

Multimodal Distinctive Behavior for Expressive Embodied Conversational Agents

Maurizio Mancini

DISSERTATION.COM



Boca Raton

Multimodal Distinctive Behavior for Expressive Embodied Conversational Agents

Copyright © 2008 Maurizio Mancini

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the publisher.

Dissertation.com
Boca Raton, Florida
USA • 2008

ISBN-10: 1-59942-699-4
ISBN-13: 978-1-59942-699-0

Contents

1	Introduction	2
1.1	Context	2
1.2	Problem definition	3
1.3	Objectives	5
1.4	Contributions	6
1.5	Thesis outline	7
2	Background	9
2.1	Nonverbal communication	9
2.2	Nonverbal signals	11
2.2.1	Gestures	11
2.2.2	Facial expressions	13
2.2.3	Torso movements	15
2.2.4	Head movements	16
2.3	Communicative intention	17
2.3.1	Information about the world	18
2.3.2	Information about the speaker’s mind	18
2.4	Embodied Conversational Agents (ECAs)	19
2.4.1	ECA general framework	20
2.4.2	Conclusion	20
I	System description	22
3	Introducing a system for distinctive behavior in ECAs	23
3.1	Introduction	23

3.2	Related work: the SAIBA project	23
3.2.1	The BML Language	24
3.3	Distinctive behavior system for the Greta ECA	26
3.3.1	Related work: the Greta ECA	27
3.3.2	Distinctive behavior system	27
3.4	Conclusion	29
4	Representation languages	30
4.1	Introduction	30
4.2	FML-APML	30
4.2.1	Related work: APML	30
4.2.2	FML-APML overview	32
4.2.3	Tags: common attributes and synchronization	33
4.2.4	FML-APML importance attribute	35
4.2.5	Emotion tag	36
4.2.6	World tag	37
4.3	BML extensions for the Greta agent	38
4.4	Conclusion	40
5	Behavior Quality Computation (BQC)	42
5.1	Introduction	42
5.2	Modality preference	43
5.3	Expressivity of behavior	44
5.4	Baseline and Dynamicline	47
5.4.1	Baseline and Dynamicline representation language	48
5.5	Dynamicline computation	50
5.5.1	Behavior qualifiers	51
5.6	Dynamicline computation example	54
5.7	Behavior qualifiers for the Greta ECA	56
5.8	Conclusion	58
6	Multimodal Signal Selection (MSS)	59
6.1	Introduction	59
6.2	Overview	60

6.3	Behavior set representation	61
6.3.1	Core signals	64
6.3.2	Implication rules	64
6.4	Multimodal Signal Selection	66
6.4.1	Steps 0, 1 and 2: Parse, Pick behavior set and Apply rules	68
6.4.2	Step 3: Selection	70
6.4.3	Step 4: Assign <i>multiplicity</i>	73
6.4.4	Step 5: Apply preference	74
6.4.5	Discussion	74
6.5	Behavior sets for the Greta ECA	75
6.6	A MSS working example	76
6.7	Conclusion	77
7	FML-APML Engine	79
7.1	Introduction	79
7.2	Overview	80
7.3	Subsystem 1: temporalization and sorting	81
7.3.1	Temporalization of the C.I.	82
7.3.2	Sorting of the C.I.	83
7.4	Subsystem 2: FML-APML Dynamicline computation	84
7.5	Subsystem 3: from C.I. to synchronized signals	86
7.6	Example	89
7.6.1	Input	89
7.6.2	Tags temporalization and sorting	90
7.6.3	Dynamicline computation	91
7.6.4	From C.I. to synchronized signals	93
7.7	Conclusion	96
8	Evaluation of the system	97
8.1	Introduction	97
8.2	Objective evaluation	98
8.2.1	Setup	98
8.2.2	Execution	100

8.2.3	Results and discussion	100
8.3	Subjective evaluation	112
8.3.1	Setup	112
8.3.2	Participants	112
8.3.3	Procedure	113
8.3.4	Results: evaluation of activity	114
8.3.5	Results: evaluation of expressivity	116
8.3.6	Discussion	117
8.4	Conclusion	118
9	Related work	119
9.1	Introduction	119
9.2	Variability depending on emotion and personality	120
9.3	Variability depending on the behavior repertoire	121
9.3.1	Preference in the use of modalities	121
9.3.2	Behavior repertoires and styles	122
9.4	Variability depending on movement expressivity	123
9.5	Summary and discussion	124
II	System Applications	126
10	Introduction	127
10.1	Greta and Psyclone	127
11	Application: GretaMusic	130
11.1	Introduction	130
11.2	State of the art in music in HCI	131
11.3	Expressivity in voice and music	132
11.4	The GretaMusic system	132
11.4.1	CUEX	134
11.4.2	Mapping acoustic parameters to expressivity parameters	134
11.4.3	From multiple emotional intentions to facial expression	135
11.4.4	Generating animation	136
11.5	Testing the mapping between emotional intention and expressivity parameters	137

11.5.1	Setup: musical stimuli	138
11.5.2	Setup: test interface	138
11.5.3	Participants	139
11.5.4	Procedure	140
11.5.5	Results	141
11.5.6	Discussion	141
11.6	Conclusion	142
12	Application 2: from video analysis to behavior generation	144
12.1	Introduction	144
12.2	State of the art	145
12.3	Application description	146
12.4	Automated extraction of video parameters	148
12.5	Expressivity mapping	149
12.6	Evaluation study	151
12.6.1	Experiment 1	154
12.6.2	Experiment 2	155
12.6.3	Discussion	157
12.7	Conclusion	159
13	Conclusion and future	160
13.1	Summary of the work	160
13.2	Future work	162
13.2.1	Short-term	162
13.2.2	Long-term	163
III	Appendices	164
A	BML Editor	165
A.1	Tool description	165
A.1.1	Applications	166
A.1.2	Usage	166
A.1.3	Example	167

B	Greta’s Behavior Qualifiers and Behavior Sets	170
B.1	Behavior Qualifiers	170
B.2	Behavior Sets	171
C	File formats	173
C.1	FML-APML	173
C.2	BML	174
C.3	BQC and MSS	176
C.3.1	Common definitions	176
C.3.2	Baseline and Dynamicline	177
C.3.3	Behavior qualifiers (BQC)	178
C.3.4	Behavior sets (MSS)	179
D	Publications	181

Chapter 1

Introduction

1.1 Context

Embodied Conversational Agents (ECAs) are a new kind of Human-Computer Interface that are embodied and have conversational skills [24]. They exhibit a human-like aspect, in both appearance and behavior, capable of exhibiting conversational functions, emotional states, personality traits, etc. People in general tend to deal with computers as if they were humans, as stated by Reeves and Nass in their book “The Media Equation” [83]. ECAs promise to increase the quality of communication between humans and computers, as they are designed to communicate and interact in a human-like manner.

The first systems implementing ECAs aimed solely at reproducing the basic skills of human-human conversation: ECAs had schematic bodies, exhibited monotonic speech, and produced few and simple gestures and facial expressions. In recent years, developers focused on refining ECAs by modeling key aspects of human-human interaction. Human communication involves verbal and nonverbal behaviors: the words we utter represent the verbal part of communication, while the nonverbal part is constituted

by a very large set of behaviors, going from speech intensity and intonation to facial expressions, hand/arm gestures, head and torso movements, posture changes, etc [4].

While communicating with others, our goal is to (both voluntarily and involuntarily) “transmit” some information from our mind to the others’ minds. Our emotional and mental states (our beliefs and goals) can be communicated with a large variety of nonverbal behaviors [2], which influence the interaction with other people. For example, while greeting someone we can simply say “Hello” and accompany it by raising our hand open, stretching both arms sideways, smiling a little bit, etc. This great variability in performing nonverbal behaviors is determined by various concurrent factors: our personality traits, our emotional state, our disposition/relation with other person/event/object [2], some physical and social constraints (for example, greeting someone in a church is different from waving at someone in a crowded place), our origins [68], our social role, our idiosyncratic, innate “way” of behaving, etc. To become more credible and usable, ECAs have been endowed with all the above characteristics [83] [68] of human nonverbal behavior. Many researchers have attempted to model some of these aspects in ECAs, considering, for example, the influences on behavior induced by the agent’s emotional state, personality traits and individualized repertoire of gestures and facial expressions [1] [7] [31] [56] [69] [71] [79] [88]. All of these works have obtained very positive results, even if considering all these aspects together is still a very challenging task. For example, how can we model an agent that is extroverted, sad and has a general tendency to raise its eyebrows while speaking? It is still not clear how we can consider these three aspects together.

1.2 Problem definition

In his book about bodily communication, Argyle [4] states that there is an underlying tendency which is constantly present in each person’s behavior: for example people that look more tend to do so in most situations, that is, there is a certain amount of consistency with the person’s general tendency. Gallaher [43] found consistencies in the way people behave. She conducted evaluation studies in which subjects’ behavior style was evaluated by friends, and by self-evaluation. In a first study, many characteristics of behavior were evaluated: tendency to use body, face, head, gestures; qualities of movement, like fast-slow, small-large, smooth-jerky, etc. The person’s behavior tendency was shown to be an innate individual characteristic that the author claimed to be a personality trait. In the second study she investigated the consistency of a person’s behavior across time and situations. Results demonstrated this consistency: people that are quick when writing have a tendency to be

quick while eating; if a person produces wide gestures then she also walks with large steps. Energy of movements is also an enduring characteristic, constant over time. The paper by Wallbott and Scherer in [100] illustrates a study on actors' body movements during the expression of several emotions. A group of people judged the actors' behaviors and annotated them. In the study, the authors found that the way actors portrayed emotional states also seemed to be actor dependent, that is, it depended on the actor's personal way of expressing those emotions. Some behavior characteristics seemed also independent from the emotion: for example the number of head movements and total activity. Finally, some actors seemed capable of showing some emotions better than others, just because their behavior style was similar to their expression for a certain emotion. Similar results have been proposed by Gross et al. [45] who found that the capacity of people in expressing their emotions depends, among other things, on the dispositional expressivity of a person. They showed that, for example, low-expressivity individuals tend to inhibit negative emotions, while high-expressivity individuals do not. This behavior tendency is maintained also across large time spans.

All the above studies suggest that the behavior of a person does not depend only on what the person is communicating, that is, their *communicative intention*. But it also depends on the person's general behavior tendencies, their personal way of behaving. To investigate how this influence can be replicated in ECAs, we propose the implementation of ECAs exhibiting *distinctive behavior*.

We say that two ECAs behave in a *distinctive* way if and only if:

- *They behave differently*, that is, given the same communicative intention (for example, the agent is *greeting* the user) they exhibit at least one of the following kinds of behavior differences:
 - in the types and number of signals produced; for example, while saying “Hello”, an agent will perform a head nod, another one will raise a hand;
 - in the quality of movement of the produced signals; for example, while saying “Hello”, even if two agents choose to perform an head nod, they will perform it with, for example, different speed/amplitude/acceleration;
- *They maintain their behavior tendencies across time and across situations*: that is, if we observe the way in which two differently defined distinctive ECAs communicate the same communicative intention (the agent is *greeting* the user) then we are able to distinguish between these two agents also in other situations (for example while describing an object, or giving directions, etc.).

1.3 Objectives

In the previous Section we propose to implement ECAs exhibiting *distinctive* behavior. Let us consider the conceptual diagram in Figure 1.1.

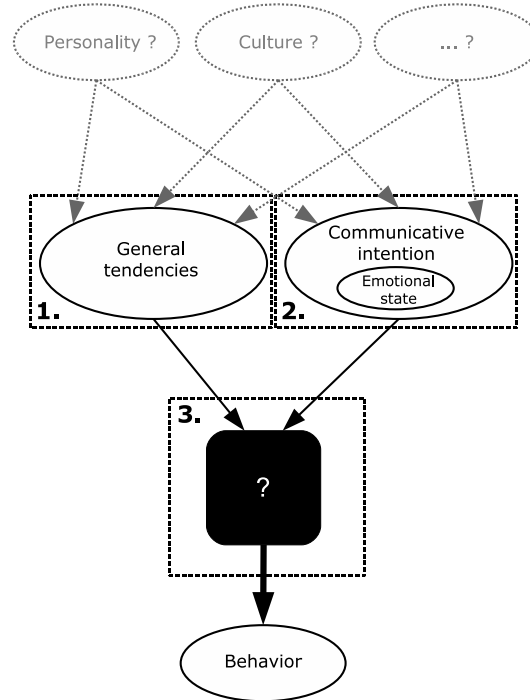


Figure 1.1: A conceptual diagram of the system implement in this thesis.

In the diagram we represent the connection between the agent's *general tendencies*, *communicative intention* and *behavior*. The black box represents a system that, receiving as input the first two kinds of data, combines them in order to determine the agent's behavior.

In this thesis, we aim to produce an implementation of a system in which the computation of the agent's behavior follows the definition of distinctiveness reported in Section 1.2. The diagram of Figure 1.1 is split in three areas, which correspond to the three main goals of this thesis:

1. *Modeling the agent's general behavior tendencies*: we aim to define the agent's general behavior tendencies. This will allow us to model, for example, an agent having the tendency to use more gestures, or facial expressions, etc.
2. *Modeling the communicative intention influences*: we seek to create a model for defining the nonverbal behaviors (e.g., gestures, facial expressions, torso movements, etc.) that could be used by an agent to communicate an intention, such as the intention to describe something, or

to give directions, etc. In our system, the agent’s emotional states are considered to be one of the communicative intention classes the agent’s can communicate.

3. *Implementing a distinctive behavior generation system*: to implement distinctive behavior in ECAs, we have to describe how the agent’s actual communicative intention may influence the way in which the agent produces nonverbal behaviors. We have to propose a method to combine this information with the agent’s general behavior tendencies.

Two additional goals are pursued in this thesis:

4. *Evaluating our distinctive behavior generation system*: we seek to evaluate the quality of our model of distinctive ECAs by performing perception tests.
5. *Obtaining an extendible/extensible system*: we aim to create a system which is extensible and flexible. Our system will allow the addition of factors influencing the agent’s behavior, for example (as reported in Figure 1.1) the agent’s personality, culture, etc.

1.4 Contributions

The contribution of this thesis is to define, implement and test ECAs exhibiting *distinctive* behavior:

- We define a model for the ECA’s general behavior tendencies. We represent it with two concepts: the agent’s *preference* in using each modality (e.g., an agent may prefer to use mainly its face, or gestures); the agent’s *expressivity* of movement, that is, a set of parameters that influence the amplitude, speed, fluidity, energy and repetitivity of the nonverbal signals produced by the agent. We define an XML-based representation language that allows us to define the agent’s behavior tendencies in a *global*, static way: “agent A has the global tendency to behave in the way W ”. We refer to the agent’s global behavior tendencies, defined with such a language, as the *Baseline*.
- We define the influence of the agent’s communicative intention on the ECA’s behavior tendencies. We do that by defining *behavior qualifiers*: these allow us to model the *modulation* of the agent’s Baseline by the agent’s *local* behavior tendencies: “agent A , having the tendency to behave in the way W and with the communicative intention C , has the local tendency to behave in the way W_C ”. We refer to the agent’s local behavior tendencies as the *Dynamicline*. Again, we use an XML-based language allowing us to add or modify these behavior qualifiers.

- We propose a representation language to describe the mapping between communicative intentions and nonverbal behaviors. A certain communicative intention can be communicated through several combinations of nonverbal behaviors. For example, to communicate the affirmation “yes”, we can produce a head nod, raise our thumb up, or perform both behaviors at the same time. Our language allows us to define all the possible combinations of signals representing a given communicative intention, called *behavior sets*. Constraints can be defined over the produced signals. By using an XML-based language, one can easily modify or extend the behavior sets associated to each communicative intention.
- We propose a system that, considering an agent with its Baseline and the intention it aims to communicate, computes the agent’s local behavior tendencies (using the behavior qualifiers defined above) and determines the nonverbal behaviors the agent has to produce, according to the communication behavior sets.
- We evaluate the quality of our model by performing evaluation studies on the output of our system. We perform two kinds of evaluation:
 - from an *objective* point of view: we look at the output of the system, to establish if it reflects the global and local tendencies of the agent;
 - from a *subjective* point of view: we ask human participants to observe and rate the agent’s behavior.

1.5 Thesis outline

In the next Chapter we give an overview of the background concepts we refer to in the thesis. We propose a definition of Embodied Conversational Agents and of nonverbal communication: we distinguish between the communicative intention, that is, what we aim to communicate, and the nonverbal signals, that is, the movements and gestures we produce with our body in conveying a particular intention.

We split the rest of the thesis in two parts. The first part is devoted to the description of our system for creating distinctive ECAs. In Chapter 3 we give an overview of the system by determining where our work is located in the general framework of an agent interacting with humans or other agents. Our work is related to the process of deriving the agent’s behavior generation based on its communicative intention. We implement our system in the framework of the Greta ECA. In Chapter

4 we describe the languages we defined to model the agent's communicative intention and nonverbal behavior. In Chapter 5 we present our definition of the agent's *global* behavior tendencies: an agent has a certain degree of preference in using each of its communicative modalities (i.e., face, head, gestures, gaze, torso) and a certain qualitative way of performing nonverbal behaviors (i.e., speed, amplitude, energy, fluidity, repetitivity, of movement). We explain how we modulate these general tendencies depending on the agent's communicative intention, to obtain the agent's *local* behavior tendencies. We introduce a notation for both of these concepts and the process of computing local from global tendencies. In Chapter 6 we focus on the generation of multimodal signals starting from the agent's local behavior tendencies. Again, we introduce a notation to represent this correspondence and we explain how we use it to perform the computation of the agent's behavior. In Chapter 7 we describe our system for the generation of distinctive behaviors in ECAs by putting together the concepts and modules presented in Chapters 5 and 6. In Chapter 8 we present a study we conducted to evaluate our system. We have performed two kinds of evaluation: an objective evaluation, in which we check if the output of our system is objectively the one we expected (that is, for example, if the Baseline of the agent really influences the multimodal signals chosen by our system); and a subjective evaluation, in which we have conducted a perceptual test by asking participants to look at animations of the Greta agent and to evaluate them on the basis of the multimodal signals produced by the agent. We conclude this part of the thesis with Chapter 9 in which we present an overview of other ECA systems exhibiting variable behaviors: some of them vary the agent's behavior by modeling its emotional state or personality profile; others assign a specific repertoire of nonverbal signals; others allow one to vary the performed nonverbal signals by changing their quality of execution.

In the second part of the thesis we describe two application scenarios for our system: in the first scenario the agent's behavior is determined by a musical performance provided as input; in the second one the agent mimics the quality of movements performed by a human while being filmed with a camera. In both cases we include a description of the system implementation specific for that scenario and an evaluation study.

Chapter 2

Background

2.1 Nonverbal communication

Nonverbal communication is an essential element of human-human communication, together with the verbal message. By nonverbal communication, we refer to all the communicative signals, with the exception of spoken words, that are produced by actions such as facial expressions, hand/arm gestures, posture shifts, and so on, and also through voice by variations in volume, pitch, speed of speech.

The general paradigm of nonverbal communication, as explained by Argyle [4], can be described as follows. Consider 2 entities, A and B. A's *intention*, or goal, is to communicate his state (e.g. his emotional state) to B. Nonverbal communication consists in encoding A's intention into nonverbal behaviors, or signals, which may be decoded by B, not necessarily as intended by A. Following the definition of Duncan [34], nonverbal communication consists in the *transmission of a mental representation from one person to another via bodily gestures*. In Figure 2.1 we represent the sender (A) on the left side of the diagram and the receiver (B) on the right side. In the diagram we highlight

the process of transmission of the sender's intention to the receiver, who interprets it. Constantly, as highlighted by the arrow going back from the receiver to the sender, nonverbal communication always need a feedback/reaction from the receiver in order to ensure that the information sent by the sender has been successfully received and/or understood by the receiver.

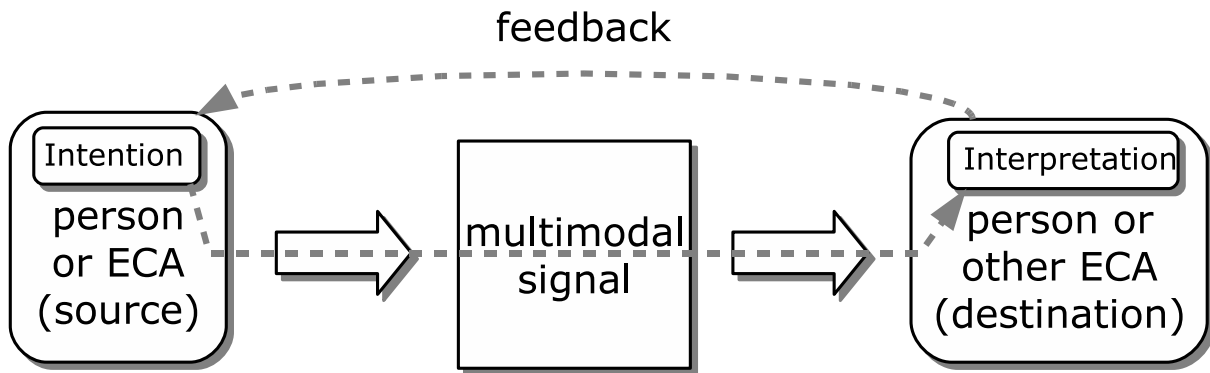


Figure 2.1: The paradigm of nonverbal communication [4].

The relation between communicative intentions and nonverbal signals is in general a many-to-many relation [2]:

- a given meaning can be communicated in a large variety of ways through nonverbal signals (to say a simple *ok* I can nod, show a thumb up, etc.);
- the same signal can serve many intentions (for example raising the palm of the hand may signify *hello*, or *stop*, etc.).

As an example of the complexity of the problem, let us consider a scenario proposed by Allwood in [2]: suppose we produce a simple verbal *yes*. We can accompany it in many different nonverbal ways, for example by nodding with the head. In this case, the global communicated meaning is one of *affirmation*. If, instead of nodding, we choose to raise our eyebrows, the communicated meaning could be *surprise*, or maybe *doubt*.

The way in which we choose some signals instead of others to convey a given communicative intention is variable, and depends on factors like: the relation/disposition between the information sender and receiver; the sender's and receiver's personalities and emotional states, their culture, social role and idiosyncratic habits; the environmental conditions.

2.2 Nonverbal signals

In this thesis we consider four categories of signals, identified by the following modalities: gesture, face, torso and head. For each modality we will now briefly describe: how the related signals are physically (i.e. how the involved parts of the body are configured) and temporally performed; which/how communicative intentions can be communicated through that modality (we will talk in this case of the *communicative function* of a certain class of signals, e.g. gestures).

2.2.1 Gestures

In this Section we provide an overview of communicative gestures. These are the nonverbal signals produced with the hand/arm accompanying speech. The following description mainly refers to the works of McNeill [67] and Kendon [54]. Gestures are produced continuously while we speak. We use gestures to represent objects as well as abstract concepts, to indicate concrete places, to give emphasis to the most important part of the discourse, and so on. While we describe objects for example, we gesticulate for approximately three quarters of the total speaking time. In this thesis we will focus on those gestures which are produced along with speech with the goal of communicating some intention in the speaker's mind. Of course there are categories of gestures, other than the communicative ones, which are used in other situations: pantomimes for example, in which gesture almost completely replace speech, and sign languages, used by deaf people to communicate. Following McNeill's theory, there are 4 major types of gestures occurring with speech:

- **Iconics:** these are related to the semantic content of the speech. The shape of the hands/arms and/or their movement trajectories are used to visualize some concrete characteristic of the uttered speech. For example, while referring to a box which was on the table, we can depict a square shape with our hands or fingers (by moving the index fingers along a squared path), while we pronounce the sentence "did you see the cards box?".
- **Metaphorics:** they are similar to iconics, but they are used to describe some abstract concept by a movement or a particular hand/arm configuration. For example, if we refer to a person to say that he/she is the *owner* of a shop, we can underline the concept of *owning* something with the following gesture: arms along the body, with a flexion of ninety degrees at the elbows (forearms toward the listener); palms up, hands opened, we bend the fingers while pronouncing the word *owner*. In this example, we want to give the idea of *owning* a shop by grasping an invisible object (the shop) with our hands and at the same time our hands are ideally placed

below that object, to suggest the idea that we *support/own/take care of* it.

- Beats: they do not have a form related to speech, and they are repeated rhythmically with usually short and quick movements along with the speech accents. Typical beats are short movements of the hand or fingers up and down. These gestures are important because they help the speaker to mark some characteristics of the spoken sentences. For example, beats are produced when discussing new or important themes, or to emphasize the uttered speech.
- Deictics: they are pointing gestures. They can point to concrete objects or highlight positions in the world, but very often they refer to abstract concepts and parts of the discourse. For example pointing down is usually to refer to the topic of the discourse. Deictics can be differentiated also by the hand shape used to show objects: an extended finger is used to point to a single object; an open hand is used to indicate multiple objects; a hand shape in which the middle two fingers are flexed with the index and little finger extended can be used to give directions or show a route.

The timing of gesture

The execution of a gesture can be divided into temporal segments or *phases* which are temporally linked to the uttered speech. In Figure 2.2 we represent the wrist position in time during the execution of a gesture, in which the hand is lifted up and moved down.

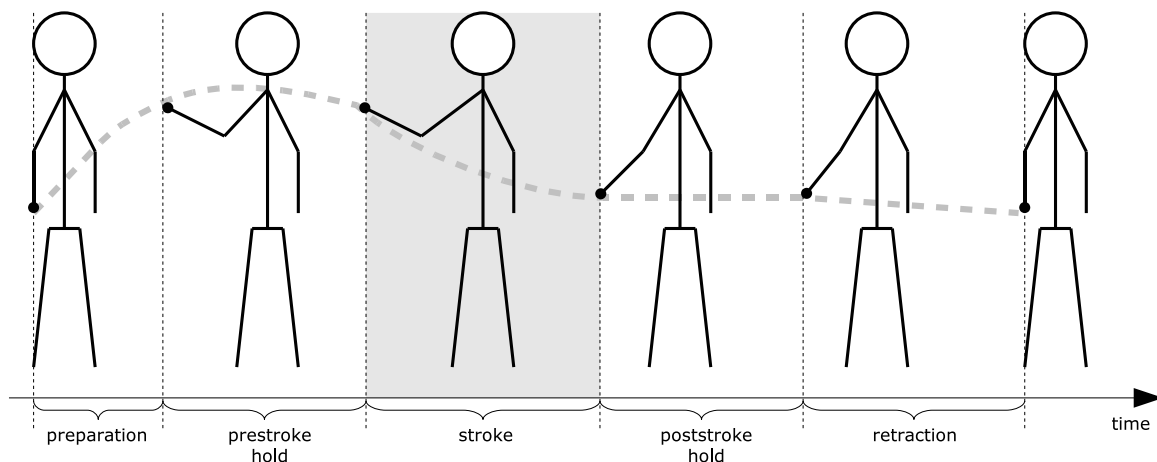


Figure 2.2: Temporal segmentation of the gesture execution.

Let us use this example to illustrate what happens while we perform a gesture. The hand moves from its initial rest position to reach the position in space in which the gesture will be produced (*preparation* phase); movement stops for a short while (*pre-stroke hold* phase); the gesture movement is performed (*stroke* phase); movement stops for a short while (*post-stroke hold* phase); the hands go back to their initial rest position, or to another rest position (*retraction* phase). The stroke phase represents the moment in which the “expression” of the gesture is accomplished. Preparation and retraction are optional phases, but the stroke must always be present during the execution of a gesture. The two hold phases (pre-stroke and post-stroke hold) have a different role in the execution of a gesture. The pre-stroke hold is used to synchronize the stroke with speech. During preparation, the gesture is set up, the hand arrives to the position in space preceding the stroke and the hold phase allows one to perform the stroke exactly at the right time (see below for discussion on the stroke synchronization). The post-stroke hold allows one to extend the meaning conveyed by the stroke for the duration of the hold.

The way in which the stroke is synchronized to speech depends on the type of gesture being performed and on the meaning to be conveyed [67]. As a general rule, the stroke of the gesture always precedes or ends with the peak syllable of the speech. The preparation phase, if present, precedes the stroke and could also start during the sentences preceding the one associated to the gesture. Sometimes this occurs to allow the arm/hands to be prepared in time to perform the stroke in time. In other cases preparation anticipation can mean for example that we are formulating the following sentence (the one carrying the most stressed syllable of speech) while still uttering the previous ones. In any case, the stroke must occur no later than the accented syllable of speech. Iconic or metaphoric gestures, instead, could present a variation of this rule: the stroke or multiple repetitions of the stroke happen while the speech containing the idea depicted by the iconic gesture is uttered (*semantic synchrony* rule [67]).

2.2.2 Facial expressions

Facial expressions are movements performed by contracting the muscles of the face. Facial signals have many functions in nonverbal communications. They are used to show our emotional state, our beliefs and goals, our attitudes or opinions towards people (or objects, places, situations, etc.), to regulate the flow of conversation, and so on.

Many researchers agree on the theory that facial expressions are the main mean for communicating our emotional states [38] [4] [2] [58]. When showing our emotional state, there is a set of facial

expressions that do not vary across many cultures in the world [38]. These expressions are associated with the emotional states of: happiness, surprise, fear, sadness, anger and disgust. When we are sad, for example, we raise the inner part of eyebrows, while the outer part is lowered; at the same time we press down our mouth's external corners. When showing anger we pull down and draw together the eyebrows; we raise our upper eyelids and we square and tighten our lips. While we speak, we always have a goal [5] [80]: to inform the interlocutor about something, warn him, suggest something, ask him to perform some action, etc.. Facial expressions can be used to aid the communication of these types of information: for example, raising the eyebrows when suggesting something, or making a small frown when warning or thinking [80]. Facial expressions are also used in regulating the flow of conversation. For example we can open our mouth to try to interrupt other people speaking and to show that we want to speak. Or we can use a smile to encourage the others to engage in a conversation. Some researchers have found that specific facial displays are associated with conveying our attitude, evaluation or opinion about people, object, events, etc. These displays are called *personal reactions* by Chovil in [27]. For example, wrinkling nose (like in the emotional emblem of “disgust” [37]) can be used to show dislike or disapproval; the same meaning can be conveyed by raising the eyebrows raising plus raising one side of upper lip and squinting the eyes. Some facial displays are used in coincidence with the syntactic punctuation of the text we are pronouncing (commas, exclamation or question marks, pauses, etc.). For example, eyebrows raising is usually used to underline exclamations while eyebrows lowering is placed at juncture pauses. Finally, the face is used by the listener to give feedback to the speaker. The listener smiles to show approval for what the speaker is saying; or he can frown his eyebrows to signal that he is not understanding, and so on. In other cases, we use facial expressions to complement or substitute speech. For example, an *ironical face* (for example, wrinkling nose like for showing “disgust”) while saying “you look nice today” allows us to communicate that we think exactly the opposite concept [80]. Sometimes instead, facial expressions emphasize what is going to be said. Raising or lowering the eyebrows is the most common emphasis signal conveyed through facial expression [36] [27]. Less frequently, emphasis is accompanied by eyes widening or tightening. These emphasizing signals can be compared to the *beats* of the gesture modality as they occur while the corresponding stressed syllable is pronounced (see Section 2.2.1). When facial expression completely replaces speech, we talk about *facial emblems*. These expressions have direct verbal translations [38] that is common to the individuals of certain cultural group. In some cultures, like in the U.S., they are mainly produced with the eyebrows. An example of a facial emblem is the *eyebrow flash* (a repeated quick eyebrow raising movement) which is part of a greeting signal in cultures of New Guinea.

Sometimes, parts of facial displays associated with the emotional states are used. For example smiling is an emblem used to greet in many cultures. Or the mouth opened like in the surprise emotional state can be used for the same purpose as the verbal “wow!”.

Temporally speaking, the execution of facial expressions is usually splitted into three phases: the temporal interval in which muscles contract is called the *onset*; after that, the facial expression is shown on the face for a certain time interval, called the *apex*; finally, the facial expression disappears from the face during the *offset* phase. Figure 2.3 shows the *trapezoid* representing the level of activation of a facial expression over time.

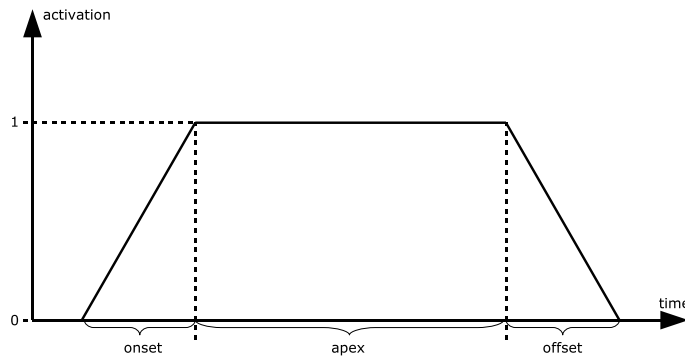


Figure 2.3: Temporal execution of a facial expression.

2.2.3 Torso movements

Torso movements in nonverbal communication have not been widely investigated. Until now, researchers have mainly focused on posture (how we change posture during conversation) and body in general (e.g., jumping if we are happy). With the word *torso* we refer to the upper part of the body, that we may also call *trunk*, including the shoulders, in the sense of the shoulder alignment and rotation. For example, “rotating the shoulders away from the listener” can be considered as a torso movement, or “bow with the trunk to greet someone” is again a torso signal. Therefore, we refer to other works in which the posture shifts have been investigated, even if a posture change involves not only the trunk but the whole body configuration. Cassell et al. [22] have observed some monologues and transcribed posture shifts which occurred at discourse turn and segment boundaries. They did not code the shifts occurring as part of a whole body gesture, for example when changing posture for

performing a deictic gesture when giving directions. The observations have shown that posture shifts occur at the beginning and the end of a discourse segment, more often when the topic of conversation is new. The same happens for the listener, who changes his posture according to the speaker's topic variation. As for the other nonverbal modalities, torso contributes to conversation regulation [90]: by leaning the torso forward we may signal we want to speak, and we may keep this posture while speaking; by leaning the torso backward we may signal we are finished, and possibly we leave the floor of conversation [77]. Emblems can also be produced with the torso, for example by shrugging the shoulders to indicate lack of knowledge or interest [90]. Allwood [2] refers to another possible function of posture changes: showing threat or shyness to give information of one's attitude. Gallaher [43] has investigated the relations between some behavior movement styles (energetic, quick, jerky, etc.) and the occurrence of signals on different modalities. For the torso, the observed movements were: "leaning when standing" and "vertically tilted posture". The tilted or collapsed posture has been found also in observations on nonverbal behavior expressing emotions, in experiments conducted by Wallbott [99]. He has found that this posture is associated with emotional states of disgust, sadness, shame and boredom. The upper trunk is moved forward when showing disgust, despair and fear. Finally, the trunk is raised mainly in case of anger, joy and fear. From all the studies listed above, it seems that torso movements are of two kinds: first, the torso can be used to adjust the actual posture, that is a movement is performed to reach a new position of the upper part of the body, including the shoulders, and then the position is maintained until a new posture is assumed; second, as also reported in [101], the torso can be used to perform gestures, sometimes associated with hand/arm/head gestures, and in this case the execution of the torso movement has a temporal execution that goes from an initial position, reaches the complete activation and then goes back to the initial position. In this sense, we can consider torso movements as being split into phases, as is the case for the arm/hand gestures: *preparation*, (opt.)*hold*, *stroke*, (opt.)*hold*, *retraction*.

2.2.4 Head movements

During conversation, we use head movements, besides other modalities, to communicate nonverbally to the other participants. In [48], Heylen reviews many researchers' works about communicative functions of head movements. One of the main functions is related to conversation regulation: the speaker usually changes his head posture just before he starts speaking, and holds it until the end of his turn; the listener uses the head to give continuous feedback to the speaker, for example with head nods and shakes. The head is also used in communicating some cognitive process made by the speaker,

for example the expression of “thinking” involves usually turning the head away from the listener. Certainty can be reinforced by head nods while uncertainty is marked by small lateral shakes. The head can perform deictic gestures, like arm/hand gestures: it rotates toward a point in space (an object, a person, a particular place, etc.) and the final part of the gesture is accompanied by a small linear movement toward the target (for example by slightly raising the head). Emblematic gestures can also be produced with the head [90]: nodding is an example of a head emblem, because it can be directly translated to the word *yes*. Sometimes head movements can have an iconic function: for example when the head is moved downward when talking about someone smaller [48]. Temporally speaking, head movements can be segmented in a way similar to hand/arm gestures. Kendon [54] cites some examples in which the head performs gestures composed by temporal phases as *preparation*, *stroke* and *retraction*. The stroke is again synchronized with the most stressed point of the accompanying speech [58]. Several researchers have identified head gestures that are used to convey emotional information [99]: head down for expressing disgust and shame; head up/backward for joy, pride and boredom; head bent sideways for cold anger and boredom.

2.3 Communicative intention

As we described in Section 2.1, nonverbal communication is the process of transmitting information from the sender to the receiver by producing nonverbal signals [4]. The goal of communicating may be internal or external, and if internal, conscious, unconscious or tacit [78]. Strictly speaking, a communicative “intention” is an internal conscious communicative goal of an agent, while a communicative “function” is an external (either biological or social) communicative goal of conveying some information, like a blushing face or the flight of a seagull warning the flock (biological communicative goals), or the police uniform (social communicative goal). Thus, in some cases the goal of conveying an emotional state may not be a real communicative intention in the strict sense; yet, in our text, for the sake of simplicity, we will use the term “communicative intention” to refer to any information that a person (or an agent) has the goal of communicating, whether internal or external, conscious, unconscious or tacit.

In this Section we provide a definition of the possible intentions the sender may aim at communicating nonverbally to the receiver during communication. We present a taxonomy of communicative intention based on the theories of Poggi et al. [78] [77] [15]. According to this taxonomy, the information that a speaking person aims to convey while communicating with others always belongs to one of

the three classes: (i) information about the world, (ii) information about his mind and (iii) information about his identity. The first type of information conveyed relates to the speaker's beliefs about the world (e.g. objects, events). The second type includes the speaker's beliefs, goals and emotions concerning what the speaker is communicating. The third type relates to the speaker's identity: sex, age, social and cultural roots. We illustrate the first and second types in the next two Sections.

2.3.1 Information about the world

While communicating with others, we seek to convey our knowledge about the world: for example about objects and their characteristics (size, shape, location, etc.), events (real or imaginary), places (location, distance, appearance, etc.), and so on. Even if this information is typically transmitted through the verbal modality, sometimes it can be expressed by nonverbal signals. For example when objects we produce gestures that mimic their shape and size (e.g. rounded, small, etc.) and we indicate their location by extending our arm and index finger toward them.

2.3.2 Information about the speaker's mind

According to Poggi [77], while we speak, we nonverbally give information about three possible aspects of our mind: our beliefs, our goals and our emotions.

When we express our beliefs, we give information, for example, about the degree of certainty we attribute to what we are saying. For example by nodding and shaking with the head we communicate that we believe something is (respectively) true or false. Or, we can provide *metacognitive* information: it represents the source of the beliefs we are communicating. For example snapping fingers can communicate the fact that we are searching in our memory, so the retrieved information may be not very reliable. The same happens for example when we look up trying to remember. We communicate metacognitive information also with body: while thinking our body posture can be slightly deflated.

While speaking, we always have a final goal [5]: to disagree with our interlocutor, to assert something, to inform about some important thing, etc. We use nonverbal behavior (facial expressions, gestures, etc.) to underlie the part of the discourse which is more relevant to our goal (also called the *rhetic* information). Or, on the other hand, if we are talking about the information which is less important (the *thematic* information) we signal this by looking away from the listener. Sometimes our main goal may be expressed by two or more sub-goals organized in a structured way. For example sometimes we want to communicate the relation between two concepts to explain that the first is the *cause* and the second is the *effect*. In this case we can communicate this *metadiscursive*