated and systematic organization of social interaction. Langage et Societe. 2017;**2**:181-197

[40] Velentzas J, Broni DG. Communication cycle: Definition, process, models and examples. In: Proceeding of the 5th International Conference on Finance, Accounting and Law (ICFA‟ 14); Vol. 17. 2014. pp. 117-131

[41] Mlakar I, Kačič Z, Rojc M. Form-Oriented Annotation for Building a Functionally Independent Dictionary of Synthetic Movement, Vol. 7403. Berlin; New York: Springer; 2012. pp. 251-265

[42] Rojc M, Mlakar I. Multilingual and multimodal corpus-based text-to-speech system PLATTOS. In: Ivo I, editor. Speech and Language Technologies. Rijeka: InTech; 2011. ISBN: 978-953-307-322-4

[43] Rojc M, Kačič Z. Gradient-descent based unit-selection optimization algorithm used for corpus-based text-to-speech synthesis. Applied Artificial Intelligence. 2011;**25**(7):635-668

[44] Mlakar I, Kačič Z, Borko M, Rojc M. A novel unity-based realizer for the realization of conversational behavior on embodied conversational agents. International Journal of Computers. 2017;**2**:205-213. ISSN: 2367-8895