TTS-driven Synthetic Behaviour-generation Model for Artificial Bodies

Regular Paper

Izidor Mlakar1,*, Zdravko Kačič 2 and Matej Rojc 2

1 Humatronik d.o.o., Slovenia
2 University of Maribor, Faculty of Electrical Engineering and Computer Science, Slovenia
* Corresponding author E-mail: izidor.mlakar@amis.net

Received 15 Jun 2012; Accepted 23 Jul 2013

© 2013 Mlakar et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract Visual perception, speech perception and the understanding of perceived information are linked through complex mental processes. Gestures, as part of visual perception and synchronized with verbal information, are a key concept of human social interaction. Even when there is no physical contact (e.g., a phone conversation), humans still tend to express meaning through movement. Embodied conversational agents (ECAs), as well as humanoid robots, are visual recreations of humans and are thus expected to be able to perform similar behaviour in communication. The behaviour generation system proposed in this paper is able to specify expressive behaviour strongly resembling natural movement performed within social interaction. The system is TTS-driven and fused with the time-and-space efficient TTS-engine, called ‘PLATTOS’. Visual content and content presentation is formulated based on several linguistic features that are extrapolated from arbitrary input text sequences and prosodic features (e.g., pitch, intonation, stress, emphasis, etc.), as predicted by several verbal modules in the system. According to the evaluation results, when using the proposed system the synchronized co-verbal behaviour can be recreated with a very high-degree of naturalness, either by ECAs or humanoid robots alike.

Keywords TTS-driven, ECA, Synthetic Behaviour Generation, Humanoid Robot Behaviour, Multi-modal Interaction

1. Introduction

Communication is a concept of information exchange incorporating visual perception, speech perception and the understanding of meaning. The complementary nature of information, as provided by the combination of speech and co-verbal gestures, has already been well researched. Researchers have shown that listeners instinctively combine acoustic, phonological and semantic information to correctly identify what is being said [1]. Further, the interaction between different entities (e.g., human-human, human-robot and human-avatar) reflects emotions [2], idiosyncratic features [3] and communicative functions [4]. Communication is, therefore, multi-modal. Most researchers agree that by co-aligning the verbal and non-verbal features, communication is better understood and faster in achieving its purpose (e.g., in inducing a social response). ECAs, such as in [5-8], are a well-researched term in the context of more natural human-machine interaction.
Several models of co-aligning the verbal and non-verbal features have already been discussed with regard to ECAs, ranging from speech-driven approaches to text-driven approaches [9, 10]. The behaviour overlaid by agents is usually limited to lip-syncing and facial expressions [11], or else incorporates gestures based on scenario-oriented behaviour generation engines, semantically tagged text and descriptions of behaviour in mark-up-languages, such as BML [12], MURML [13], FML [14] and APML [15]. The idea of controlling humanoid robots by similar principles as virtual-agents arises from the fact that most ECAs are controlled by similar elementary structures (e.g., a kinematic (skeletal) model). In the field of social robotics, the co-alignment of speech and non-verbal expressions represents an important and challenging task. Where robots tend to look more “human”, they are also expected to make some communicative gestures and exhibit a degree of expressiveness [16]. Such robots should be able to communicate by at least using their body (e.g., arms, hands, torso, and head) [17] and by expressing some meaningful visual signals (e.g., to illustrate spoken information, to signal a direction, etc.). For the purpose of generating expressive co-verbal gestures, several techniques have been researched. For instance, the puppeteer technique [46] might be used to compute and drive the movement performed by the robot. This technique enables the computation of walking (leg movement), arm-swinging (shoulder movement used for expressing balance), arm moving (in order to reach an “object” space) and collision avoidance. However, multimodal communication is much more than just the recreation of movement. In particular, it requires different body-parts to form and express a meaning. For such tasks, robotic systems require a large degree of flexibility in control and efficient mechanisms to adapt and drive the control in such a way whereby robots can also achieve communicative goals.

There has been previous work on speech and co-verbal gesture synchronization. The most elementary approach is to use predefined speech and specific movement for each co-verbal sequence. Such systems use hand-crafted animations or animations generated by capture technology. However, such approaches require extensive effort in designing new animations for each new speech sequence. More affordable and efficient are text-driven approaches. For example, in [49] and [50] researchers have used the syntactic and semantic structure of the text along with additional domain knowledge and mapped them to various gestures. Similarly, in [51] researchers present an approach that is based on lexical affiliation (semantics and syntax) and an underlying virtual agent platform GRETA [26]. Gestural prototypes are described symbolically (in the form of intentions and emotions), based on the annotation of human-human dialogue and extended to be used by a robotic unit NAO. The prototypes are stored in a gesture database, called a ‘lexicon’. Given a set of intentions and emotional states to communicate, the system selects from the lexicon and executes expressive robot behaviour. However, the aforementioned approaches do not consider the prosodic features in the verbal speech. The approach presented in [52] selects animation segments from a motion database based on the audio input, and synthesizes these selected animations into a gesture animation. In [53], the authors present a gesture generator that produces conversation animations for virtual humans conditioned on prosody. The gesture generator derives features describing human gestures and constrains the animation generations to be within the gesture feature space. The prosody is used as a motion transition controller.

Gestures do not have a linguistic nature and have no grammatical rules to shape and propagate the movement. For instance, two shapes produced together do not necessarily represent a gesture expressing complex meaning. Furthermore, several gestures represented by different shapes may represent a similar meaning, whereas gestures that are quite similar in shape and propagation may represent a totally different meaning.

The goal of this paper is to propose an automatic, TTS-driven gesture generation process that maps gestures to speech by using linguistic and prosodic features on general text and semiotics. The proposed approach links speech and co-verbal gestures through intention, affiliation and prosody. In contrast to most similar text-driven approaches, the intent is automatically derived by using semiotic classification based on morphological structure, semiotic grammar and prosody. The intent and prosody features further identify those words that carry meaning (meaningful words) and those surrounding words that further shape the meaning into viable, fluid co-verbal movement.

The paper is structured as follows. Firstly, the research relating to communicative behaviour performed by ECAs and humanoid robots is discussed. The related-works section is followed by the proposed non-verbal behaviour generation system. It is well known that human conversations represent the most accurate source for understanding and describing the nature of behaviour. The discussion is, therefore, followed by a methodology of how expressive resources and behaviour rules can be formed based on a video-corpus of multi-speaker conversational dialogue. This is followed by a detailed description of the proposed algorithm, the used data structures and all the processing steps that have to be performed within the system. Additionally, we also discuss how several types of synchronization are implemented. Finally, the experimental results are
presented by performing the re-creation of natural co-verbal behaviour descriptions on an ECA - called ‘EVA’ [18] - and by simulating the robot behaviour using the iCub framework [19]. At the end, the evaluation process is described and discussed and conclusions are drawn.

2. Related Works

Research on human-like humanoid robots’ communicative behaviour (in the area of social robotics) represents a very exciting topic. In general, it is based on theories of communicative behaviour [4, 20-23] that are transformed into specifications of robotic (physical) movement/embodied (virtual) movement. Several challenges, however, can be found in the limited domain of temporal dynamics and the limited domain of overlaid shape imposed by humanoid robots. In order to co-align the verbal and non-verbal information, an ECA-based communicator engine ACE has been used in [24]. Their integrated model of a speech-gesture production model is based on the following three steps: content-planning, behaviour-planning and behaviour-realization. The displayed (performed) behaviour is then specified externally using the MURML mark-up language [13]. The WBM controller framework [25] is then used in order to adapt the virtual specification of expressions into the robot’s motor controls. BEAT [49] is a rule-based system that can - based on the input text - generate co-verbal gestures. The system analyses the syntactic relation between the surface text and the gestures. The BEAT maps sets of gestures and text based on the tree structure (containing information, such as clauses, themes, objects and actions) and a knowledge base containing additional information about the world. In [26, 51], the authors present a story-telling humanoid robot with human communicative capabilities that performed in a similar way. The co-verbal alignment of non-verbal behaviour is implemented by using external behaviour-planning and a behaviour-realizer engine [7]. The different phases of behaviour generation are linked through BML and FML mark-up languages. The gesture prototypes used by the system are described symbolically (in the form of intentions and emotions), based on the annotation of human-human dialogue, and extended to be used by a robotic unit NAO. The systems in [49] and [51] provide a clear distinction between the communicative intent embedded in the surface text and the realization of gestures. This is achieved by integrating a nonverbal behaviour generator (NVBG) [50]. In order to enrich humanoid robot with communicative capabilities, the authors in [16] try to take advantage of the similarities between a humanoid robot and an articulated ECA. Namely, they both share a similar structure (movement controller-oriented kinematic model). The system addresses the representational gestures [53], e.g., movements pointing to a referent in the physical environment (deictic gestures) or gestures describing a concept with the movement or shape of the hands (iconic gestures). The concept underlying their multi-modal production model is based on an empirically suggested assumption [54]. Similar to our own approach, they assume that speech and gesture are produced in successive segments. Each of these segments represents a single idea. However, their approach involves prosody and speech-features in a post-production (realization) phase performed by ACE [24]. The speech features and phonological structure are mostly used for temporal synchronization and are not involved in the definition of intent (and/or meaning). In human interaction, the same speech spoken in different manners can express different meanings [32]. The intent may, therefore, be partially identified through prosody.

Several studies also show that the kinematic features of a speaker’s gestures (e.g., such as speed and acceleration) are also correlated with the prosody of speech [35].

Current research on communicative behaviour within the scope of social robotics mostly suggests a concept of three independent systems providing content-planning (what to display), behaviour-planning (how to display) and a behaviour-realization system (the physical/virtual generation of the artificial behaviour). The three processes are mostly separated in different engines and interlinked using mark-up-languages. However, in [27] a text-driven approach is suggested. It is based on text processing and semiotic indicators, such as content and planning indicators. Their gesture-model processes input text as an internal process. The input text is POS-tagged. The semantic correlation used to establish the correlation of speech and gesture is based on additional semantic tags assigned to the input text. There are also some more general platforms that drive human-like movement. For instance, in [28] a set of techniques for the movement of upper body, extendable to several different robotic platforms, is presented, while in [29] a lip-sync concept adaptable to service-robotics is presented.

The major advantages of the proposed algorithm are that it tries to combine intent and the prosodic and syntactic features of speech into a semiotic grammar. The semiotic grammar classifies gestures into combinable semiotic dimensions, as proposed by McNeil [22] and Pierce [21]. In contrast to the related research, the proposed algorithm incorporates language-dependent rules based on morphology, syntax, prosodic and semantics. It automatically identifies the intent and any meaningful words. The gesture prototypes used are extrapolated based on the video-annotation of communicative behaviour [36] and grouped into a semiotic space. The system also does not require any semantic or symbolic tagging of the input text, and the co-verbal gesture selection and gesture planning are fused into a single engine with the corpus-bases TTS system PLATTOS [30] in an efficient and flexible way. Therefore, no separate
The engine is required for speech generation and the extrapolation of prosodic features. The disadvantage of the system could be its language dependency, since both gestures and speech-synthesis systems are language-dependent and require additional language-dependent resources. Nevertheless, language-dependent resources are separated from the core TTS-driven behaviour generation algorithm by using the FSM formalism as proposed in [30]. Therefore, although the system is currently developed for the Slovenian language, it can also be adapted for other languages when the required language resources are available. Currently, for the PLATTOS TTS engine Slovenian language resources, and for co-verbal gestures, annotations on a talk-show-database in the Slovenian language are available. However, when these resources are also available for other language (e.g., specifying additional language-dependent semiotic rules for intent identification, content selection and/or gesture planning, etc.), the modular architecture of the system allows for the flexible reconfiguration of the system. Therefore, the proposed TTS-driven behaviour generation system can also be used for other languages. To summarize, the core platform of the proposed algorithm is TTS PLATTOS, which is further extended with several additional modules devoted to gesture synthesis. The system also performs three types of synchronization between speech and gestures:

- **Synchronization of the form:** the meaning expressed by the gesture is compatible with the one expressed by the connected speech fragment. Shapes preceding and following the dominant expression are non-mandatory and depend on the temporal and semantic features of the surrounding meta-text.
- **Synchronization of propagation:** the meaningful part of a gesture (the so-called “stroke”) co-occurs with (or slightly proceeds) the prosodically most prominent segment of the accompanying speech (alignment is performed on the level of phoneme/viseme, syllable, word and word-phrase levels).
- **Pragmatic synchronization:** gesture and speech work together to achieve the goals of communication (the alignment of consecutive conversational expressions and verbal information into a communicative act).

### 3. TTS-driven Non-verbal Behaviour Generation System

The proposed TTS-driven non-verbal behaviour generation system in Figure 1 is able to transform general input text into a speech signal that is co-aligned with very close-to-human-like gestures (including the head, hands, arms and torso). As can be seen, several external language-dependent resources are needed: an expressive dictionary of gestures, language resources related to text-to-speech synthesis, and a general input text. The system generates at the end two outputs: a synthesized speech signal and a hierarchical expressive behaviour specification (co-aligned with the generated speech signal) in EVA-Script format [18]. Both outputs can be fed into an animation/realization engine (e.g., a personalized multi-modal output on a robotic platform or a conversational agent).

In general, any information about the form of movement (content) or the co-alignment with generated speech (presentation) is assumed to be extrapolated from the input text only. The proposed TTS-driven algorithm can improve the gesture generation performed by text-driven approaches. In particular, the algorithm uses prosodic and linguistic features (e.g., morphology, part-of-speech tags, segment durations, pauses positions and durations, phrase break positions and their strength, syllables and prominence, etc.) available within general TTS engines. Further, in order that the system’s outputs can be used by different virtual/physical interfaces at interactive speeds, it is most efficient when the TTS-driven algorithm synthesizes the co-verbal gestures and corresponding speech signal simultaneously. In the proposed system, this is achieved by fusing the non-verbal expressions’ synthesis stream with the TTS engine’s verbal expressions’ synthesis.
stream. In this way, both streams retain modularity and are efficiently interlocked within a common processing chain, enabling for much better synchronization possibilities.

The resulting system is completely modular, time-and-space efficient, and consists of several deques performing all verbal and non-verbal expressions’ processing steps. The verbal expressions’ synthesis stream is used to linguistically analyse the general input text, predict the prosodic features, perform the unit selection algorithm and synthesize the speech signal as presented in [30]. The non-verbal expressions’ synthesis is used to select suitable visual content, the form of presentation and for synchronization with the generated speech signal. The co-verbal gestures are finally transformed into a form understandable by the actors that perform the interaction in multi-modal-interfaces (either humanoid robots or ECAs).

Further, the language-dependent resources for the TTS modules and the gesture generation modules are separated from the language-independent engine. In this way, the system can be used for several languages, providing the language-dependent resources. The language-dependent resources for the TTS modules include lexicons, corpora, linguistic rules, several machine-learned models and an acoustic database [30]. The non-verbal part relies on language-dependent resources, such as lexical affiliates (including gesture model relations and gesture prototypes classified into semiotic dimensions) and semantic/semiotic rules used in order to identify the intent and semiotic dimension. The gesture prototypes are stored in gesture repositories for the ECAs and robots.

4. Building the Expressive Communicative Capacities of Virtual/Physical Agents

The ECAs’ and robots’ gesture repositories in Figure 1 represent the description of the prototypical shape of movement classified into a semiotic dimension. In order to construct the prototypical movement patterns - exhibiting close-to human-like patterns - the proper audio/video data has to be available. Video-corpora are best for performing such analysis, especially when they contain spontaneous interaction and include multi-speaker dialogues. The multi-modal corpus for such analysis was developed for the Slovenian language, as proposed in [36]. The communicative behaviour was then studied and described expressively at a high level of detail. The results of the annotation were stored as lexical affiliates. The affiliates contain the elementary linguistic relations between the shape, movement-phases [22], classification into semiotic dimensions and identification of the meaningful word. In the following subsection, the construction of a gesture dictionary based on lexical affiliates is discussed in more detail.

4.1 The lexical affiliates

The lexical affiliates have been described by Schegloff as the “word of words deemed to correspond most closely to a gesture in meaning” [23]. The affiliates are used in order to establish explicit semantic relationships between speech and gestures. The semantic relation used in the proposed system is limited to the identification of meaningful words (e.g., the words and the gesture were observed to co-align with). In addition, the affiliates used within our system also symbolically describe the gesture prototype in the form of semiotic classification (e.g., intention, morphology). The most natural way to establish these relations is by observing natural conversational behaviour (gesture analysis by the annotation of a multi-modal corpus). In particular, the corpus-based nonverbal behaviour models are, in general, the closest possible approximations of natural behaviour. The annotation schema proposed in [36] is already oriented towards text-centric (text-driven) synthesis for more “natural” human-like co-verbal gestures. The proposed topology and the formal model of the annotation as shown in Figure 2 suggest that each expression be encoded as a series of consecutive prototypical shapes encoded with additional linguistic information (e.g., segmentation hypothesis [24]). The formal model of the annotation schema captures semantic and semiotic features separately for each body-part and in the form of movement phases and movement phrases [22]. Although there are some significant differences between speaker-styles (e.g., intensity, frequency, particular shapes used, etc.), some general trends can be identified [27]. For instance, specific gesture prototypes have the tendency to co-occur within the same semiotic dimensions and over certain meaningful words, phrases or sequences. If not done so semantically, the gesture prototypes may be linked based on similar morphological patterns representing a semiotic dimension.

During the annotation procedure, the gesture prototypes were identified in the Slovenian multi-modal database and classified into the following semiotic dimensions:

- **Iconic expressions**: gestures indicating objects, processes, metaphorical or qualitative meanings. Gestures that indicate (outline) concrete objects are usually triggered by an isolated/intonated noun or a preposition-noun (PN). The “object” gestures are directly co-aligned with the trigger (beginning and the end of the noun). The gesture illustrating processes are triggered by verb or verb-preposition (VP) word phrases, and are synchronized with the beginning and the end of the verb. “Metaphorical” (qualitative) gestures are triggered by phrases containing adjectives or noun-preposition (NP) combinations. In general, the “metamorphic” gestures are usually co-aligned with the beginning and the end of the adjective/NP or adverb, if an adverb-adjective combination is indicated [42, 47].
• **Symbolic expressions (emblems):** gestures that represent some sort of a conventional expression and/or conventional sign. These gestures are triggered by conventionalized key-word phrases. The “symbolic” gestures are, in general, co-aligned with the beginning and the end of the conventionalized phrases.

• **Indexical (deictic) expressions:** gestures that represent determination, reference to the addressee, a direction, and any turn-taking communicative function in the dialogue [43]. “Determination” gestures are triggered by typical determiners and key-words, such as ‘your’, ‘his’, ‘mine’, etc. The references to objects/persons (addressee) are triggered with pronouns and, therefore, keywords, such as ‘you’, ‘me’, ‘this’, ‘that’, etc. These gestures are usually directly aligned with the trigger. The turn-taking gestures are triggered by turn-taking signals [44, 45] and expect some-sort of co-speaker response (e.g., longer pauses, phrases intonated as questions, such as *alright?*, *ok?*, etc.). Such gestures are co-aligned over the whole trigger phrase and, in general, signal an existence of a pause (hold) at the end of the conversational expression [43-45].

• **Adaptors (manipulators):** gestures that address/outline speech flow and interruptions in the flow. The most common interruptions are include speakers searching for words in the lexical dictionary and instances of thinking or re-thinking, etc. These gestures are usually triggered by fillers (phrases such as *hmm…*, *oh…*, *well…*, etc.), and are co-aligned with the triggering filler. Additionally, these gestures will also maintain their shape during the silent period (if it exists), following the trigger until the next phrase is produced. The following word-phrase will usually be also accompanied by a gesture movement (e.g., *Ohhh* [silence], *right…*).

4.2 Describing shapes of co-verbal expressions

The concept of ‘gesture’ as proposed in [36] describes the shape and propagation of movement in very high detail. In the context of shape, the used annotation schema shown in Figure 2 is pose-oriented, and captures the spatial configurations of key-poses in the form of 3D configurations of the articulators manifesting the shape. This is also very important for the representation of the gestures on humanoid robots. In particular, and as in the case of skeleton-based articulated conversational agents, the humanoid robots can also be driven by a kinematic model [38] (Figure 1). Therefore, similar movement controllers (e.g., joints) can be used to drive the desired movement. The general shape of the movement refers mostly to the shape of the hand, manifested and observed within the communicative expression. As a guideline, and as a base, we defined prototypical (referent) hand-shapes based on the Ham-no-sys [39] and SignPhon [40] databases. The general position of the arm is described in the form of abstract spatial dimensions: the radial-orientation of the elbow (e.g., far-out, side, front, inward, etc.), the height of the palm position (e.g., head, shoulders, chest, abdomen, belt), the distance of the palm from the torso (e.g., fully-extended, normal, close, touching, etc.), and the elbow inclination (e.g., orthogonal, normal, etc.). The annotation also allows for spatial configurations of movement controllers to be stored under each abstract dimension. The shape prototypes, therefore, describe the kinematic model that will manifest them in either the physical or the virtual domain. Since we wish for the physical, virtual and real-life movement to represent the same meaning, the gesture repositories use the notion of similar shapes and identification based on the Hamming-distance [41].

![Figure 2. Form-oriented annotation of communicative sequences as constructed in ANVIL [37] (Slovenian multi-modal database).](image-url)
In the following section, we will present the proposed behaviour generation algorithm that co-aligns verbal and non-verbal behaviour. The algorithm also performs re-synchronization that arises as a result of behaviour adaption due to restrictions provided by physical agents (humanoid robots).

5. The TTS-driven Non-verbal Behaviour Generation Algorithm

The system’s architecture is shown in Figure 1. All the modules in the system are easy to maintain and improve, and they allow the easy integration of new algorithms and processing steps by using the flexible queuing mechanism that is the core of the engine. The TTS engine’s queuing mechanism is further exploited in order to fuse verbal and non-verbal modules in a flexible way. Moreover, temporal synchronization and acoustic information, extracted or predicted from the general input text. They are also used for the flexible and transparent construction of the complex features needed by several machine-learned models (CART trees) used in several of the engine’s modules. FSMs are used for the symmetric and space-efficient representation of external language-dependent resources, linguistic rules, etc., as already suggested in [31]. The generation of non-verbal behaviour co-aligned with verbal information is performed by the proposed algorithm as follows in the following subsections.

5.1 TTS-driven non-verbal behaviour generation stream

In order to recreate/perform different co-verbal gestures within either virtual or physical interfaces, three tasks must be taken into account. The first task is to select and prepare the visual content. Based on the understanding of the verbal information, this task must identify the intent, the meaningful words (or phrases), and also the suitable gesture prototypes representing the meaningful words. The second task then addresses the content presentation. Based on the identified words and additional prosodic features, this task must identify exactly where the meaningful part will be utilized and what visual cues (if any) will surround it. Finally, and for the third task, by using elimination techniques (e.g., communicative functions, redundant gestures, etc.), the produced density of the visual cues must be reduced and only those visual cues that actually contribute to the meaning of the communication are to be kept. In the proposed system, the first task – also called ‘semantic synchronization’ - is performed in the phase-tagger deque. The second task (pragmatic synchronization) and the third task (temporal synchronization) are performed in the gesture-planner deque (Figure 1).

5.2 Data structures

In the system, a common HRG data structure is used that is accessible by all the modules. In this way, they are all able to access, change, store or enrich the information while performing several algorithm steps on general input text. The HRG data structure is not static and can dynamically change its structure through the queuing mechanism (e.g., non-verbal modules are able to create new relation structures while processing the input sentence) [55]. Relation structures can be in the form of linear lists or else in the form of trees, depending on the type of information that is stored in the HRG structure by the specific module. The HRG structure is ultimately in the expression-processing deque for the EVA-Script-based behaviour specification. Figure 3 presents the complete structure of the HRG (when all the processing steps of the proposed algorithm are already performed). The relation structures in the form of linear lists are Segment, Syllable, Word, Phrase, IntEvent, SynUnits, ContentUnits and MovementPhases. In addition, those in the form of trees are SyllableStructure, PhraseStructure, IntonationStructure, SynUnitsStructure, CESTructure, CoverbalExpressionStructure and MPDynamicsStructure. For example, SynUnitsStructure relates items in the SynUnits relation layer and items in the Segment relation layer. Next, CESTructure relates items in the MovementPhases relation layer and items in the ContentUnits relation layer. SyllableStructure relates items in the Segment relation layer and items in the Syllable relation layer. As can be seen, it also relates items in the Syllable relation layer and items in the Word relation layer. IntonationStructure relates items in the IntEvent relation layer and items in the Syllable relation layer. MPDynamicsStructure relates items in the MovementPhases relation layer and items in the Syllable relation layer. CoverbalExpressionStructure relates items in the ContentUnits relation layer and items in the Word relation layer. Finally, PhraseStructure relates items in the Phrase relation layer and items in the Word relation layer.

Figure 3. A common HRG structure as a storage for verbal and non-verbal data in the algorithm.
TTS-driven non-verbal behaviour generation stream consists of phase-tagger deque, gesture-planner deque, and expression-processing deque (Figure 1). Each deque performs several processing steps, as specified by the proposed algorithm, and described in detail in the following subsections.

5.3 Phase-tagger deque

The decision as to whether a specific text sequence represents a possible gesture trigger is performed based on semiotic grammar, as deduced by the annotation of the conversational corpora [36]. The semiotic rules (needed in order to identify gesture triggers) are based on relations established by the linguistic properties of semiotic gestures, as proposed by Pierce [21]. Some of these properties are language-dependent. However, language dependencies are resolved through relations/rules maintained within the grammar. The algorithm’s processing steps performed in the phase-tagger deque are presented in Figure 4. As can be seen, the deque has to perform semiotic tagging, semiotic processing and matching.

The input represents already-POS-tagged word sequences (stored in the HRG structure) that are matched against the semiotic grammar’s rules and relations as defined in the gesture dictionary. Firstly, the semiotic tagging step tags those words that could trigger co-verbal movements (mandatory word-type-units). At the morphology level (based on the semiotic grammar’s rules), the procedure searches for nouns, verbs, pronouns, fillers and determiners. The mandatory word-type-units and their neighbours are then grouped into word-phrases that are compared against the classification of semiotic dimensions. The classification into a semiotic dimension defines the intent of the phrase, the form of visual representation, and it also suggests the structure of the co-verbal gestures. As shown in Figure 4, the semiotic processing step traverses through the HRG’s Word relation layer and, by considering the preceding POS tags and successive POS tags, tries to detect those patterns that match one of the word-type-orders defined by semiotic classification. Finally, during the matching step, the lexical affiliates are chosen for those words identified as carriers of meaning. This is done by performing a semantic lookup in the gesture lexicon and selecting an appropriate physical representation of the word(s). In the external gesture dictionary, a suitable physical shape is selected based on the following rules: a) the shape that has been observed to precede the meaningful manifestation, and b) the shape that has been observed to precede the meaningful manifestation co-aligned with similar semiotic patterns. The properties of the selected co-verbal movement are then stored as additional attributes to the items in the Word relation layer, as seen in Figure 5. Additionally, the matching step creates a new ContentUnits relation layer in order to store generated content units (CUs). The phase-tagger deque may identify multiple semiotic patterns and multiple possibilities for representation. This means that each CU may be described by multiple word items and that each speech phrase may be defined by multiple CUs (e.g., they can be identified by multiple intents). The final selection is performed in gesture-planner deque, when all the prosodic features are already available.

As seen in Figure 5, the information stored within the CUs is also in the form of attribute-name/value specifications, where each attribute has a predefined set of values that subsequent deques can select from. In particular, the set of values for the attribute semi is based on the semiotic types the system can classify (e.g., iconic, adaptor, symbolic, etc.). The attribute mwords holds those key-words that may, in the observed sequence, represent meaning. For example, a word SL: iste (EN: the same). In addition, the attribute pwords indicates those words that the preparation movement phases may be triggered by. For example, a word SL: vedno (EN: always). Word items are enriched at this level with new attributes, describing the symbolical co-verbal alignment and movement-structure (propagation) of shapes of non-verbal expressions with regard to the verbal content. These attributes are: movement phase, rhshape (right-hand-shape), rashape (right-arm-shape), lhshape (left-hand-shape) and lashape (left-arm-shape). These attributes have their values defined based on the selected gesture prototypes available in the external gesture dictionary.
In the Word relation layer, the processed word sequence SL: vedno iste in samo iste obraze (EN: always the same and only the same faces) is tagged as: R, P, C, Q, P and N (adverb, pronoun, conjunction, particle, and noun). The semiotic tagging step recognizes three mandatory word-types: two pronouns and a noun. Next, the semiotic processing step indicates two semiotic patterns: adverb-pronoun and particle-pronoun-noun that may trigger a co-verbal expression. Therefore, two iconic CUs are inserted into the HRG structure. Within both semiotic patterns, the pronoun carries the meaning that might be reflected by the physical manifestation. The pronouns also indicate the position of the stroke and the particle in the second pattern indicate the preparation movement-phase. Moreover, the noun in the second pattern indicates the hold movement-phase. By considering semiotic relations, implicit rules and lexical affiliation, the matching step gives us the co-verbal movement for each semiotic pattern, which is identified as 1-handed, the lexical affiliates manifested over the word-carrying meaning as K7 (for the hand shape), and Fr\Ab\Cl\O (front, abdomen, close, orthogonal) for the position of the right arm. The phase-tagger deque in this way identifies what non-verbal content can be represented on a specific input sentence and the starting and ending points in the Word relation layer that drive the manifestation of the physical shapes. Within the proposed algorithm, it takes care in identifying those words that carry meaning (or a functional role) and the physical manifestation that the meaning may be represented by.

5.4 Gesture-planner deque

As seen in Figure 1, the phase-tagger deque is followed by the grapheme-to-phoneme conversion deque, which assigns phonetic transcriptions and inserts syllable markers by considering the cross-word contexts and by using phonetic lexica and CART trees [30]. Next, the symbolic prosody deque performs the following three algorithms on the syllable level: the prediction of phrase breaks (position and level), the prediction of prominence labels, and the prediction of Tilt intonation labels, by using CART trees [30] [34]. The phrase break prediction step inserts phrase break labels, the prominence prediction step marks the prominent syllables, and the intonation prediction step assigns symbolic Tilt intonation labels to each syllable, as proposed in [33]. The next acoustic prosody deque then also predicts the duration of segments (phonemes), the duration of pauses at phrase break positions, and the Tilt acoustic parameters for the intonation events assigned by the symbolic prosody module. All these deques generate the prosodic features and temporal information required by the gesture-planner deque. By using the information already stored in the HRG structure for a given input sentence, the gesture-planner deque performs the elimination of multiple possible intents and multiple representations of the intent. The elimination process uses features such as prosodic word phrases, accents and intonation. The gesture-planner deque also performs the temporal alignment of triggers and conversational expressions in the form of movement-phases [22]. Furthermore, it also temporarily co-aligns the consecutive series of conversational expressions into a conversational act, as follows. Firstly, the deque performs the alignment of the propagation of the selected co-verbal gestures and the input-sentence on the syllable level by using information in the following HRG layers: ContentUnits, Phrase, Word and Syllable relations. It also creates a new HRG layer, called ‘MovementPhases’ for describing the relation between the propagation of co-verbal movement and the input sentence on the syllable level, as can be seen in Figure 6. The starting and ending points of the CUs are compared against assigned prosodic word phrases, as predicted by the symbolic prosody deque. In the system, the B3 label is used for labelling major phrase breaks, and the B2 label for minor phrase breaks. The prosodic word phrases indicate those text-sequences that have a complete meaning. In addition, sentences with more than one prosodic word phrase can indicate additional explanation, emphasis, or even the negation of the preceding meaning [30]. Presented in Figure 7 are the processing steps that have to be performed within this deque, according to the proposed algorithm. The first step, called Search for word-phrase breaks, adjusts (extends or even removes) the starting and ending points of the selected CUs. Specifically, it aligns CUs with the predicted prosodic counterparts. In this step, the rule is applied such that each prosodic word phrase can contain only one (or none at all) co-verbal gestures. Further, the movement-propagation of each co-verbal expression must be maintained within the predicted prosodic word phrase. As seen in Figure 6, the sentence SL: vedno iste in samo iste obraze (EN: always the same, and only the same faces) at this point of the algorithm contains two CUs: CU-1 and CU-2. Their starting and ending points are determined by the phase-tagger deque as: SL: vedno iste (CU-1) (EN: always the same), and SL: samo ... obraze (CU-2) (EN: and only the same faces). However, and obviously, the predicted prosodic word phrases demand necessary adaptation at this level. As can be seen, the predicted word phrases are SL: vedno iste (WP-1), and SL: in ... obraze (WP-2). Clearly, the CU-1 matches the WP-1, whereas the CU-2 has to be extended in order to also include the word SL: in (EN: and). However, if the CUs include more words than the predicted WPs, those words are automatically disregarded.
The second step, called Search for emphasized words, identifies the emphasized word, as predicted by the symbolic prosody deque. This is a word that contains a syllable assigned with a PA label (primary accent - the most accentuated syllable within the specific prosodic phrase). The stroke-movement phase is then co-aligned with the word containing the PA syllable. When the CU has no PA syllable (i.e., no emphasized word), the CU is simply disregarded. Those words preceding the emphasized word then represent preparation or pre-stroke-hold movement-phases. Moreover, those words following the emphasized word represent the post-stroke-hold or retraction movement-phases. On all the remaining stressed syllables within the intonation prosodic phrase, the secondary accent (NA) is assigned. The NA labels identify the preparation movement-phase (when the word contains a NA-tagged syllable and precedes the prosodic gesture trigger) or the retraction movement-phase (if this is the last co-verbal expression or if a longer pause is predicted). In addition, when silences (sil) are predicted by the acoustic prosody deque, they indicate the existence of the hold-movement-phases. The third step, called Search for stressed syllables, aligns the stroke-movement phases according to the PA syllable. In this step, the starting and ending points of the stroke phase are determined by the beginning of the emphasized word and the end of the PA syllable. In Figure 6, there are two emphasized words containing a PA syllable. The first syllable ‘is’ in the emphasized word SL: *iste* (EN: the same) identifies the position of the first meaningful shape defined by CU-1 (e.g., the physical representation of the word SL: *iste*). Therefore, the first stroke phase will be propagated over the PA syllable. The second syllable ‘mo’ in the emphasized word SL: *samo* (EN: only) identifies the position of the manifestation of the second meaningful shape defined by CU-2. The fourth step, called Align stroke, aligns the preparation movement-phases. In particular, the CUs already contain the definition of the shape preceding the meaningful shape (e.g., the “initial” physical manifestation from which the body transforms throughout the stroke phase). The fifth step, called Align preparation, aligns the preparation movement-phase. The preparation phase is identified by the first word with the NA syllable that precedes the specific prosodic gesture trigger. The start and end points of the preparation phase are further identified by the position of the NA syllable and the end of the “preparation” word. As seen in Figure 6, the words SL: *vedno* (EN: always) and SL: *in* (EN: and) are indicated as preparation words for the first and second prosodic word phrases, respectively. Both words also influence the physical manifestation of shapes within the preparation movement-phases of CU-1 and CU-2. For CU-1, the NA syllable is ‘ve’ and the start and end points of the movement phase are determined by the syllables ‘ve’ and ‘dno’. For CU-2, there is no word preceding the prosodic gesture trigger that would contain the NA syllable. Moreover, and according to the aforementioned rule, this indicates that no preparation movement phase is required. Nevertheless, since CU-1 and CU-2 represent similar meanings (in terms of gesture affiliates), the preparation phase over the word SL: *vedno* (EN: and) is artificially inserted into the movement propagation scheme. Finally, the last step, called Align hold/retraction, aligns the hold (both pre- and post-stroke) and retraction movement phases. The retraction movement phase is identified by the last meaningful word-phrase via the word that contains the NA syllable and which precedes the B3 phrase break or else precedes a longer pause. Additionally, the start and end points of the movement-phase are determined by the NA syllable and the end of the “retraction” word. In Figure 6, the starting and ending points are the NA syllables ‘ra’ and ‘ze’ of the word SL: *obraze* (EN: faces). Moreover, the hold movement phase is determined by the pauses (the predicted sil units, stored in the Syllable relation layer) and the residual (those content syllables not yet used in the preparation, stroke or retraction movement phases) of the meaningful words not yet aligned with the stroke movement-phase. Further, in addition to the sil units, the residual of the word phrase WP-1 is the unused syllable ‘te’, while the residuals of the word phrase WP-2 are the syllables ‘is’, ‘te’ and ‘ob’. As can be seen in the MovementPhases relation layer of the HRG structure in Figure 6, the residual content and the sil units are assigned as hold movement-phases. The movement structure as generated by all six steps is presented in detail in Figure 8. The movement phases such as preparation (P), stroke (S) and retraction (R) indicate a physical manifestation of a shape, and the hold (H)
movement phase indicates the maintaining of the closest previous physical manifestation.

Figure 8. The generated movement structure with the sequence of shapes for conversational expressions.

The conversational expression CE-1 is aligned with the prosodic boundaries (phrase breaks) of the word phrase SL: *vedno iste* (EN: always the same) marked as WP-1. As seen in Figure 8, during propagating, two shapes will manifest. The shape CU1-P will manifest itself over the syllables ‘ve’ and ‘dno’. Next, the preparation movement phase is followed by a hold (H), indicated by the sil unit. The next stroke shape CU1-S will manifest itself over the syllable ‘is’. The residual syllable ‘te’ then indicates the post-stroke hold phase (H). Obviously, an additional post-stroke hold phase (H) is inserted due to the presence of the predicted sil unit after the word SL: *iste* (EN: the same). In a similar way, we describe the conversational expression CE-2 aligned with the prosodic boundaries of the word phrase WP-2. The MovementPhases relation layer in this way stores the generated movement structure of the observed sequence (Figure 6).

Accordingly, the layer outlines what shape should manifest over specific words, where the movement should be withheld, and where it should be retracted to its neutral (rest) state. However, these steps do not yet specify any temporal boundaries for movement generation. Ultimately, there remain several repeated holds within the movement sequence which should either be filtered or merged. Therefore, in the proposed algorithm the temporal alignment of verbal and co-verbal gestures is also integrated as part of the gesture-planner deque. As presented in Figure 9, these additional processing steps benefit from the temporal properties at the phoneme/viseme level, as predicted by the acoustic prosody deque, and stored within items in the Segments relation layer in the HRG structure. Additionally, the Segments relation layer also stores the temporal information about the predicted pauses (sil units). The temporal information in the Segments relation layer is already used by the TTS modules (verbal stream). However, within the proposed algorithm the duration of each movement phase can also be determined by predicted durations of phonemes/visemes and silences. The temporal processing steps benefit from the following HRG layers: ContentUnits, Segments and MovementPhases. Additionally, new units called ‘phase units’ (PUs) are stored within the MovementPhases relation layer of the HRG structure, while the CUs (stored in the ContentUnits relation layer) are enriched with additional attributes.

Figure 9. The processing steps for the temporal alignment in the gesture-planner deque.

Each sil unit represents a hold movement-phase in the movement structure (MovementPhases relation layer). The first step in Figure 9 is called *Filter CEs* and searches for those sil units that have a predicted duration of 0 ms, since they are unnecessary and should be removed. In order to filter any corresponding hold-movement phases, the vertical HRG’s connections linking each sil unit (in the Segments relation layer) and the corresponding hold movement-phase are deleted, and the corresponding PUs removed from the MovementPhases relation layer. For example, in Figure 10 the sil unit positioned between the syllables ‘dno’ and ‘is’ has a predicted duration of 0 ms. Therefore, the corresponding connected hold movement PU has to be removed from the MovementPhase relation, and new links between the neighbouring preparation and stroke movement-phases are established (dotted links). This step also merges repeated hold-movement phases into a single hold. When this is done, the start and end points of the merged hold are adjusted to the starting point of the first hold and the ending point of the last hold. The second step, called *Align CEs*, temporally aligns each conversational expression (CE) with the input text. In Figure 10, CE-1 is represented by the co-verbal gesture CU-1. Its attributes tell us that the assigned co-verbal gesture is iconic and that the meaningful shape represents the word SL: *iste* (EN: the same) (lexical affiliate). Considering the context, the word units contained within

Figure 10. Temporal alignment in the HRG structure.
CE-1 are the words SL: vedno (EN: always) and SL: iste (EN: the same). Next, the starting and ending point of CE-1 are determined by the syllables ‘ve’ and ‘te’. Considering the vertical link between the syllable and movement-phases relation, we are also able to assign to CE-1 the sequence of the following movement phases: Preparation-Stroke-Hold. In addition, the physical manifestation within the stroke-movement phase represents the word SL: iste (EN: the same). In particular, the physical manifestation within the preparation movement-phase represents the initial shape of the physical representation of the word SL: iste (EN: the same) triggered over those words marked as in the pwords attribute.

The CUs are temporally defined by the following attributes: delay, duration, persistence and duration-down. Moreover, the PUs are temporally defined by the attributes duration-up and persistence. The temporal values of the CUs and PUs are calculated by using the temporal values of their “children” in the HRG structure (phonemes and sil units), as predicted by the acoustic prosody deque. In order to calculate all the attributes’ values, equations (1)-(6) are suggested to be used in this step. The delay attribute represents the absolute duration, measured from the beginning of the input text sequence, for which the co-verbal gesture is withheld prior to execution. When the first conversational expression (CE-1) does not start with the beginning of the input sentence, a default CE-0 is inserted by assigning that duration which is equal to the sum of the segment units preceding the first segment of CE-1. In addition, when CE-1 starts with the first segment in the input sentence, the corresponding delay is always set to 0. For this attribute, the following equation is used:

\[ CE_{i}\text{delay} = \left(\sum_{j=1}^{n} CE_{j}\text{delay} + CE_{j}\text{duration}\right) \]  

where \( i \) represents the observed content item. The duration of the last hold-movement phase of each CE represents the persistence attribute’s value for the CE. The value is calculated as follows:

\[ CE_{i}\text{persistence} = r(MP_{j}) \]  

when the last movement-phase is hold, otherwise its value is zero. The persistence attribute represents the duration for which the shape of the preceding movement-phase is maintained prior to the execution of the next CE. Further, the duration attribute represents the overall duration of the conversational expression. However, it does not contain the duration of the final hold-movement phase. It is calculated as follows:

\[ CE_{i}\text{duration} = \left(\sum_{j=1}^{n} r(MP_{j})\right) - CE_{i}\text{persistence} \]  

where \( n \) represents the number of movement phases. The next duration\_down attribute represents the duration of the retraction phase, and is calculated as follows:

\[ CE_{i}\text{duration\_down} = r(MP_{j}) = \left(\sum_{j=1}^{n} r(\text{sil})\right) \]  

where \( n \) represents the number of segments and \( k \) the first segment in the given CE. The PU stored within the MovementPhases relation layer, in addition to shape manifestation, also stores the temporal values of the preparation/stroke movement-phase and the temporal values of those holds that immediately follow the observed preparation/stroke movement-phase. The corresponding attribute duration\_up represents the duration of the preparation/stroke movement-phase - this is how long the transformation between shapes has to last. The attribute’s value is calculated as a sum of predicted temporal values on segment units (phonemes/visemes) contained within the specific preparation/stroke movement-phase:

\[ PU_{i}\text{duration\_up} = r(MP_{j}) = \left(\sum_{j=1}^{n} r(\text{sil})\right) \]  

where \( n \) represents the number of segments and \( k \) the first segment in the given PU. Likewise, the attribute persistence is determined by the first hold phase that follows the observed preparation/stroke movement-phase. Its value is calculated as a sum of temporal values for segments contained within the hold phase:

\[ PU_{i}\text{persistence} = r(MP_{j}) = \left(\sum_{j=1}^{n} r(\text{sil})\right) \]  

where \( n \) represents the number of segments and \( k \) the first segment in the given PU. In Figure 11, the generation of conversational expressions is presented in more detail, including the temporal alignment of movement on the level of phonemes/visemes. As seen in Figure 11, the required links can be efficiently and flexibly established between the Segment relation layer, and the MovementPhases relation layer.

In this way, we have available relations between conversational expressions, words, movement-phases, and...
phase units. The PU fuses the preparation/stroke movement-phases and the hold/retraction movement phases that follow the preparation/stroke. As mentioned before, the exceptions are the hold movement-phases, which are the last in the observed sequence. In Figure 11, the conversational expression CE-1 consists of the phase units PU-1 and PU-2, while the execution of CE-1 is finished by a hold (H) movement-phase. The CE-1 can then be temporally described by equations (1)-(3). Since the CE-1 starts with the first segment in the input text sequence, the delay value is set to zero. Next, the persistence value is determined by the duration of its last hold movement-phase (in this case, the persistence value is 0.068 s). Moreover, the duration value is determined by summing the durations of PU-1 and PU-2. The PUs’ durations are determined by the predicted duration of those segments it encapsulates. In the case of CE-1, the duration_up value for PU-1 is determined by summing the durations of t(‘v’), t(‘e’), t(‘d’), t(‘r’) and t(‘o’). In addition, in the case of PU-2, this is determined by summing the durations of t(‘t’1) and t(‘s’). Since none of the PUs encapsulates the hold movement-phase in this case, the corresponding persistence values are, in both cases, set to zero.

By using equations (1)-(6), any conversational expression and corresponding shape manifestation can be temporally aligned with the speech signal, as synthesized on the general input text. Both processing steps finally generate two types of units that contain all the information necessary for the recreation of the co-verbal gestures. Each CU stores a global temporal structure for corresponding conversational expressions, whereas each PU stores local information regarding the overlaid shapes. In the following section, we discuss the final non-verbal deque that has to transform CUs and corresponding PUs into suitable procedural descriptions of synthetic gestures, as required for the recreation of co-verbal gestures by using synthetic embodied conversational agents and robots.

5.3 Expression-processing deque

ECA-based animation engines generally recreate non-verbal behaviour based on procedural animation description mark-up languages, such as BML [12] and EVA-SCRIPT [18]. These mark-up languages all require at least the temporal specification (e.g., relative position, duration) of behaviour and the description of shape provided in at least abstract notation. These behaviour descriptions are then able to be fed into the animation engine, and can be recreated by a synthetic ECA. In the proposed system, the expression-processing deque is needed in order to transform the HRG data structure into a form understandable to the ECA-EVA animation engine (or into the engine driving the movement of a robotic unit). Therefore, within this deque the HRG structure is transformed into EVA-Script format for the behaviour descriptions, supporting both lip-sync and co-verbal gesture animation processes. Since the HRG structure contains very detailed information about co-verbal gestures, transformation into other mark-up languages is possible and straightforward. In Figure 12, the processing step of this deque is presented. As input, the CUs in the Content relation layer and the PUs in the MovementPhases relation layer are used. The output is a behavioural script, written in EVA-SCRIPT. In order to re-create (animate) the co-verbal gestures as described in the HRG structure, those shape models determined by the PUs’ attributes (rshape, rashape, lshape and lshape) are selected from the gesture dictionary. The selected prototypes are temporally adjusted according to the temporal specification, stored in the CUs and PUs. The processing step in the deque simply traverses through the CUs, recalls the aligned PUs, and generates the corresponding XML description (EVA-Script). As can be seen (in the EVA-Script), each CU is represented by a gesture tag. The obligatory attributes for this tag are thus the following attributes: name, type and delay. In the context of recreation, name value is used only for easier access to the selected procedure in the animation graph. The type value is generically set to ‘arm_animation’, since hand and arm gestures are to be animated, as specified by the PUs. The start value defines for how long the animation of the described behaviour has to be withheld prior to its execution. It is equivalent to the delay value stored in the CU. The durationDown value, stored in the CUs as duration-down, describes the duration of the controller’s retraction (e.g., its transition from an excited to a neutral (or previous) state).

Each PU in the HRG structure then represents a sequence of body-parts moving in a parallel manner. The configuration of body-parts is maintained within the UNIT tags. The value of the attribute name represents a gesture prototype stored within the gesture dictionary. The temporal specification for each UNIT is provided by specifying the following two obligatory EVA-Script attributes. The durationUp value, stored in the PU as duration_up, represents the absolute temporal value and describes how long it takes for the specified shape to be fully overlaid on the synthetic agent (e.g., the duration of the transition between shapes). Meanwhile, the value, stored in the PU as the persistence value determines persistence for the period after the transition between sequential shapes is completed and in which the overlaid shape (the excited state) has to be maintained. Additionally, this deque also forms a description of facial expressions required for the-lip-sync process, where the visual sequence of visemes/phonemes is described in the form of viseme in EVA-Script’s tags, encapsulated by using the speech tag. The corresponding information is stored in the Segment relation layer. In this layer, each segment (including sll) represents one viseme tag within a lip-sync specification, and at the same time a gestural
affiliate for a facial expression which, in the mouth region, overlays the pronunciation of the segment. Additionally, the duration of each segment is also available and is equivalent to the duration attribute of the EVA-Script’s *viseme* tag. The appropriate transition between segments is handled internally by the EVA animation engine, as proposed in [18].

![Figure 12. Transformation of the HRG structure into an EVA-script behaviour description.](image)

The expression-processing deque also integrates an optional capabilities-conversion processing step. It ensures that the behaviour specification also becomes compatible with the target robotic platform (in our case, iCub [19]). As already mentioned, there are several similarities between ECA-EVA-based movement-control mechanisms (virtual specification) and iCub movement specifications (physical specification). The similarities range from the movement controllers used to the techniques used to simulate (recreate) the specified movement. In particular, both the concept of ECAs (based on the EVA-Framework [18]) and the concept of robots specify movement controllers as the joints (bones) controlled by the HPR (H-heading, P-pitch and R-roll) mechanism. Although the physical systems mostly use the RPY (P-pitch, R-roll and Y-aw) mechanism, the analogy in terms of 3D space exists. In particular, e.g., the “yaw” is analogous to the “heading” attribute. These joints are a part of skeleton structure approximating the human skeleton structure (a concept of forward-kinematics). A minor difference between the physical and virtual model may be observed regarding the HPR configuration. The virtual model specifies the HPR configuration as a single controller, whereas the physical model configures the HPR values by using three movement controllers (one for each dimension in the HPR space). For instance, the left shoulder joint of the virtual model is represented by the following joints: 0 (shoulder pitch – P), 1 (shoulder roll – R), and 3 (shoulder yaw – Y/H). The dynamics of movement are then defined by the time it takes to perform the transition. The same principles are also shared by the EVA-Script. The behaviour is specified in a context of forward-keying. Major parts of movement are defined as sequential movement phases (sequences). Each such sequence also describes the HPR configuration of the particular joints that are responsible for the formation of shape. In addition to shape and movement articulators, the major disadvantages of robot movement are the dynamics and fluidity of the movement, as well as the gradualness of its trajectory. The dynamics are addressed mostly by the relation between the temporal distribution of the co-verbal expression (movement-phases) and the spatial configuration of the shapes. Virtual movement can be specified as almost instantaneous, whereas the movement controllers of the robotic unit require longer temporal periods. The responsiveness of the servo-motors (the robot’s physical movement controllers) is defined by the motors’ maximum angular velocities and by their maximum angular acceleration. These values are generally specified by the manufacturers of the servo-motors [48]. Since the stroke-phase is the most important phase (the co-verbal shapes carry meaning in the stroke phase), it can, if required, gain additional temporal space. The temporal-adjustment-process shortens (or even eliminates) those movement-phases that surround the stroke-phase. The process takes into account the responsiveness of the robot’s physical movement controllers. The surrounding movement-phases are shortened to the minimal requirements for the stroke-phase to adjust to the movement controllers specifications. For instance, the stroke-phase (Figure 13) is extended over the preparation-phase in order to include the whole duration of the trigger word. In order to temporally adjust the movement within the stroke-phase, the minimal time required to perform the physical angular adjustment is calculated and compared against the duration of the stroke. The stroke-phase is then (if necessary) extended into the preparation and/or hold/retraction movement phases. The temporal remainder of those optional phases is then compared again against the responsiveness of the robot. Those phases that cannot be performed in the suggested time are automatically disregarded.

![Figure 13. Recreating communicative acts by a robot and an ECA-EVA.](image)
6. Experiment

In this experiment, the text sentence SL: *Kača je bila tako velika* (EN: the snake was that big) was used as input to the proposed system in order to test the suggested algorithm. By analysing the system’s output, the possible triggers are identified to be word SL: *kača* (EN: snake/noun), and the word phrase SL: *tako velika* (EN: this big/adverb-adjective). However, the word SL: *tako* (EN: this/adverb) is at the same time the only word in the text being emphasized. The word SL: *kača* was, therefore, eliminated as a possible trigger. Further, the temporal alignment was performed by using the concept of stressed syllable and the output from the *acoustic prosody deque*. Considering that the stressed syllable identifies the beginning and the end of the movement phase. The phoneme sequences used to pronounce the indicated verbal sequences and their pronunciation rate (identified by the *acoustic prosody deque*) then determined the absolute duration of the movement-sequence. The duration of the stroke-phase was determined as the absolute duration measured for the pronunciation of the sequence starting at first stressed phoneme and ending at the last phoneme of the trigger word (including the short pause after the triggered word in this case). In this experiment the system predicted the needed duration 109 ms for the stressed syllable ‘ko’. Considering that the stressed syllable also represents the end of the trigger word’s phoneme sequence, the duration of the stroke movement-phase was also specified as 109 ms. The hold-phase was identified as part of the trigger-phrase that follows the stroke phase. Then follows the trigger word-phrase SL: *tako velika* (EN: that big). In this case the hold sequence that follows the stroke trigger was identified as phoneme sequence used to pronounce the word SL: *velika* (EN: big) (and its duration was predicted to be 451 ms). The preparation movement-phase was then identified as a phoneme sequence (including short pauses) that precedes the phonemes already co-aligned with the stroke-movement phase. In this case the unused-phoneme sequence of the trigger word-phrase was SL: *ta* (EN: this). And its duration was predicted to be 191 ms. The algorithm also identified the word sequence SL: *je bila* (EN: was/verb) as part of the preparation movement-phase. The duration of the phoneme sequence indicating the preparation movement-phase was determined to last 784 ms (593ms + 191 ms). In this way the proposed system temporally co-aligned the non-verbal expression with the pronunciation of the word sequence SL: *je bila tako velika* (EN: it was that big). However, no expressions are identified to co-insight with the word SL: *kača* (EN: snake). Since concept of absolute durations is used, the non-verbal behaviour (movement) will start to appear after 488 ms (the time used to pronounce the word *kača*).

In Figure 13 behaviour as animated by ECA-EVA, and simulated in the iCub based on generated system’s output, is shown. The upper sequence of images represents the key poses captured while simulating the behaviour when using iCub Simulator. The lower sequence of images represents the key poses of the same behaviour as generated by ECA-EVA. And in the middle the temporal, semantic, and pragmatic synchronization of the verbal and non-verbal behaviour is presented. In order to simulate the same behaviour on both the humanoid robot and virtual agent, also the temporal dynamics of the virtual behaviour specification must be adjusted, and aligned with the capabilities of the specific robot. For instance, the stroke phase has to be extended, at the expense of the preparation phase and hold phase. The preparation phase also has to be extended to begin with the beginning of the verbal sequence. In order to prevent instant configuration changes an additional hold phase can be inserted between the preparation and the stroke phase. In the context of physical manifestations, the word-phrase SL: *tako velika* (EN: that big) is in the gesture lexicon represented by the corresponding key-shape (last shape in the shape-sequence in Figure 13). The shape representing the word-phrase SL: *je bila* (EN: it was) was selected as most suitable among representations stored within the gestural dictionary.

7. Evaluation

In order to evaluate the proposed system, we conducted a perceptive experiment by using an embodied conversational agent EVA and an emulation of a robotic unit. We wanted to gain an insight into the quality (viability) of the gestures produced by the proposed algorithm. Additionally, we wanted to gain a deeper understanding into how the proposed gesture production algorithm impacts upon the perception of the multi-modal content. The perceptive experiment comprised a series of isolated (unrelated) statements that were synthesized into co-verbal behaviour, executed first by ECA-EVA and followed by simulation performed in iCub.

Thirty participants were involved in the experiment. Five participants were members of our staff and twenty-five participants were students. Among them all there were seven females and twenty-three males, ranging in age from 22 to 40 years (Mean = 26.73, Standard Deviation = 4.88). All the participants were native Slovenian speakers with no (or little) insight knowledge about the design, functionality or limitations of the proposed system. The participants were working and studying in several areas, including computer science, telecommunications, electronics and language technologies.

In the experiment, the participants were instructed to observe and evaluate the information communicated by the target synthetic entity (ECA-EVA and the simulated robotic unit). The communication setup was unidirectional...
and no face-to-face interaction was intended. The complete system synthesized speech output in the Slovenian language by using the TTS PLATTOS engine. Based on answers to the post-experiment questionnaires, the quality of presentation and the naturalness of the perceived visualized statements were investigated.

| Dependent measure | Questions | Scale |
|-------------------|-----------|-------|
| Content match (C1) | Did the gesture represent the correct speech segment? | 1 – no match, 5 – matched as expected |
| Synchronization of form (C2) | Did the visualized form represent the co-aligned speech segment? | 1 – no correlation, 5 – high correlation |
| Synchronization of propagation (C3) | Was the execution of the movement phases fluid and appropriate? | 1 – not appropriate, 5 – highly appropriate |
| Synchronization between consecutive gestures (C4) | Was the start position of the consecutive gesture aligned with the end position of its predecessor? | 1 – no alignment, 5 – highly aligned |
| Temporal synchronization (C5) | Was the visual execution of the gesture temporally aligned with the speech? | 1 – no alignment, 5 – highly aligned |
| The speed of execution (C6) | How was the movement executed? | 1 – too slow, 5 – too fast |
| Amount of synthesized gestures (C7) | How was the density of the executed movement? | 1 – too low, 5 – too high |

Table 1. Measures used to evaluate the quality of the gestures generated by the proposed algorithm.

The measures defined and used in order to evaluate the proposed system are presented in Table 1. As can be seen, Table 1 consists of several dependent measures, questionnaire items and scales used to evaluate the quality and accuracy of the proposed algorithm, respectively, for ECA-EVA and the iCube simulation. The content match measure (C1) was defined in order to evaluate whether the visualized gesture represented a viable speech segment (e.g., the visualized segment provided some sort of complementary meaning). The synchronization of the form measure (C2) was used to evaluate whether the selected speech segment was visualized correctly (e.g., the selected speech segment can be visualized by using the selected movement sequence/sequences). Further, the synchronization of propagation measure (C3) was used to evaluate the inner-fluidity of the represented movement. For example, the visualized movement was fluid and the executed movement phases (especially stroke, hold and retraction) were performed correctly over the visualized speech segments. The synchronization between consecutive gestures measure (C4) was used to evaluate the outer-fluidity of the movement (e.g., did consecutive gestures retract and transit in a fluid and viable manner?) The temporal synchronization measure (C5) was defined in order to evaluate the temporal alignment between generated speech and gestures. The speed of execution measure (C6) was defined in order to evaluate the rate of movement propagation. Finally, the amount of synthesized gestures measure (C7) was used to evaluate whether the density of the movement performed was too low or too high.

The participants listened to speech sequences prior to observing the visualized sequences. Fifteen participants were randomly selected to evaluate co-verbal sequences as performed by the ECA-EVA, while the other fifteen evaluated the co-verbal sequences executed by the iCub-simulated robotic unit. The auditory information in the sequences was exactly the same for both ECA-EVA and the iCub simulation. All the participants in the experiment evaluated the observed sequences on a Likert scale from 1 to 5.

8. Discussion

The results obtained from the perception experiments are given in Table 2. The table summarizes the investigated quality of the system with regard to the alignment of speech and non-verbal gestures for ECA-EVA and the iCube simulation. Figure 14 visualizes the obtained scores.

With regard to C1, the mean values show that most of the participants were quite satisfied with the selection of meaningful segments, regardless of the entity that executed them. The overall mean value $M = 3.37$ (SD = 0.05) shows that the symbolical synchronization implemented by POS-pattern-matching identifies many appropriate semiotic phrases that carry the most meaning.

With regard to C2, the mean values show that most of the participants found at least some correlation between the shapes and the speech segments. The overall mean value $M = 3.48$ (SD = 0.11) shows that lexical affiliation that is dependent on semiotic gesture types and meaningful word types may be a viable process whilst generating non-verbal behaviour over text-sequences.

As regards C3, the participants also evaluated the movement as being generally continuous and fluid. The stroke and hold movement-phases were performed viably. The lower rating for the iCub simulation may be attributed to the minimization concept - due to the physical limitations of the simulated robotic unit, several movement phases were eliminated (or prolonged) in
order to lower the dynamics of the simulated movement. However, the overall mean value $M = 3.26 \ (SD = 0.47)$ suggests that the proposed model provides an adequate dynamics of movement.

| Measure | ECA-EVA | iCub simulation | Overall |
|---------|---------|-----------------|---------|
| C1      | 3.40(0.83) | 3.33(0.89) | 3.37(0.05) |
| C2      | 3.56(0.81) | 3.40(0.99) | 3.48(0.11) |
| C3      | 3.6(0.91) | 2.95 (0.80) | 3.26 (0.47) |
| C4      | 3.33 (0.98) | 2.87 (0.74) | 3.10 (0.33) |
| C5      | 3.53 (0.74) | 2.8 (0.94) | 3.17 (0.52) |
| C6      | 3.6(0.74) | 2.87(0.99) | 3.24 (0.52) |
| C7      | 3.53(1.13) | 2.93(1.03) | 3.23 (0.43) |

Table 2. Mean values for evaluating the quality of presentation (standard deviation values in parentheses).

Measure C4 was used in order to evaluate how the gestures align and propagate over the communicative act as a whole. This measure generally describes whether the preceding gesture retracted to a proper position (if at all) and whether the transition between consecutive gestures was executed fluidly and viably. The overall mean value $M = 3.1 \ (SD = 0.33)$ suggests that the participants were quite satisfied with the post-filtering process that eliminates duplicated/unwanted movement phases.

Measure C5 was used in order to evaluate the temporal synchronization between speech and gestures. The mean value $M = 2.8 \ (SD = 0.94)$ for the iCub simulation suggests that in some cases the algorithm eliminates (or prolongs) too many movement phases. However, the overall rating of C5 ($M = 3.17, \ SD = 0.52$) shows that the participants were quite satisfied with gesture timings that were extrapolated based on syllables and their structure.

Measure C6 was used in order to evaluate the speed of movement. The overall rating ($M = 3.24, \ SD = 0.52$) shows that the speed of the execution of the gestures was perceived as viable. The lower rating ($M = 2.87, \ SD = 0.99$) received by the iCub simulation might be again attributed to the limited dynamics and gesture repository intended for robotic units.

Finally, the overall rating for measure C7 ($M = 3.23, \ SD = 0.42$) suggests that the proposed algorithm generally produces an adequate density of movement. In the case of ECAs, some of the gestures could be skipped, whereas in the case of robotic unit simulations additional gestures could be inserted.

To summarize, the post-experiment study shows that the proposed algorithm can generate more natural and viable co-verbal behaviour for ECAs and robotic units alike. In general, the lower scores given to the iCub simulations suggest that the transformation from ECA behaviour to robotic units’ behaviour was regulated using too strict limitations. For instance, if the C5 measure is compared against the C6 measure, it can be observed that in most cases the robotic movement was restricted too tightly and that it was designed so as to be executed too slowly. Similarly, the scores of measures C3 and C4 (for iCub) shows that the restrictions used whilst adapting the gestures to a robotic unit were probably too strict. The duration of some movement phases was extended for too long. In some cases, the algorithm also eliminated some movement phases in order to provide a more fluid movement. Less satisfactory results whilst observing the segments performed with the iCub simulations might generally be attributed to the limitations in the dynamics used as well as the limitations of the gesture repository intended for robotic units.

9. Conclusions

This paper presented a novel TTS-driven non-verbal behaviour system for co-verbal gesture synthesis. The system’s architecture and the linguistic concepts used to identify and better synchronize the non-verbal expressions with verbal information were presented in detail. The system is capable of the semantic, temporal and pragmatic synchronization of verbal and non-verbal behaviour. The synthetic behaviour reflects iconic, symbolic and indexical expressions as well as adaptors. The system can be used in either virtual or physical worlds, in the form of an animation generated by an ECA or as a movement generated by humanoid robotic unit. Further, this paper also discussed and presented the similarities between ECAs and robotic systems (humanoid robots) and the required transformations in the spatial and temporal domains that are required for the TTS-driven non-verbal behaviour generation system to be able to specify behaviour that is completely compatible with robotic units. The proposed system, the concept of annotating communicative behaviour, and the generation of the gesture dictionary, enable us to accurately recreate not only behaviour performed by a real human, but so too does it allow us to generate human-like behaviour based only on pure text sequences. Nevertheless, the behaviour performed by the humanoid
robotic unit may be – at least when compared with the ECA - less natural, less diverse and less intuitive. This issue arises mostly due to the limitations of the physical movement controllers. As a part of our future tasks, therefore, we will continue with the proposed mechanism in building an even larger expressive gesture dictionary, and try to store as many gesture-instances as possible. This will further contribute to the diversity (that is typical of naturalness) and expressive capabilities of both ECAs and humanoid robots.

10. Acknowledgments

This operation was part-financed by the European Union European Social Fund.

11. References

[1] Sargin, M. E., Aran, O., Karpov, A., Ofli, F., Yasin, Y., Wilson, S., Erzin, E., Yemez, Y., Teka, A. M. (2006). Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis. In Proc. of the International Conference on Multimedia and Expo (ICME), July 2006. Toronto, Ontario, Canada, IEEE, pp. 893-896.

[2] Wallbott, H. G. (1998). Bodily Expression of Emotion. In European Journal of Social Psychology. Volume 28, Issue 6, John Wiley & Sons, pp. 879-896.

[3] Gallaher, P. E. (1992). Individual differences in nonverbal behavior: Dimensions of style. In Journal of Personality and Social Psychology. Volume 63, Issue 1, pp. 133-145.

[4] Allwood, J., Ahlsgen, E., Lund, J., Sondqvist, J. (2007). Multimodality in Own Communication Management. In Journal of Current Trends in Research on Spoken Language in the Nordic Countries, Volume 2, pp. 10-19.

[5] Rojc, M., Mlakar, I. (2011). Multilingual and Multimodal Corpus-Based Text-to-Speech System - PLATTOS -. Speech and Language Technologies. Ivo Isp. (Ed.). InTech, ISBN: 978-953-307-322-4, DOI 10.5772/16245.

[6] Bergmann, K., Kopp, S., Eyssel, F. A. (2010). Individualized Gesturing Outperforms Average Gesturing - Evaluating Gesture Production in Virtual Humans. In Proc. of the 10th International Conference on Intelligent Virtual Agents (IVA). LNCS, Volume 6356, Springer, pp. 104-117.

[7] Pelachaud, C. (2009). Modelling Multimodal Expression of Emotion in a Virtual agent. In Philosophical Transactions of the Royal Society B. Volume 364, Issue 1535, pp. 3539-3548.

[8] Poggi, L., Pelachaud, C., De Rosi, F., Carofiglio, B., De Carolis, B. (2005). Greta: A Believable Embodied Conversational Agent. Multimodal Intelligent Information Presentation. Part I. Oliviero Stock and Massimo Zancanaro (Eds.). Volume 27, Springer Netherlands, pp. 3-25.

[9] Zoric, G., Smid, K., Pandzic, I. (2006). Automated Gesturing for Virtual Characters: Speech-driven and Text-driven Approaches. In Journal of Multimedia. Volume 1, Issue 1, pp. 62-68.

[10] Tang, H., Fu, Y., Tu, J., Hasegawa-Johnson, M., Huang, T. S. (2008). Humanoid Audio–Visual Avatar with Emotive Text-to-Speech Synthesis. In IEEE Transactions on Multimedia. Volume 10, Issue 6, pp. 969-981.

[11] Goranka, Z., Pandzic, I. (2008). Towards Real-time Speech-based Facial Animation Applications Built on HUGE Architecture. In Proc. of the International Conference on Auditory-Visual Speech Processing (AVSP). Tangalooma, Australia.

[12] Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Maurizio, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thorisson, K. R., van Welbergen, H., van der Werf, R. J. (2007). The Behavior Markup Language: Recent Developments and Challenges. In Proc. of the 7th international conference on Intelligent Virtual Agents (IVA). LNAI, Volume 4722, Springer Berlin Heidelberg, pp. 99-111.

[13] Krans, A., Kopp, S., Wachsmuth, I. (2002). MURML: a Multimodal Utterance Representation Mark-up Language for Conversational Agents. Technical report SFB 360 Situated Artificial Communicators. Universität Bielefeld.

[14] Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., Vilhjalmsson, H. (2008). Why do Conversational Agents do what they do? Functional Representations for Generating Conversational Agent Behavior. In Proc. of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Estorial, Portugal.

[15] De Carolis, B., Pelachaud, C., Poggi, I., Steedman, M. (2004). APML, a Markup Language for Believable Behavior Generation. Life-like Characters. Tools, Affective Functions, and Applications. Part II. Helmut Prendinger and Mitsuru Ishizuka (Eds.). Springer Berlin Heidelberg, pp. 65-85.

[16] Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., Joublin, F. (2012). Generation and Evaluation of Communicative Robot Gesture. In International Journal of Social Robotics. Special Issue on Expectations, Intentions, and Actions. Volume 4, Issue 2, Springer Netherlands, pp.201-217.

[17] Miyashita, T., Shinozawa, K., Hagita, N. (2006). Gesture Translation for Heterogeneous Robots. In Proc. of the 6th IEEE-RAS International Conference on Humanoid Robots. Genova, pp. 462-467.

[18] Mlakar, I., Rojc, M. (2011). Towards ECA’s Animation of Expressive Complex Behaviour. Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues. Anna Esposito, Alessandro Vinciarelli, Klara Vics, Catherine Pelachaud, Anton Nijholt (Eds.). LNCS, Volume 6800, Springer Berlin Heidelberg, pp. 185-198.
[19] Tikhanoff, V., Cangelosi, A., Fitzpatrick, P., Metta, G., Natale, L., Nori, F. (2008). An Open Source Simulator for Cognitive Robotics Research: The Prototype of the iCub Humanoid Robot Simulator. In Proc. of the 8th Workshop on Performance Metrics for Intelligent Systems (PerMIS), August, 2008. New York, USA, ACM, pp. 57-61.

[20] Mehu, M., D’Errico, F., Heylen, D. (2012). Conceptual Analysis of Social Signals: the Importance of Clarifying Terminology. In Journal on Multimodal User Interfaces. Volume 6, Issue 3-4, Springer-Verlag, pp. 179-189.

[21] Peirce, C. S. (1958). Collected Papers of Charles Sanders Peirce, 1931-1958. 8 Vols. MA: Harvard University Press.

[22] McNeill, D. (2005). Gesture and Thought. The University of Chicago Press, ISBN: 9780226514628.

[23] Schegloff, E. A. (1985). On Some Gestures’ Relation to Talk. Structures of Social Action: Studies in Conversation Analysis. J. Maxwell Atkinson and John Heritage (Eds.). Cambridge University Press, pp. 266-298.

[24] Kopp, S., Wachsmuth, I. (2004). Synthesizing Multimodal Utterances for Conversational Agents. In Journal of Computer Animation and Virtual Worlds. Volume 15, Issue 1, pp. 39-52.

[25] Gienger, M., Janßen, H., Goerick, S. (2005). Task-oriented Whole Body Motion for Humanoid Robots. In Proc. of the IEEE-RAS International Conference on Humanoid Robots. Tsukuba, pp. 238-244.

[26] Pelachaud, C., Gelin, R., Martin, J. C., Quoc, A. L. (2010). Expressive GesturesDisplayed by a Humanoid Robot during a Storytelling Application. Second International Symposium on New Frontiers in Human-Robot Interaction (AISB), Leicester, UK.

[27] Ng-Thow-Hing, V., Pengcheng, L., Okita, S. (2010). Synchronized Gesture and Speech Production for Humanoid Robots. In Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Taipei, pp. 4617-4624.

[28] Miyashita, T., Shinozawa, K., Hagita, N. (2006). Gesture Translation for Heterogeneous Robots. In Proc. of the 6th IEEE-RAS International Conference on Humanoid Robots. Genova, pp. 462-467.

[29] Luo, R. C., Chien-Chieh, H., Shu-Ruei C., Yi-Jeng T. (2011). Speech Synchronization between Speech and Lip Shape Movement for Service Robotics Applications. In Proc. of the 37th Annual Conference on IEEE Industrial Electronics Society (IECON). Melbourne, pp. 2261-2266.

[30] Rojc, M., Kačič, Z. (2007). Time and Space-efficient Architecture for a Corpus-based Text-to-speech Synthesis System. In Journal Speech Communication. Volume 49, Issue 3, pp. 230-249.

[31] Arnaud, A. (2000). Study and Implementation of a Weighted Finite-state Library – Application to Speech Synthesis. Faculté Polytechnique de Mons, Belgium.

[32] Chiu, C. C., Marsella S. (2011). How to Train Your Avatar: A Data-driven Approach to Gesture Generation. Intelligent Virtual Agents. Hannes Högni Vihjálmsson, Stefan Kopp, Stacy Marsella, Kristinn R. Thorisson (Eds.). LNCS, Volume 6895, Springer Berlin Heidelberg, pp. 127-140.

[33] Taylor, P. (2000). Analysis and Synthesis of Intonation using the Tilt Model. In Journal of the Acoustical Society of America. Volume 107, pp. 1697-1714.

[34] Breiman, L., Freidman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees. Chapman & Hall/CRC, New York.

[35] Valbonesi, L., Ansari, R., McNeill, D., Quek, F., Duncan, S., McCullough, K. E., Bryll, R. (2002). Multimodal Signal Analysis of Prosody and Hand Motion: Temporal Correlation of Speech and Gestures. In Proc. of the European Signal Processing Conference (EUSIPCO). Tolouse, pp. 75-78.

[36] Mlakar, I., Kačič, Z., Rojc, M. (2012). Form-oriented Annotation for Building a Functionally Independent Dictionary of Synthetic Movement. Cognitive Behavioural Systems. Anna Esposito, Antonietta M. Esposito, Alessandro Vincarelli, Rüdiger Hoffmann, Vincent C. Müller (Eds.). LNCS, Volume 7403, Springer Berlin Heidelberg, pp. 251-265.

[37] Kipp, M. (2001). Anvíl - A Generic Annotation Tool for Multimodal Dialogue. In Proc. of the Interspeech Conference. Aalborg, pp. 1367-1370.

[38] Salem, M., Kopp, S., Wachsmuth, I., Joublin, F. (2011). A Multimodal Scheduler for Synchronized Humanoid Robot Gesture and Speech. Book of Extended Abstracts of the 9th International Gesture Workshop (GW2011), pp. 64-67.

[39] Prillwitz, S. (1989). HamNoSys. Version 2.0 Hamburg Notation System for Sign Languages: An Introductory Guide. Signum-Verlag.

[40] Crasborn, O., van der Hulst, H., van der Kooij, E. (2001). SignPhon: A Phonological Database for Sign Languages. Sign Language and Linguistics. Volume 4, Issue 1-2, pp. 215-228.

[41] Hogrefe, K., Ziegler, W., Goldenberg, G. (2011). Measuring the Formal Diversity of Hand Gestures by their Hamming Distance. Integrating Gestures: the Interdisciplinary Nature of Gesture. Stam Gale and Mika Ishino (Eds.). John Benjamins Publishing Company, pp. 75-88.

[42] Kipp, M. (2001). From Human Gesture to Synthetic Action. In Proc. of the Workshop on Multimodal Communication and Context in Embodied Agents, held in conjunction with the 5th International Conference on Autonomous Agents (AGENTIS). pp. 9-14.

[43] Allwood, J. (2001). Dialog Coding – Function and Grammar: Göteborg Coding Schemas. In Gothenburg Papers in Theoretical Linguistics. Volume 85, University of Göteborg, pp. 1-67.
[44] Gravano, A., Hirschberg, J., Benus, S. (2012). Affirmative Cue Words in Task-oriented Dialogue. In Journal Computational Linguistics. Volume 38, Issue 1, MIT Press, Cambridge, pp. 1-39.

[45] Luke, K. K., Thompson, S. A., Ono, T. (2012). Turns and Increments: A Comparative Perspective. Special Issue: Cross-Linguistic, Multi-Modal, and Grammaticization Perspectives on Turn Continuation in Conversation. Volume 49, Issue 3-4, pp. 155-162.

[46] Hoffman, G., Kubat, R., Breazeal, C. (2008). A Hybrid Control System for Puppeteering a Live Robotic Stage Actor. In Proc. of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Munich, pp. 354 - 359.

[47] Kopp, S., Tepper, P., Cassell, J. (2004). Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output. In Proc. of the 6th International Conference on Multimodal Interfaces (ICMI). ACM, New York, pp. 97–104.

[48] Lima, J. L., Goncalves, J. C., Costa, P. G., Moreira, A. P. (2008). Humanoid Robot Simulation With a Joint Trajectory Optimized Controller. In Proc. of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). Hamburg, pp. 986-993.

[49] Cassell, J., Vilhjalmsson, H. H., Bickmore, T. (2001). Beat: The Behavior Expression Animation Toolkit. In Proc. of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). ACM, New York, pp. 477–486.

[50] Lee, J., Marsella, S. C. (2006). Nonverbal Behavior Generator for Embodied Conversational Agents. In Proc. of the 6th International Conference on Intelligent Virtual Agents (IVA). Springer-Verlag, Berlin, pp. 243–255.

[51] Le, Q. A., Pelachaud, C. (2011). Expressive Gesture Model for a Humanoid Robot. Affective Computing and Intelligent Interaction. Sidney D’Mello, Arthur Graesser, Björn Schuller, Jean-Claude Martin (Eds.). LNCS, Volume 6975, Springer-Verlag, Berlin, pp. 224-231.

[52] Levine, S., Krähenbühl, P., Thrun, S., Kolton, V. (2010). Gesture Controllers. In Journal Transactions on Graphics (TOG). Volume 29, Issue 4, ACM, New York, pp. 1-11.

[53] Hostetter, A. B., Alibali, M. W. (2008). Visible Embodiment: Gestures as Simulated Action. In Journal Psychonomic Bulletin & Review. Volume 15, Issue 3, Springer-Verlag, pp. 495-514.

[54] McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. Psychology/cognitive science. University of Chicago Press, ISBN: 978-022-656-132-5.

[55] Taylor, P., Black, A. W., Caley, R. (2001). Heterogeneous relation graphs as a formalism for representing linguistic information. In Journal Speech Communication. Volume 33, Issue 1-2, pp. 153-174.