Capturing Form of Non-verbal Conversational Behavior for Recreation on Synthetic Conversational Agent EVA

[1]IZIDOR MLAKAR, [2]MATEJ ROJC
[1]Roboti c.s. d.o.o, [2]Faculty of Electrical Engineering and Computer Science, University of Maribor
[1]Tržaška cesta 23, [2]Smetanova ulica 17
SLOVENIA

Abstract: - Several features of human-human conversation have to be accounted for in order to recreate conversational behavior on a synthetic model, as natural as possible.. Spontaneous conversations are a combination of multiple modalities (e.g. gestures, postures, gazes, expressions) in order to effectively convey information between participants. This paper presents a novel process for capturing the forms of motion performed during spontaneous conversations. Furthermore, it also addresses the process of transforming the captured motions' descriptions into high-resolution, expressively transformable behavioral scripts. The aim of the research was design a process that will allow building a high-resolution motion dictionary. The dictionary is to be presented as a set of expressively transformable behavioral scripts, each capturing the expressive details from a spontaneous conversation (e.g. spatial, repetitive, structural, and temporal features).

Key-Words: - multimodal annotation, conversational behavior, high-resolution annotation, motion dictionary, analysis of non-verbal behavior

1 Introduction

Conversational behavior consists of verbal and non-verbal features correlated by a sophisticated mechanism driven by communicative and non-communicative functions [1]. The concept of conversational behavior is, therefore, based on the presumption that people communicate by using their voices and their bodies [2]. The most natural and user-friendly human-machine interaction (HMI) interfaces have proven to be those that incorporate embodied conversational agents (ECAs).

ECAs have already been used within a wide-range of applicative scenarios, such as games [3], virtual-worlds [4], web-based interfaces [5] educational [6] and other commercial/non-commercial applications. From amongst these multimodal interfaces, lip-sync is regarded as one of the elementary processes for the imitation of non-verbal conversational behavior. The goal of the lip-sync process is to move lips and face muscles, and to correlate them with the spoken utterance [7]. Another of the primitives is gaze. In spontaneous conversation, gaze is involved in several communicative functions of behavior (turn taking, accentuation, and organization). By propagating the mechanism of gaze, humans are also capable of attracting visual attention [8]. Finally, gestures [9] and facial expressions [10] are the elements of non-vernal behaviour that most evidently provide additional information about the spoken utterances. Gestures have the capacity of transforming the speakers' thoughts into visible objects. Together, however, facial expressions and gestures are used to express emotions, attitudes, ease, exhortment, approval and other states of mind.

ECAs may be a widely used concept within human-machine interaction interfaces; however, they still lack naturalness [11]. The wooden appearance is even more evident when motion is reproduced based on unknown input text sequences (text-centric). The major disadvantage of text-centric motion generation is that most of the contextual information (acoustic signal information, emotion, speaker-listener relations, intent etc.) is missing. Furthermore, if the text is previously unknown to the system, the scenario-oriented generation of synthetic non-verbal behavior may be impossible. However sequences of utterances carry some linguistic information in the forms of syntax, morphology, and semantics. If the linguistic rules are derived as based on the observation of human-human spontaneous conversation, the resulting synthetic behavior has the potential to evoke some-sort of social-response.

The motivation of the presented work is a more natural, text-centric synthesis of non-verbal behavior. The concept has already been briefly discussed in [12]. This paper discusses the concepts of annotation and transformation in detail. The paper is structured as follows; section 2 reports on several studies relating to capturing and coding the different features of non-verbal behavior, and the recreation of the described behavior on a synthetic agent. Sections 3 and 4 describe the presented concepts of capturing, coding and recreation in detail. Section 5 describes the multimodal-corpora used and the results of the annotation. This paper is concluded by a discussion, and details of future plans.

2 Background

Several annotation schemas and annotation tools have emerged in order to examine the communicative and non-communicative functions of conversation. These processes have provided further insights into how non-verbal behavior is structured, organized, and how it is synchronized with verbal information. At the highest level (functional-level) of understanding, researches have explored human mechanisms used for managing communication [13]. The MUMIN coding scheme [14], for instance, focuses on the annotations of three communicative functions: the feedback, turn-management, and sequencing functions. Bergmann & Kopp in [15] studied the correlations between contextual factors (referent features, discourse) and gesture features. These correlations were then classified as systematic (shared amongst speakers), or idiosyncratic (inter-individually different). Empirical investigations on a functional level have offered insights into motives and a correlation between verbal and non-verbal behavior. However, these annotations only coarsely describe the forms and dynamics of motion.

Form-oriented systems have been developed in order to capture more-detailed information on the structure and dynamics of the produced motion. The representative form-oriented system is FORM [16]. The temporal and spatial dynamics of motion are encoded within FORM based on those articulators propagating the motion. Due to the level of detail and the complexity that FORM describes within the annotation graphs, the motion coding is highly time consuming. As a solution, the authors in [17] compensate for FORM's complexities by introducing a 3D pose-editor integrated within an ANVIL annotation tool [18]. This concept allows hand-gestures to be fine-tuned based on different end-poses and movement phases [19], and then interpolated as gesture phrases and units [20] into gradual synthetic-motion. Similarly, the authors of [21] have defined an annotation schema for studying those spatial references occurring during conversation (e.g. gesture, gaze, and posture).

Existing annotation approaches and schemas provide different levels of understanding non-verbal behavior and its forms. The functional annotations provide detailed data on correlations between the produced non-verbal behaviors in the forms of: 1) functions within dialogue, 2) semantic/morphological structures and 3) state of the body/mind of the observant. Form-oriented annotations provide detailed data about the structure, power, and other expressive features of motion. The semi-functional forms are the most economical. These annotations combine some parts of the functional- and some parts of the form-oriented schemas.

The primary goal of our work was to synthesize non-verbal human-like behavior based on pure text sequences, and by using ECA EVA [22]. In order to do this a concept for describing and transforming annotated behavior into EVA's expressive motion templates is presented in this paper.

This novel concept is based on manually-annotating informal multi-speaker dialogs within a form-oriented annotation schema. It adapts the concepts presented in [17] and [21]. The presented schema captures the expressive features of moving body parts at high-resolution, and can be directly transformed into co-verbal motion generated by ECA EVA. It also captures several linguistic contextual factors e.g. at what utterance is the motion initiated, propagated, and subsided, what combination of utterances triggers motion sequences, at what part of the utterance is the motion phase triggered, etc.

The following section addresses the concept of capturing conversational movement and its expressive details at high-resolution.

3 Capturing movement at high-resolution

A 2-staged approach is suggested in order to be able to reproduce co-verbal motion, as presented in figure 1. This approach consists of empirically analyzing the spontaneous informal multi-speaker dialog and the generation of procedural animations in the forms of EVA EVENTS. These events can be directly animated on the embodied conversational agent EVA.

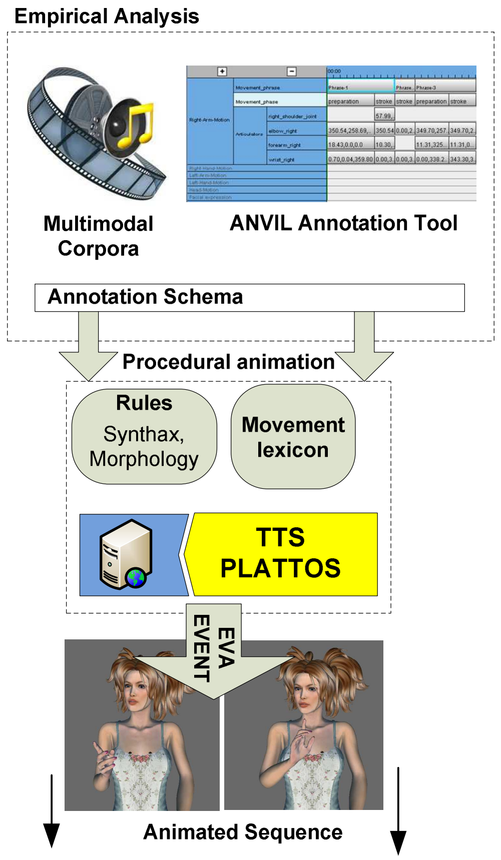


Fig. 1: Architecture of annotation and transformation of communicative behavior into synthetic movement

The empirical analysis of a motion's expressive details is performed manually and involves: (a) observing the multimodal corpora and (b) coding the observed data based on an annotation schema within an annotation tool. The spatial, repetitive, and temporal features of motions are annotated separately for different body parts and their corresponding articulators. However, in contrast to [17] and [21], the topology and the formal model for the annotation do not differ between body-parts. Figure 1 shows the topology of the synthetic conversational behavior generation process. The empirical analysis of spontaneous informal multi-speaker dialog was performed by using the ANVIL annotation tool, and multimodal corpora of spontaneous conversational behavior in the Slovenian language [23], and the annotation topology, as discussed in this paper.

The main advantage of ANVIL is its rich tier system. This tier system is hierarchically-oriented with an underlying XML level. The tier system is also compatible with SQL-based databases. Since EVA-Script is also a hierarchically-oriented XML concept, the transformations from annotation to EVA-Script based movement lexicons are quite straightforward. In addition the ANVIL tool is also well-suited for capturing motion details within form-oriented systems. It provides freedom of attributes and the possibility of forming hierarchical relationships between tiers.

TV interviews and theatrical plays have shown themselves to be very usable source of real-life behavior. However, based on corpora of spontaneous behavior more credible (more believable) co-verbal sequences may be produced [24]. The annotation corpora used was based on informal dialog with a high-degree of spontaneous co-verbal movement. It contained four accurately-transcribed sessions, each with durations of about 50 minutes (approximately 200 minutes of the material). Within each session, there were five different participants; however, only two of the participants were always present during all four sessions, whereas the other participants were different in each talk-show. In each session at least 3 participants actively contributed to the communicative dialog.

Manual annotation is then performed in the form of a series of main tracks that hierarchically group the expressive features of movement, based on body-parts. The empirical analysis is performed offline and results in a) linguistic rules for the correlation between verbal and non-verbal behavior and b) a movement lexicon containing expressively-transformable EVA-SCRIPT based movement descriptions.

The behavior generation process suggests text-centric generation. The text-to-speech engine (TTS) PLATTOS [25] is used to transform unknown input text sequences into voiced. The TTS engine is also used to correlate the spoken utterances and the motion sequences performed by the conversational agents. The correlation between utterances and motion is performed contextually and temporarily. The contextual correlation therefore selects motion models, based on linguistic rules and adjusts them to. In the context of text driven animation's these rules include morphological relations, syntactic relations, and semantic relations. The contextual correlation therefore selects what movements within the motion lexicon are going to be generated. The temporal correlation is performed based on the duration of utterances and the linguistic rule being used for movement selection. The result of the procedural animation block is an EVA-EVENT, animated by the synthetic agent EVA.

3.1 Annotation Schema: The formal model and topology

The formal model and the topology of the annotation schema are derived based on [17] and

[21]. The spatial and temporal configurations of body-parts are defined in form of end-poses. Figure 2 shows the general topology of the presented formal model.

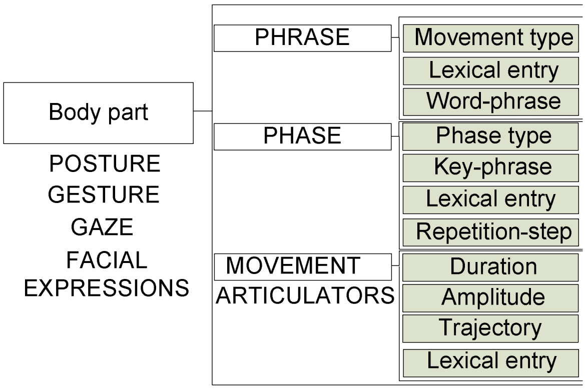


Fig. 2: The Formal model of annotation of body movement

As shown in Figure 2, the observation and coding of conversational behavior is performed separately for each body-part. The body-parts are also used to define the four concepts of non-verbal behavior:

- **POSTURE** [left and right arms]
- **GESTURE** [left and right hands]
- **GAZE** [neck and eye regions of the head]
- **FACIAL EXPRESSIONS** [facial region]

The scheme allows for annotating the movement lemmas (movement phases, movement phrases, and movement units). The movement of the observed body-part is therefore described by *movement phase*, *movement phrase* and the *articulators* propagating the observed movement. The *articulators* are in terms of [22] control units that model the final-poses overlaid by the articulated 3D model.

3.2 Annotation Schema: Articulators

Each articulator is described by its spatial configurations (*trajectory*) and the duration interval within which the articulator traverses from its previous spatial configuration to the described spatial configuration. The *trajectory* also defines whether the transition between spatial configurations is linear (direct), circular elliptic or rectangular.

The lexical entry describes the abstract label of the configuration (of the end-pose) described within the movement lexicon. Each lexicon entry can store several similar spatial configurations. The amplitude is used when the lexicon already stores a quite similar (or even exactly the same) configuration. It describes the spatial extent of the lexical entry being generated by the conversational agent.

3.3 Annotation Schema: movement phases

The movement phase [19][26] (Figure 2) hierarchically groups the articulators into a sequence of elements contributing to the transformation of the observed body-part. T Movement phases were extrapolated based on the definition of gesture phases. The movement phase therefore describes those spatial configurations within the four possible stages of motion (phase-types):

- *stroke*: phase of movement, where the dynamics and shape are manifested by the greatest clarity (the part of the motion with the most energy). The stroke phase is synchronous with the co-expressive speech. If however strokes are asynchronous, they slightly precede the speech to which they link semantically.
- *preparation*: phase of movement that leads-up to the stroke (initiates stroke). Within this phase the movement is prepared for, withheld if need be until the co-expressive speech is ready.
- *recovery(retraction)*: phase of movement that transfers the gesture into a relaxed or withdrawn state (rest pose). If the speaker moves to a new stroke, the retraction phase may not exists (repeated strokes).
- *pre- and post-stroke hold (hold)*: arrives before and/or at the end of the stroke, as a nucleus of the gesture phrase (maintains the semantic activity of stroke). The existence of hold phase suggests that the stroke and the co-expressive speech express an idea created in advance.

Each movement-phase is defined by its phase-type, key-phrase, lexical entry and the repletion step. The key-phrases (e.g. key-words) define the temporal dynamics of movement. For instance each of the movement phases (preparation, stroke, hold, and retraction) can have a key-utterance (e.g. key-word, syllable) that triggers its propagation. The full span of phases (from preparation to retraction) defines the so called lifetime of movement.

It is quite important for the exact moment of the phase initiation to be defined, whilst generating text-centric conversational behavior. The lexical entry, similarly as under articulators, describes the abstract label of the body-configuration. The repetition-step is important within the repetitive motions. It defines the how many times the lifetime of movement is repeated when linked with co-verbal speech. It can consequently be used whilst observing the changes of spatial configurations depending on the stage of repetition (e.g. the cycle of movement).

The amplitude is used when the lexicon already stores quite similar (or even exactly the same) configuration. It describes the spatial extent of the lexical entry being generated by the conversational agent.

3.4 Annotation Schema: movement phrases

The movement phrase (Figure 2) joins sequential movement phases into continuous movement linked with an idea unit. The movement phrases describe the full span of phases (from preparation to retraction). Each movement phrase, therefore, contains a mandatory stroke and optional preparation, hold, and retraction phases. The optional retraction and preparation phases can be observed only on the borders of the movement phrase.

The movement types are used in order to describe the dimension of movement. These types are mostly derived based on McNeill's works on hand-gestures ([20][26]). The *adaptor* comprises non-verbal behavior not participating directly within the meaning of speech. However, the adaptors are used within the communicative functions of the conversational behavior. In multi-speaker dialogs such a movement-type is quite frequent and is used in listener behavior, turn-taking/turn-giving and sequencing functions.

The *iconic motion* (displaying images of objects and actions), the *metaphoric motion* (displaying images from the abstract usage of form and space), *the deictic* motion (pointing in some direction), and *emblems* (conventionalized signs) group non-verbal behavior that directly participates in the meaning of speech. Such non-verbal behavior is used within non-communicative functions of conversational behavior.

The word-phrase attribute indicates for which words the movement is propagated. If these words/phrases are morphologically-labeled and grouped into semantic/syntactic rules, they provide the basic correlation between a movement phrase and hte general text (e.g. which movement phrases ECA displays during different word sequences).

The following section of this paper discusses the coding process performed within the ANVIL annotation environment.

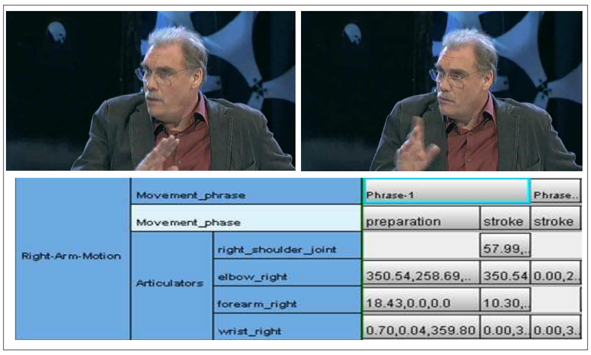
4 Coding the conversational behavior using ANVIL

Most of the information carried by co-verbal movement is presented through stroke and post-stroke-hold movement phases (nucleus of movement). However, the events prior and post the nucleus are also important when recreating annotated motion. Namely, these events carry part of the dynamical features and also information on the best way to transit from one motion phrase to another. The border phases can also indicate the notion of mental processes such as thinking, consideration, hesitation etc. The coding process takes into account only those end-poses that occur at the borders of movement phases. The transition (in-between poses) between two end-poses is generated by the ECA automatically within animation period. Figure 3 shows an example of the ANVIL interface and the encoded motion for the right-arm.

The movement phases are firstly defined. The stroke phase is, in general, defined based on the significance of the motion. The hold-phases are defined by those segments, where there is no movement performed pre/after the stroke phase (end-poses are relatively static). The retraction-phase labels those motion segments that drive the observed body part into a relaxed (neutral) state. The preparation-phase denotes those segments that drive the observed body part into a stroke phase. As defined by the formal-model, each motion-phase is assigned a utterance (key-phrase) found at the borders of phase and utterances co-occurring with movement (indicator of shared co-verbal overlay of an idea). Additionally, each phase is also assigned a lexical name. If hands are observed, then hand-shapes denote the lexical name of the motion. When the observed body-part is an arm or head, the trajectory of the end-pose defines the lexical name. The lexical names are mostly derived based on the Posture Scoring System [27].

In the second stage of annotation process, the movement phrases are defined and labeled based on coded movement phases. All the words occurring during the phrase are also codded in addition to the movement type and lexical entry.. As already mentioned, these sequence indicate when and which movement lexicon the ECA should use whilst generating text centric conversational behavior.

Finally, the spatial configuration is encoded for each articulator of the observed-body parts (Figure 3).



Right-Arm-Motion	Movement_phrase		Phrase-1		Phrase..
	Movement_phase		preparation	stroke	stroke
	Articulators	right_shoulder_joint		57.99,.	
		elbow_right	350.54,258.69,..	350.54	0.00,2..
		forearm_right	18.43,0.0,0.0	10.30,.	
		wrist_right	0.70,0.04,359.80	0.00,3.	0.00,3.

Fig. 3: Anvil interface for coding conversational behavior

The example in Figure 3 shows how conversational video segments are analyzed and coded in ANVIL. The annotator observes a movement of right arm that consists of 2 movement phrases and three movement phases. The first image of the speaker (Figure 3 top-left-hand-side), shows the end-pose at the ending-border of the movement phrase named *Phrase-1*. The second image (Figure 3 top-right-hand-side) shows the end-pose at the ending border of the movement phrase named *Phrase-2*. Figure 3 also shows that both movement phrases contain a mandatory stroke-phase. However only *Phrase-1* contains less-energetic movement and can thus be described as the preparation-phase. The phenomenon observed in Figure 3 is called the *repeated stroke* mechanism. This movement transforms from one stroke configuration to the next stroke configuration with no intermediate hold or retracted spatial configurations. The word sequence intthe example of Figure 3 relates as: "Tako že **ne! Ampak**…" *(Not like that! But..)*.

The right-arm motion is, according to EVA-Script, defined by the spatial configurations (trajectory) of the following articulators: "*right_shoulder_joint*", "*elbow_right*", "*forearm_right*", and "*wrist_right*". However, during the preparation phase of Phrase-1, the "*right_shoulder_joint*" has no influence on the spatial configuration of the end-pose. Therefore no coding is performed for this articulator. The same also applies to both, the "*right_shoulder_joint*" and the "*forearm_right*" articulators within the stroke phase of *Phrase-2*.

The following section describes the coding of the articulator's spatial configuration.

4.1 Capturing and coding spatial configuration of end-poses

The spatial configurations of the end-poses describe the spatial features of those articulators manifesting the observed end-pose (and the in-between poses). As described in [22] there are three types of articulators, the *joint-based* articulators, the *morphed-shape-based* articulators and the *inverse-kinematics-based (IK)* articulators.

Those articulators that are *joint-based* are mostly used when generating head, eye, and arm and hand synthetic movements. The spatial configurations of such articulators are defined by adjusting their HPR values (*Heading*, *Pitch and Roll*). These types of articulators can define any given spatial configuration with no predefined rules (except spatial restrictions). However, the capturing of their spatial configurations requires the largest amount of resources (especially time). The *morphed-based* articulators are mostly used whilst generating facial expressions, emotions, and other facial and eye-region configurations (e.g. eye blinking, raising brows, etc.). The spatial configurations of such articulators are defined by their amplitude (the translation on the X axis). These types of articulators are also convenient and require a lot less coding time. However, they only enable a finite set of forms for coding into. The *inverse-kinematics-based (IK)* articulators require the least coding time. They offer a limited set of full body-part spatial configurations. For instance, by translating one IK articulator in the 3D space, a proper HPR configuration can be achieved for the right-arm's articulators. However, the *(IK)* articulators are based on *inverse kinematic rules* and therefore offer only a limited set of body postures.

Joint-based and morphed-shape-based articulators were mostly used within the context of the presented work. Two methods were devised in order to capture the spatial configuration. The first method of movement coding is shown in Figure 4. It involves the use of an animation engine's built-in pose editor, and its placer panel.

The animation engine, provided by EVA-Framework allows for any spatial approximation to be built on-line. This means that by visually-adjusting the schema-specified articulators, the annotators can model the observed end-pose directly onto the embodied conversational agent. *The panel for selecting articulators* allows the annotator to select the configuration options for the observed articulator, and automatically opens the placer panel. The *''placer''* panel allows the annotators to adjust the articulator values in the forms of adjusting the HPR and the translation attributes of the articulator. All the values are relative to the parent of the articulator. When the spatial configuration of the parent changes, the spatial configuration of the child adjusts accordingly. The values obtained by the placer panel are inserted into the annotation editor as trajectory attributes.

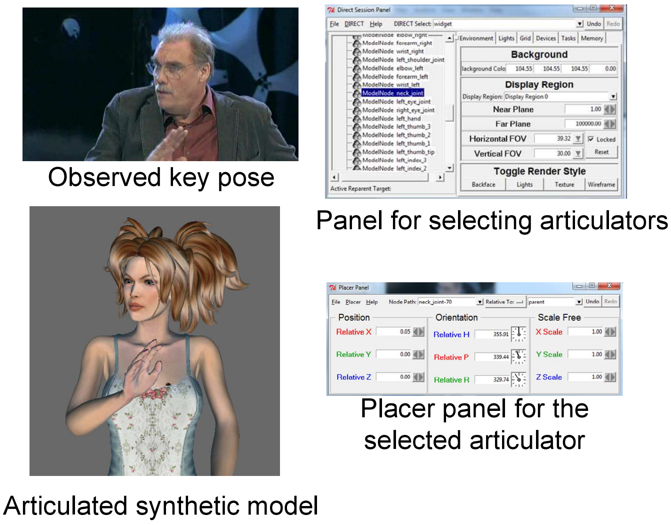
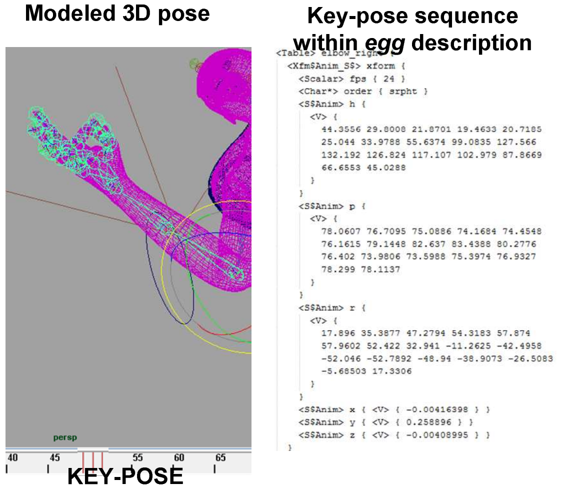


Fig. 4: Interface for capturing the spatial configurations of end-poses, by using the built-in method

The EVA-Framework's animation engine also supports the exchangeable articulator models (e.g. *X, egg, bam*). The final pose (or even the movement sequence) can therefore also be modeled in any 3D modeling tool (e.g. Maya 3D[1], Blender[2], Daz3D[3]) and exported that supports those types of exchangeable models. Figure 5 shows an example of modeling the end-pose by using an external 3D modeling tool. By using an external 3D modeling tool the time extensively reduces for approximating the articulated model's end-pose to the observed conversational pose.



```
<Table> elbow_right {
  <Xfm$Anim_S$> xform {
    <Scalar> fps { 24 }
    <Char*> order { srpht }
    <S$Anim> h {
      <V> {
        44.3556 29.8008 21.8701 19.4633 20.7185
        25.044 33.9788 55.6374 99.0835 127.566
        132.192 126.824 117.107 102.979 87.8669
        66.6553 45.0288
      }
    }
    <S$Anim> p {
      <V> {
        78.0607 76.7095 75.0886 74.1684 74.4548
        76.1615 79.1448 82.637 83.4388 80.2776
        76.402 73.9806 73.5988 75.3974 76.9327
        78.299 78.1137
      }
    }
    <S$Anim> r {
      <V> {
        17.896 35.3877 47.2794 54.3183 57.874
        57.9602 52.422 32.941 -11.2625 -42.4958
        -52.046 -52.7892 -48.94 -38.9073 -26.5083
        -5.68503 17.3306
      }
    }
    <S$Anim> x { <V> { -0.00416398 } }
    <S$Anim> y { <V> { 0.258896 } }
    <S$Anim> z { <V> { -0.00408995 } }
  }
```

Fig. 5: Defining end-poses using Maya 3D, key-framing and egg-based exchangeable 3D articulated model

The process of coding the spatial configurations using an external 3D tool requires the observed conversational pose to be firstly modeled as a key-frame pose (*key-framing* [28]) within the external 3D modeling tool. Figure 5, left-hand-side, shows an example interface and key-framing process within Maya 3D. After the pose is modeled, the selected joint- chain can be exported as a pose (or an *animated sequence*) in the form of an exchangeable 3D model. Figure 5, *right-hand-side*, shows an example of an egg-based description of an animated sequence. The description within the exchangeable 3D model stores the spatial configurations for all of the selected articulators (within the joint chain). The temporal information, however, is ignored.

Depending on the format of the exchangeable 3D model, the spatial configuration values can be transferred as annotation coding either manually or automatically. Automatic transformation is currently provided for *egg*- and *bam-based*, exchangeable 3D model formats. This process searches for the schema defined articulators (e.g. elbow right) and transfers their spatial values into the annotation tier of the articulator. Each spatial entry is temporally mapped to the pre-marked phase of the observed conversational behavior.

4.2 Capturing facial expressions

Facial expressions are annotated based on facial action points (FAPs), predefined facial expressions, or even emotions. The models for defining and describing expressions are based on the MMI facial-expressions database [29]. The level of exposure ranges from 1-10. When encoding facial-expressions, the annotators encode the associated level of the exposure for the articulators manifesting the overlaid facial spatial configuration.

4.3 Capturing and coding the temporal information of end-poses

The proposed annotation schema captures temporal information in the forms of the durations of the movement phases and phrases. In the context of procedural animation, the duration of a movement phase relates to the time (number of animated frames), within which the articulated 3D model is transformed from its current spatial configuration to the spatial configurations of end-poses described within the tiers of the articulators.

The following section of this paper discusses a process for the automated transformation of

[1] Maya 3D - http://usa.autodesk.com/maya/

[2] Blender - http://www.blender.org/

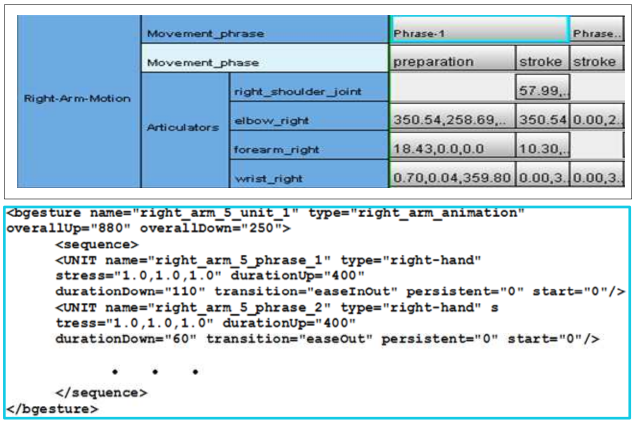
[3] Daz3D - http://www.daz3d.com/i/products/daz_studio

annotated data into expressively adjustable movement description.

5 Transforming annotation into synthetic behavior - imitating the speaker

The EVA framework and PLATTOS TTS system support the concepts of expressive motion, and co-verbal behavior generation from a general text. Any ‘’annotated’’ speaker can be imitated by ECA EVA. By using the process of automatic transformation the annotated data is converted into movement phases, movement phrases, and also into complete conversational behavior movement segments. The correlation between annotation and EVA-Script is shown in Figure 6.

Annotated data



Right-Arm-Motion	Movement_phrase		Phrase-1		Phrase..
	Movement_phase		preparation	stroke	stroke
	Articulators	right_shoulder_joint		57.99,..	
		elbow_right	350.54,258.69,..	350.54	0.00,2..
		forearm_right	18.43,0.0,0.0	10.30,..	
		wrist_right	0.70,0.04,359.80	0.00,3..	0.00,3..

```
<bgesture name="right_arm_5_unit_1" type="right_arm_animation"
overallUp="880" overallDown="250">
      <sequence>
      <UNIT name="right_arm_5_phrase_1" type="right-hand"
      stress="1.0,1.0,1.0" durationUp="400"
      durationDown="110" transition="easeInOut" persistent="0" start="0"/>
      <UNIT name="right_arm_5_phrase_2" type="right-hand" s
      tress="1.0,1.0,1.0" durationUp="400"
      durationDown="60" transition="easeOut" persistent="0" start="0"/>

            •    •    •

      </sequence>
</bgesture>
```

Movement template

Fig. 6: Correlation between annotation and EVA-Script

Figure 6 shows the correlation between an EVA-Script-based description of expressive movement, and the annotation topology proposed in this paper. It demonstrates the correlation between annotation and movement template for right-arm motion (based on EVA-Script).

Each movement phrase (for each body-part) is identified based on its movement phrase track. However, if no movement phrase is specified, the movement phrase is defined automatically, based on the observed movement phases. The borders of the movement phrases are defined by those phases surrounding the stroke-phase. In general, each movement phrase contains at least one stroke-phase. Other phases are optional. However, a movement phrase comprises at most 5 movement phases.

Movement phrase, in terms of EVA-Script, defines a motion template of a complete idea unit that ECA can express int the forms of conversational behavior. The overall duration of the template is denoted by the durations of the preparation, stroke, and hold movement phases. The retraction phase then defines the overall duration of the subsidence. If there is no retraction-phase indicated, the retraction time is denoted as half of the last stroke’s duration

Movement-phases’ articulators propagate the motion of the observed body part into sequences of parallel motions (<sequence><parallel> blocks in EVA-Script). In addition, movement phases define the key-frame interpolation of animated motion. EVA-Script defines three types of motion interpolation:

- “easeIn” – reserved for the preparation movement phase and for any motion followed by post stroke hold,
- “easeOut” – reserved for post stroke hold and retraction phases,
- “easeInOut” – reserved for any movement phrases that contain stroke- hold combinations.

The annotated articulators are mapped into EVA Script’s UNIT tags and inserted into EVA-Script’s “<sequence><parallel>” blocks. The annotated values of each articulator define its transition type (animation interpolation), and the value to which it transits. In addition to the spatial expressive dimension, the articulators can, within the temporal borders of the movement-phase, also define their own “local” temporal features. For instance, if the modeling of an articulator is delayed, a start-attribute of UNIT tag is set. The annotated articulators and their attributes are processed individually for each phase block.

The persistency and delay expressive features of the movement phrases are also handled within the movement unit’s template. A movement unit is defined as a sequence of movement phrases that complement the co-expressive verbal idea (e.g. a set of idea units within a sentence). If a certain movement phrase within the annotation diagram is delayed, the delay value is reflected in the “start” attribute of the phrase. Similarly, if a movement-phrase is maintained over a certain time, the persistency attribute will reflect the duration of the maintained pose. The complete movement segment’s descriptions are formed based on the temporal relations between movement units. This process is quite similar to the transformation of movement phrases into movement units. The movement segments comprise the co-verbal motions expressing a set of ideas (e.g. a passage, paragraph, a speaker turn, etc.) into a verbally synchronized conversational behavior.

The importance of movement templates also lies in the fact that each movement template is, when reused, expressively adjustable. The end-pose (poses) it forms may vary in temporal, power, and repetitive expressive domains. When used in non-verbal behavioral generation, these templates can be adjusted to any general text and any rule (or to a combination of rules). E.g. at certain combination of words, the same template can be performed faster, it can be repeated or even generated with more/less enthusiasm (power).

6 Results

37 minutes of spontaneous informal conversation had already been annotated. Based on this annotation, several movement templates were generated for movement phases, movement phrases, movement units, and also as complete movement segments. The co-verbal movement was also evaluated by visually comparing the original sequences and those synthetically produced based on the annotated values.

The achieved performance, by using the presented annotation scheme and the reproduction capabilities of the EVA-framework, is demonstrated in Figure 7. There it can be seen how the annotation is reproduced as synthetic-motion, as generated by ECA EVA.

The sequence in Figure 7 is a movement segment of a speaker's turn. In this turn the speaker responds to the idea proposed by the co-speakers. The image sequence in figure 7 contains several frames of representative end-poses produced by a human speaker (upper sequence), and corresponding frames of end-poses performed by the synthetic agent (lower sequence). The frames were captured during the same time points. Both, the form and dynamics of each synthetic sequence matched its corresponding annotated sequence. The proposed annotation scheme and reproduced results, therefore, show high potential for reproducing more natural non-verbal behavior in ECAs.

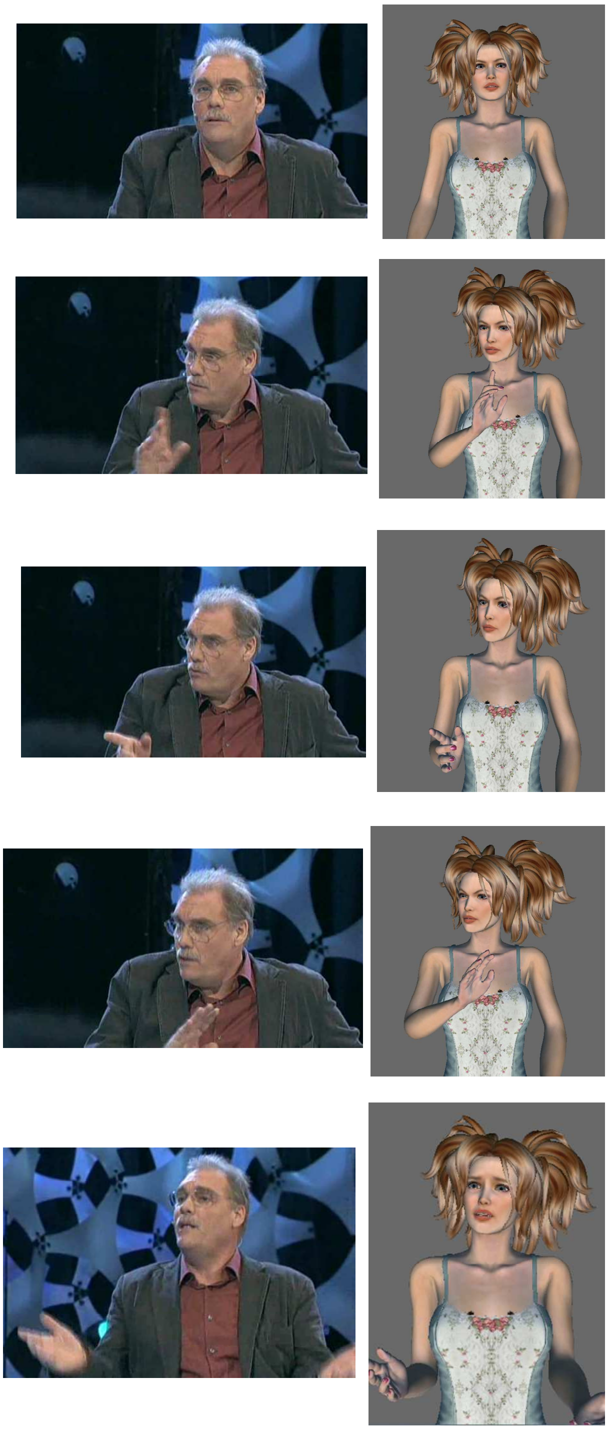


Fig. 7: Synthetic imitation of conversational behavior

7 Conclusion

This paper presented the process of annotating real-life, spontaneous non-verbal behavior and its reproduction on synthetic ECAs. This process could be used to build a high-resolution, functionally-independent movement dictionary. The discussed annotation was form-oriented and captured the expressive details of motion at high- resolutions.

The topology of the presented annotation schema extends the general hand-gesture oriented topologies in several ways. Firstly, it adopts the notion that any type of body movement can be regarded as carrying meaning. It is designed in a way that enables for posture, gesture, gaze and facial expressions to be described during a single session and under a shared time-line. The shared timeline enables the annotators to establish relations between different movement types, especially between arm-postures and hand gestures.

Secondly the schema integrates several aspects of pure form-oriented systems (e.g. FORM). The movements of body parts are not only described in the forms of shapes and poses, but also with spatial configurations of articulators propagating movement. By using movement description in the form of spatial configurations regarding its corresponding articulators, any movement lexical can define several similar movements with slightly different spatial configurations. The exact lexical entry can then be chosen randomly or influenced, based on certain contextual information (e.g. attitude, emotion, etc.).

Finally, the annotation schema also defines those word-phrases and key-phrases to be captured within a single annotation session. The word-phrases define the utterance sequence based on which movement-phrase/phase (e.g. gesture unit) can be re-produced.

Annotated movements present a small part of the dictionary that ECA should use. Therefore, we intend to annotate the entire available multimodal corpora (around 200 minutes). When necessary, additional video samples of informal dialogue will also be incorporated. The manual annotation process is, however, time consuming. As a solution the scheme already enables the re-usage of movement templates. However, in order to further optimize the process, we are developing a system for semi-automatic annotations. The spatial features of indicated end-poses are going to be approximated based on –pose- recognition and pose-tracking techniques (e.g. [30], [31]).

Acknowledgment:
Operation part financed by the European Union, European Social Fund.

References:

[1] R. M. Krauss, Y. Chen, & P. Chawla, *"Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?"*, Advances in experimental social psychology, 1996, pp. 389-450

[2] C. Navarretta, "Annotating Non-verbal Behaviours in Informal Interactions", Analysis of Verbal and Nonverbal Communication and Enactment, 2011, LNCS 6800/2011, pp. 309-315

[3] O. S. Goh & C. C. Fung, "The Design of Interactive Conversation Agents", *WSEAS Trans. Info. Sci. and App.*, Vol. 15, No. 6, 2008, pp. 901-912.

[4] S. Pita & L. Pedro, "Verbal and Non-Verbal Communication in Second Life", *Virtual Worlds and Metaverse Platforms: New Communication and Identity Paradigms*, 2011, pp. 100-116

[5] H. M. El-Bakry, A. M. Riad, & N. Mastorakis, "Adaptive User Interface for Web Applications", *In proc. of 4th WSEAS International Conference on business administration (ICBA '10)*, 2010, pp. 190-211

[6] E. André & C. Pelachaud, "Interacting with Embodied Conversational Agents", *Speech Technology*, pp. 123-149, Springer US, 2010.

[7] M. Malcangi, "Text-driven avatars based on artificial neural networks and fuzzy logic", *International journal of computers*, Vol. 4, No. 2, 2010, pp. 61-69.

[8] G. Bailly, S. Raidt & F. Elisei, "Gaze, conversational agents and face-to-face communication", *Speech Communication*, Vol. 52, No. 6, 2010, pp. 598-612

[9] E. André, E. Bevacqua, D. Heylen, R. Niewiadomski, C. Pelachaud, C. Peters, I. Poggi & M. Rehm, "Non-verbal Persuasion and Communication in an Affective Agent", *Cognitive Technologies*, Vol. 6, 2011, pp. 585-608

[10] D. Rigas & N. Gazepidis N," An empirical Investigation for the Role of Facial Expressions and Body Gestures in Interactive Environments", *In proc. of the 7th WSEAS International Conference on Applied Computer and Applied Computational Science*, 2008, pp. 672-677.

[11] N. Novielli, F. de Rosis & I. Mazzotta, "User attitude towards an embodied conversational agent: Effects of the interaction mode", *Journal of Pragmatics*, Vol. 42, No. 9, 2010, pp. 2385-2397.

[12] I. Mlakar & M. Rojc, "Recreation of Spontaneous Non-Verbal Behavior on a Synthetic Agent EVA", In proc. of the 11th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data bases (AIKED '12), 2012

[13] J. Allwood, "Dialog Coding - Function and Grammar" Gothenburg Papers in Theoretical Linguisics, 85, 2010.

[14] J. Allwood, L. Cerrato, K.Jokinen, C. Navarretta & P. Paggio, “The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena”, *Journal of Language Resources and Evaluation*, Vol. 41, No. 3, 2007, pp. 273-287.

[15] K. Bergmann, & S. Kopp, “Systematicity and Idiosyncrasy in Iconic Gesture Use: Empirical Analysis and Computational Modeling”, *Gesture in Embodied Communication and Human-Computer Interaction*, 2010, pp. 182 – 194.

[16] C. Martell, “Form: An Extensible, Kinematically-Based Gesture Annotation Scheme”, *Advances in Natural Multimodal Dialogue Systems*, Vol. 30, 2005, pp. 79-95.

[17] Q. Nguyen & M. Kipp, “Annotation of Human Gesture using 3D Skeleton Controls” *In proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[18] M. Kipp, “Anvil - A Generic Annotation Tool for Multimodal Dialogue”, *In proc. of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, 2001, pp. 1367-1370.

[19] A. Kendon, *Gesture: Visible action as utterance*, Cambridge University Press, 2004.

[20] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992.

[21] N. Tan, G. Ferré, M. Tellier, E. Cela, M. A. Morel , J. C. Martin & P. Blache, “Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversation”, *In proc. of International Conference for Language Resources and Evaluation*, 2010

[22] I. Mlakar & M. Rojc, “EVA: expressive multipart virtual agent performing gestures and emotions”. *International journal of mathematics and computers in simulation*, Vol. 5, No. 1, 2011, pp. 36-44.

[23] D. Verdonik, A. Zwitter-Vitez, M. Romih & S. Krek, “Konkordančnik za govorni korpus GOS = Concordancer for the speech corpus GOS”, *In proc. of the 13th International Multiconference Information Society - IS 2010*, volume C., 2010, pp. 12-15.

[24] X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt & M. Pantic, “A Multimodal Database for Mimicry Analysis”, *Affective Computing and Intelligent Interaction*, LNCS vol. 6974, 2011, pp. 367-376.

[25] M. Rojc & I. Mlakar, “Multilingual and Multimodal Corpus-Based Text-to-Speech System - PLATTOS-“. *Speech Technologies / Book 2*, Chapter 7, InTech, 2011.

[26] D. McNeill, “Gesture and Thought”, *University of Chicago Press*, 2005.

[27] P. Bull, “Posture and gesture”, *International series in experimental social psychology*, Vol. 16, 1987.

[28] S. Vosinakis & T. Panayiotopoulos. “Design and Implementation of Synthetic Humans for Virtual Environments and Simulation Systems”, *Advances in Signal Processing and Computer Technologies*, 2001, pp. 315 - 320.

[29] M. Pantic, M.F. Valstar, R. Rademaker & L. Maat, “*Web-Based Database for Facial Expression Analysis*”. *In proc. of Multimedia '05*, 2005, pp. 317-321.

[30] Y. Minghai, Q. Xinyu, G. Qinlong, R. Taotao & L. Zhongwang, “Online PCA with adaptive subspace method for real-time hand gesture learning and recognition”, *WSEAS Transactions on Computers*, Vol. 9, No. 6, 2010, pp. 583-592.

[31] G. Peng, A. Weiss, A. O. Balan & M. J. Black, “Estimating human shape and pose from a single image”, 2010, pp. 1381 – 1388.