Recreation of spontaneous non-verbal behavior on a synthetic agent
EVA

IZIDOR MLAKAR, MATEJ ROJC
1Roboti c.s. d.o.o, 2Faculty of Electrical Engineering and Computer Science, University of Maribor
1Tržaška cesta 23, 2Smetanova ulica 17
SLOVENIA
izidor.mlakar@revolutionary-robotics.com, matej.rojc@uni-mb.si

Abstract: This paper presents a novel process of transferring the human-generated communicative behavior onto an embodied conversational agent. The aim of our work is to build a high-resolution motion dictionary based on empirical analysis of non-verbal behavior performed in multi-speaker informal dialogues. The verbal and non-verbal behavior is recreated by using this motion dictionary and on pure, unprocessed text. It involves animating hand and arm gestures, head movement and gaze, and facial expressions with lip-sync. The generated non-verbal behavior is speech driven and in synchrony with verbal information passed in form of synthetically generated speech.

Key-Words: multimodal annotation, non-verbal behavior, human-speaker imitation, motion dictionary

1 Introduction
The importance of embodied conversational agents (ECAs) within advanced user-interfaces has already been proven by several researches and research programs. ECAs can be used can be used within a wide range of application scenarios, e.g. games [1], educational applications [2], and commercial interfaces [3]. Although widely used, non-verbal behavior generated by ECAs still lacks naturalness [4]. The “wooden” appearance is even more evident when motion is reproduced based on general text. The major disadvantage of motion directly generated based on general text is that most of the contextual information is missing (acoustic signal information, emotion, speaker-listener relation, etc.). In [5], Malcangi tries to overcome this problem by developing a rule-based, text-to-viseme synthesis system. It employs artificial neural networks and fuzzy logic to generate phoneme and viseme information that drives facial movements. Although successful, his approach is limited only to facial region and still lack information necessary to drive other body regions (e.g. head, arms, and hands).

Several annotation schemas and annotation tools emerged in order to provide an insight 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 explore human mechanism used for managing communication [6]. The MUMIN coding scheme [7], for instance, focuses on the annotation of three communicative functions: the feedback, turn-management, and sequencing functions. Bergmann & Kopp in [8] study correlations between contextual factors (referent features, discourse) and gesture features. The correlations are then classified as systematic (shared among speakers), or idiosyncratic (inter-individually different). Empirical investigations at functional level offer insights into motives and a correlation between verbal and non-verbal behavior. However, these annotations only coarsely describe the form and the dynamics of motion. In most cases the annotations at functional level use gestural lexicons and semantic classes in order to describe/code the motion being produced.

Form-oriented systems were developed in order to capture more-detailed information on the structure and the dynamics of the produced motion. The representative form-oriented system is FORM [9]. The temporal and spatial dynamics of motion are in the FORM encoded based on articulators that are propagating the motion. Due to the level of detail and the complexity that FORM describes, the motion coding is highly time consuming. As a solution, the authors in [10] compensate for FORM’s complexity by introducing a 3D pose editor integrated into ANVIL annotation tool [11]. Their concept allows gestures to be fine-tuned based on different-end poses and movement phases [12], and interpolated, as gesture phrases and units [13], into gradual synthetic-motion.
To sum up, the existing annotation approaches and schemas provide different levels of understanding non-verbal behavior and its form. The functional annotations provide detailed data on correlations between 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 and are capturing part of the functional, and part of the form-oriented features.

The primary goal of our work is to synthesize non-verbal human-like behavior based on pure text sequences, and by using ECA EVA[14]. In order to do that a concept of describing and transforming annotated behavior into EVA’s expressive motion templates is presented in the paper. The novel concept is based on manually annotating informal multi-speaker dialogues in a form-oriented annotation schema. The schema captures expressive features of moving body parts at high resolution and can be directly transformed into co-verbal motion generated by ECA EVA.

2 Capturing movement at high-resolution

In order to be able to reproduce co-verbal motion we suggest a 2-staged approach, presented in figure 1. The approach consists of empirical analysis of spontaneous informal multi-speaker dialog and generation of EVA-Script-based procedural animation in form of EVA EVENTS. These events can be directly animated on embodied conversational agent EVA.

The empirical analysis of motion’s expressive details is performed manually and involves: (a) observing multimodal corpora and (b) coding observed data based on an annotation schema in an annotation tool

The empirical analysis results, at the end, in movement lexicon. It stores the different motions that ECA can reproduce, and rules that define in what context those movements should be generated. In the context of text driven animation’s these rules include morphological relations, syntactic relations, and semantic relations. If, however, the TTS-system’s output driving the animation is able to generate e.g. prosody on general text, the communicative rules regarding pitch and amplitude can also be included as additional information.

The generation of EVA-Script-based procedural animation is performed within the PLATTOS TTS-system [15]. It generates EVA-Script-based animation descriptions based on synthetic behavior generation rules (Rules module in Figure 1), and general text as input. In the following sections basic building blocks of the proposed concept will be presented in more detail.

2.1 Annotation Corpora and Annotation tool

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 [16].
The annotation corpora [17] used was based on informal dialog with 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 in all four sessions, whereas the other participants were different in each talk-show. In each session at least 3 participants were actively contributing to the communicative dialogue.

The form-oriented annotation was performed in ANVIL [11]. The primary advantage of ANVIL is its rich, structurally oriented tier-based system. The annotators can specify the different attributes of the tiers, and establish different relations between the tiers. In regard to the work presented in this paper, we used ANVIL to perform form-oriented analysis of only those dialog acts that had relevance to the observed conversation (functionally, semantically, syntactically, or morphologically synchronized verbal and non-verbal behavior).

2.2 Annotation Schema
The proposed annotation scheme consists of a series of main tracks based on body-parts. Body-parts merge articulators within the following body-regions: head, face, right and left arms, right and left hands. Each body-part is designated with a separate tier.

As shown in Figure 1, the movement of the observed body-part part is described by movement phrase, movement phase, and the articulators propagating the observed movement. The articulators are in terms of [14] control units that model the end-pose.

The movement phase describes the power and the repetitive expressive-details, and joins movement phases into a lexical group of sequential motions. Movement phrases can be expressively reused during annotation and animation processes. 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 general text (e.g. what movement phrases ECA displays during different word sequences).

The movement phase describes the temporal features of those articulators influencing the observed movement. In connection with articulators it defines temporal borders of motion propagation (time limit during which, the motion must reach its end-pose). Each articulator can model the end-pose only within these temporal borders. However, the modeling of articulator can be delayed, or can even end before the maximally allowed duration is reached. If the modeling of articulator finishes before the phase’s “left” temporal border, the state of the articulator is maintained until the phase finishes. The pose identifies an abstract description (e.g. closed-fist, hand-wave) of the movement being propagated. It can be used for the classification of movement and identification of similar movement.

Movement phases and phrases define all expressive dimensions, except the spatial dimensions. The spatial dimension is defined by the spatial features of articulators that model an end-pose at the borders of each movement phase. Hand and arm movement is described in the form of 3D rotational vector (HPR), head movement involves HPR and an optional attribute for gazing, whereas the facial expressions are defined by coding the level of expression.

2.3 Annotation process – capturing expressive details of non-verbal behavior
Most of the information, carried by co-verbal movement, is presented through stroke and post stroke hold movement phases. However, the events prior and post to stroke 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 annotation of corpora is performed in ANVIL.

As already mentioned, the movement within motion segment is observed separately for each body-part. The coding is based on end-poses at the borders of movement phases. The stroke phase is, in general, defined based on the significance of motion. The post stroke hold phase is defined on those segments, where motion is performed after the stroke phase and it also remains relatively static. The retraction phase labels those motion segments that drive the observed body part into a relaxed (neutral) state. Finally, the preparation phase denotes those segments that drive the observed body part into a stroke phase.

When movement phases and phrases are clearly defined, separately for each body-part, the motion’s temporal, repetitive and fluidic expressive features are captured. The annotators now have to capture
the spatial features of the end-poses. Figure 2 presents the interface used to visually adjust the ECA’s pose to the pose displayed in the dialog.

![Observed key pose](image1)

![Panel for selecting articulators](image2)

![Articulated synthetic model](image3)

**Fig. 2:** Interface for capturing spatial features of key-poses

Further, Figure 2 also presents the observed key-pose (end-pose), the approximation of the pose on the articulated model, and the EVA-Framework’s interface for the manual modeling of end-poses. In order to describe the key-pose in high-resolution, the annotator selects the articulator from the list of articulators provided by the panel (for selecting articulators), and adjust its values in the placer panel. When the modeling of the key-pose is finished, the annotator has to transcode those values into the annotation schema.

E.g. when annotating head movement, the annotators can specify eye-gazing using eye articulators. In addition, gazing can be described as a general direction the ECA is looking at, or even as automatic (gazing is automatically adjusted to the neck articulator). 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 [18]. The level of exposure ranges from 1-10. When encoding facial expressions, the annotators code the options of the articulator, and the associated level of the exposure.

**3 Transforming annotation into synthetic behavior - imitating the speaker**

EVA framework and PLATTOS TTS system support the concepts of expressive motion, and non-verbal behavior generation on 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 units, and also into complete motion segments. The correlation between annotation and EVA-Script is shown in Figure 3.

![Annotated data](image4)

![Motion template](image5)

**Fig. 3:** Correlation between annotation and EVA-Script

Figure 3 demonstrates 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 movement phrases are defined by those phases surrounding the stroke phase. In general, each movement phrase contains at least a stroke phase. Other phases are optional.

Movement phrase, in terms of EVA-Script, defines a motion template that ECA can animate. The overall duration of the template is denoted by the durations of the preparation, stroke, and post stroke hold movement phases. The retraction phase then defines the overall duration of the subsiding. If there is no retraction phase indicated, the retraction time is denoted as half of the last stroke’s duration.

Movement phases group articulators that are propagating motion of the observed body part into sequences of parallel motion (<sequence><parallel> blocks in EVA-Script). In addition, movement phases define the key-frame interpolation of animated motion [14]. 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 only stroke, or stroke, post 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 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.

As shown in Figure 3, the movement phases are merged into movement phrases as sequential motion. Each movement phrase is also labeled. This label is then reflected in the name of the phrase. Movement units are then formed by merging and correlating movement phrases, separately for each body part. In the context of EVA scripts, a movement unit is formed by hierarchical and temporal correlation of sequential movement phrases.

Within the movement unit’s template, the persistency and delay expressive features of the movement phrases are also handled. 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 segments descriptions are formed based on the temporal relations between movement units. The process is quite similar to the transformation of movement phrases into movement units. The movement segments merge motion of all body-parts into a verbally synchronized non-verbal behaviour.

The importance of movement templates also lies in the fact that each movement template is, when reused, expressively adjustable. The end-pose (pose’s) it forms may vary in temporal, power, and repetitive expressive domains. When used in non-verbal behaviour generation, these templates can be adjusted to any general text and any rule. 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).

4 Results
37 minutes of spontaneous informal conversation have already been annotated. Based on this annotation, several movement templates have been generated for movement phrases, movement units, and also as complete movement segments. The motion segments were 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 4. There it can be seen how the annotation is reproduced as synthetic motion as generated by ECA EVA.

Fig. 4: Imitating a speaker

The sequence in Figure 4 contains several frames of a motion segment performed by a human speaker (upper sequence), and corresponding frames of end-poses performed by the synthetic agent (lower sequence). The frames were captured at the same time points. Both, the form and dynamics of the 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.

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

Annotated movement presents 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.
Manual annotation process is, however, time consuming. As a solution the scheme already enables the re-use 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 the pose recognition and the pose tracking techniques (e.g. [19], [20]).

Acknowledgements
Operation part financed by the European Union, European Social Fund.

References: