University of Zagreb
Faculty of Electrical Engineering and Computing
Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Goranka Zorić

Hybrid approach to real-time speech driven facial gesturing of virtual characters

Hibridni pristup stvaranju gesti lica virtualnog lika govornim signalom u stvarnom vremenu

Doctoral thesis
Doktorska disertacija

Zagreb, 2010.

Doctoral thesis written at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb and at the Division of Information Coding at Linköping University

Advisor: Prof. dr. sc. Igor S. Pandžić

This thesis has 85 pages.

Thesis no.:

The dissertation evaluation committee:

1. Professor Maja Matijašević, Ph.D.
   Faculty of Electrical Engineering and Computing, University of Zagreb

2. Associate Professor Igor Sunday Pandžić, Ph.D.
   Faculty of Electrical Engineering and Computing, University of Zagreb

3. Assistant Professor Krešimir Matković, Ph.D.
   VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Wien

The dissertation defense committee:

1. Professor Maja Matijašević, Ph.D.
   Faculty of Electrical Engineering and Computing, University of Zagreb

2. Associate Professor Igor Sunday Pandžić, Ph.D.
   Faculty of Electrical Engineering and Computing, University of Zagreb

3. Associate Professor Robert Forchheimer, Ph.D.
   Linköping University, Sweden

4. Professor Dragan Jevtić, Ph.D.
   Faculty of Electrical Engineering and Computing, University of Zagreb

5. Associate Professor Davor Petrinović, Ph.D.
   Faculty of Electrical Engineering and Computing, University of Zagreb

Date of dissertation defense: 09^{th} July 2010.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With an increasing everyday use of various computer devices, the question of human-computer interaction (HCI) becomes more and more important. Designing interfaces to be more natural, effective, as well as user oriented, improves our interaction with technology. One way to approach this is through believable virtual humans which HCI has been researched for almost two decades. Virtual humans can take over roles as virtual guides, teachers, newscasters, story tellers or sales persons, act in films or be avatars in chat rooms, teleconferences and games. For the users to benefit from the use of virtual humans, they need to look and behave as real humans – and humans are social beings who have developed complex ways of communication.

Human communication consists of information exchange where different modalities are used to transmit messages, i.e. it is multimodal communication. This includes speech, facial expressions, body posture and gestures. A human face capable of showing a diversity of facial expressions is among the most important modalities of communication. Not only lip and tongue movements, but also head and eyebrow movements, blinking and the like contribute to the message understanding. Under conditions when one communicative channel experiences interference like a noisy environment, or in cases of a human disability, e.g. a hearing impairment, HCI benefits from multimodal interfaces due to information redundancy.

Virtual humans, i.e. human-like virtual characters (VCs), are graphical simulations of real or imaginary persons capable of a human-like behaviour, most importantly talking and gesturing. In human-computer interaction, they represent a multimodal interface where modalities are the natural modalities of human conversation such as speech, facial expressions, gestures, body movements and body posture.

1.1 Contributions

The focus of this thesis is in the area of virtual human development. More precisely, it is about the modelling of nonverbal, speech-related facial gestures for virtual humans, i.e. its use in the automatic generation of facial animation from the speech signal in real-time. Facial gestures include various nods and head movements, blinks, eyebrow gestures and gaze, i.e., all facial displays used in communication, including head movements but excluding explicit verbal and emotional displays. Virtual news presenters have been chosen as a representative application for speech driven facial gestures, since presenters generally do not show emotions. In order for virtual humans to be properly built, a knowledge from various disciplines is needed – psychology, sociology, communication science, anthropology, linguistics and computer science. Figure 1.1 symbolically shows research fields on which this thesis largely relies: research on nonverbal communication, digital speech processing and embodied conversational agents (ECAs).

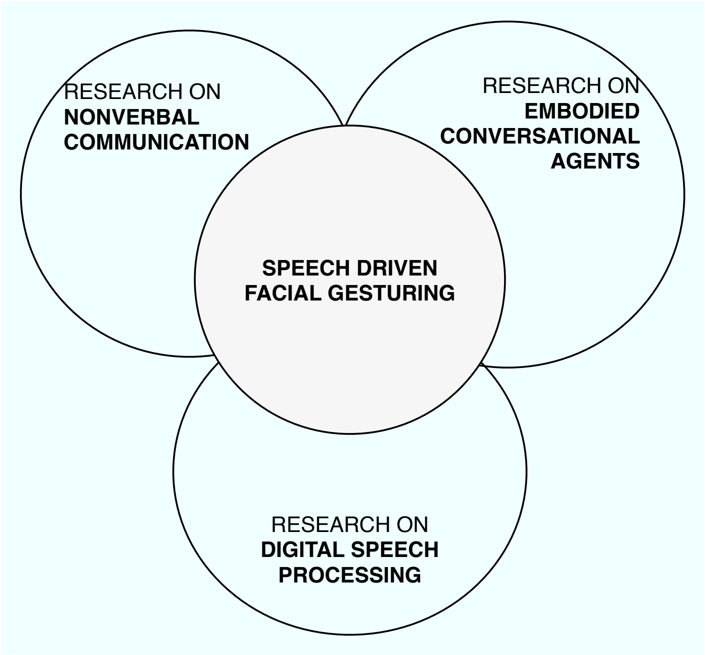

RESEARCH ON
NONVERBAL COMMUNICATION
RESEARCH ON
EMBODIED CONVERSATIONAL AGENTS
SPEECH DRIVEN FACIAL GESTURING
RESEARCH ON
DIGITAL SPEECH PROCESSING

Figure 1.1: Research on speech driven facial gesturing in this thesis relies on three main fields

The summarized contributions of this thesis are:

1. A systematic overview of facial movements and their connection with the speech signal. Movements included are those used in nonverbal communication that are

not connected with emotions or semantics.

2. A hybrid method for mapping from the speech signal to statistically correct facial gestures correlated with the speech prosody in real-time.

3. A system for automatic facial gesturing for virtual characters in real-time based on the proposed hybrid method.

4. A validation of the method and the system using perceptual evaluation. The following hypothesis is set up: *adding facial gestures driven by the speech signal to the animation of virtual characters using the proposed hybrid approach in real-time will result in more coherent and appropriate facial movements than random or no movements.*

1.2 Thesis outline

The remaining chapters of this thesis are organized as follows:

- Chapter 2 (*Speech related facial gestures*) presents an overview of nonverbal communication and the speech related facial gesturing. It includes related work and summary of the most common facial gestures.

- Chapter 3 (*Proposed method for generating visual prosody*) describes a proposed hybrid approach for mapping the speech signal to facial gestures.

- Chapter 4 (*System for speech driven facial gesturing for virtual characters*) describes the system for speech driven facial animation of virtual characters based on the proposed hybrid method, including relevant implementation details, and the method information.

- Chapter 5 (*Perceptual evaluation of visual prosody*) gives a description of the experiment designed for perceptual evaluation of visual prosody, together with a statistical analysis of results and a matching discussion.

- Chapter 6 (*Speech driven virtual news presenters in networked environment*) puts the system in a context of virtual news presenters and describes the issues that arise in networked environments.

- Chapter 7 (*Conclusions, and future work*) - concludes the paper and suggests future works.

The majority of the work done for this thesis has been carried out at the Department of telecommunications and the *Human Oriented Laboratory* (HOTLab), at FER in Zagreb. Design and implementation of the proposed method for speech driven facial gesturing (Chapters 3 and 4) present the work performed when the author was visiting the *Division of Information Coding* at *Linköping University* in collaboration with professor Robert Forchheimer.

Chapter 2

Speech related facial gestures

For a human-like behaviour of a talking head all facial expressions need to be accurately simulated synchronously with an underlying speech signal. Most efforts, so far, have been focused on the articulation of lip and tongue movements, i.e. lip synchronization. Besides a visible speech, nonverbal component needs to be included in order not to have a "stone-cold" face. Emotional displays have been investigated to a great extent, yet, it is important to consider other roles of the face in communication, e.g. situations such as when facial displays are used to emphasise or punctuate what is being said.

This chapter gives background information and the idea behind this thesis, and summarizes related work. Starting from a brief introduction on nonverbal communication in Section 2.1, and in particular facial displays in Section 2.1.1, it continues with speech-related facial gestures and visual prosody in Section 2.2. Section 2.3 gives a relevant work for generating facial gestures correlated with the speech signal in Section 2.3.1, and reports on related systems in Section 2.3 2. Section 2.4 summarizes collected knowledge on the most used facial gestures.

2.1 Nonverbal communication

Bearing in mind various modalities that humans use for communication, it is obvious that much can be said without the words, nonverbally. Nonverbal communication refers to all aspects of the message exchange excluding only communication through words. It includes all expressive signs, signals and cues apart from manual sign language and speech. Examples are: body language including facial expressions, eye movements, gestures, posture and the like, eye contact, touch and smell, paralanguage, vocal features, physical appearance and artefacts.

Although the exact amount of nonverbal information exchanged during a face-to-face conversation is still unknown, it is clear that the nonverbal channel plays an important role in understanding a human behaviour. If people rely only on words to express themselves, it can lead to difficulties in communication. Words are not always associated with similar experiences, similar feelings or even meaning by listeners and speakers. A situation when there is no consistency, i.e. one signal is being said and another is shown can be very misleading to another person. When listeners are in a doubt they tend to trust the nonverbal message since, according to psychiatrist Jurgen Ruesch, as cited by Givens [2], disturbances in nonverbal communication are *more severe and often longer lasting* than disturbances in verbal language.

Nonverbal communication has multiple functions. It can repeat, accent [3], complement or contradict the verbal message, regulate interactions or substitute for the verbal message, especially if it is blocked by noise, interruption, etc. Nonverbal communication can be subconscious, meaning that message can be transmitted even if the speaker or the listener are not aware of the cues used, or it can also be unintended, e.g. transmitted message is against speaker's will. Moreover, culture, gender and a social status might influence the way nonverbal communication is used (e.g. some cultures forbid direct gaze while others find gaze aversion an offence [4]). Nonverbal cues may be learned, innate or mixed [2]. Some of them are clearly learned (e.g. eye-wink, thumbs-up) and some are clearly innate (e.g. eye blinking, facial-flushing). Most of nonverbal cues are mixed (e.g. laugh, shoulder-shrug), because they originate as innate actions, but cultural rules and environment shape their timing, energy and use.

2.1.1 Facial displays

The human face is an important communication channel in face-to-face communication. Even our brain is adapted to such communication – a particular region in primates' brains is in charge for facial information processing [5]. Through the face emotional and communicative signals are displayed. For those signals, a facial displays term is usually used, e.g. in [6, 7], since the facial expressions term has connotation of emotions.

Starting with Charles Darwin's book *The Expression of the Emotions in Man and Animals*, published in 1872., facial displays have been studied in the great extent, e.g. in [8, 9, 10]. Paul Ekman in [11], as cited by Pelachaud in [12], characterizes facial expressions as follows:

- **Emblems.** Correspond to movements whose meaning is very well known and culturally dependent. They may replace, or repeat a word (or small group of words) and can be directly translated into a verbal statement. Example is nodding instead of saying "yes".

- **Emotion emblems.** Correspond to signals used to convey emotion. They reference particular emotion without requiring the person mentioning emotion to feel it at the moment of expression. An example is the eyebrow raising as a part of the facial expression of surprise.

- **Conversational signals.** Correspond to the signals that clarify and support what is being said. They are synchronized with accents or emphatic segments. An example is the eyebrow raising occurring to signal an exclamation mark.

- **Punctuators.** Correspond to the signals that support pauses. They group or separate the sequences of words into discrete unit phrases, thus reducing the ambiguity of the speech. An example is special head movements occurring during pauses.

- **Regulators.** Control the flow of conversation (e.g. a speaker breaks or looks for an eye contact with a listener, or a speaker turns his head towards or away from a listener during conversation).

- **Manipulators.** Correspond to the biological needs of a face (e.g. eye blinking to wet the eyes) and have nothing to do with the linguistic utterance.

- **Affect displays.** Correspond to facial expression of emotion, e.g. a smile as a part of the facial expression of happiness.

Facial displays can also be characterized by the amount of time that they last [13]. Some facial displays are linked to the personality and remain constant across a lifetime (e.g. frequent eye blinking), some are linked to the emotional state, and may last as long as the emotion is felt (e.g. looking downward and frowning in case of sadness), and some are synchronized with the spoken utterance and last only a very short time (e.g. eyebrow rising on the accented word).

Considerable research is being done on the topic of facial displays with focus on the relationship between facial displays and emotional states [14]. Still, according to Ekman and Friesen, as cited by Chovil in [14], less than one third of nearly 6000 facial displays of psychiatric patients in their experiments were classified as emotional.

Accordingly, many works, like [6, 15, 4, 7, 16, 17, 18] agree that great attention has to be paid to communicative signals since they play significant role in the communication.

2.2 Visual prosody

In this thesis, communicative facial displays, named facial gestures, are observed. Facial gestures include various nods and head movements, blinks, eyebrow gestures and gaze, i.e. all facial displays, including head movements, but excluding explicit verbal and emotional displays like visemes or expressions such as smile [19].

According to the P. Ekman's characterization (*Section 2.1.1*), facial gestures that happen as conversational signals, punctuators, and regulators (e.g. eye blinking as a signal of a pause in the utterance or the eyebrow raising to emphasize what is being sad), and not those happening as emblems, emotion emblems, manipulators or affect displays (e.g. the eyebrow action as "but" or the eyebrow raising as a signal of surprise) are observed. According to the general linguistic function as used in [14, 8, 20, 6]), such facial gestures serve as syntactic displays conveying grammatical information, e.g. eyebrow raising on exclamation mark or eye contact at the end of story. Conversational context and timing, synchronization or delay with the speech and other communication modalities is important for interpretation of these facial gestures, since it depends upon the situation in which they appear.

Thus, the focus of research are those facial gestures that are connected to the syntactic and prosodic structure of the speech and not to the semantics (words or their meaning), emotion or personality. They are connected to prosody and paralinguistic information, and are called visual prosody.

Visual prosody assumes various facial gestures continuously used during speech articulation, i.e. it includes head and facial movements analogous to prosody in speech analysis. To model visual prosody, the relationship between prosodic parameters of the speech signal and the facial gestures correlated with the speech needs to be considered.

While acoustic prosody has been investigated to a great extent, much about visual prosody is still unknown. On one hand there is a relationship between the speech signal and lip movements which is obvious (e.g. phoneme-viseme), and on the other there is a relationship between facial gestures (also gestures in general) and the speech signal which is not so strong and obvious. Still, there is psychological and paralinguistic research relevant to synthesizing visual prosody, as described in *Section 2.3.1*.

2.2.1 Speech prosody

Speech prosody refers to characteristics of speech signal, such as intonation, rhythm, and stress, which cannot be extracted from the characteristics of phoneme segments. It's typical acoustical correlates are fundamental frequency (pitch), intensity (loudness) and syllable length. Paralinguistic refers to the nonverbal elements of communication which are used to modify meaning and convey emotion. It includes pitch, loudness and intonation of the speech.

Prosodic speech events can be described by the temporal variations of loudness and pitch, as well as by the insertion of pauses, phoneme durations and timing. Among these, the most expressive is pitch which represents the perceived fundamental frequency (F0) of a sound. In this work changes in pitch and loudness, as well as long and short pauses were used. In the next paragraph, relevant findings on the use of pauses are described.

Pauses

According to [21] there are two classification of pauses generally used: physical and linguistic classification and psycholinguistic. The first, physical and linguistic classification, looks at the pause as physical entity – a normal speech flow is interrupted whenever a brief silence can be observed in the acoustic signal. As a silence, a segment with no significant amplitude is considered, whereas exact duration of such segment depends on its linguistic context. Differentiated are intra-segmental pauses, like the voice onset time of a stop consonant with an upper threshold of around 100 ms, as stated by [22], and inter-lexical pauses which may appear between two words, and tend to be longer. However, perceived pauses are not really the equivalent of physical pauses. Zellner in [21] further states:

Some pauses are more easily perceived than others, and generally, such pauses appear to support particular functions within the message, such as grammatical functions, semantic focus, hesitation, and so on. Also, pauses are more easily perceived if their duration is around 200 – 250 ms. That appears to be the standard auditory threshold for the perception of pauses.

Psycholinguistic classification distinguishes silent pauses corresponding to the perception of unvoiced segments in the speech signal, and filled pauses corresponding to the perception of any spoken sound or word used to fill gaps in the speech. Generally, silent and filled pauses occur between words.

The occurrence and the length of the pauses is to a great extent individual and depends on the nature of talk. In [23], as referenced by [21], it is shown that in the task like it is describing cartoons, pauses are both longer, with mean value 1320 ms, and more frequent than when answering interview questions, with mean value 520 ms. As well, there is a significant difference in the length of pauses when comparing professional speakers reading the news, non-professional reading and spontaneous speech.

Pauses are one of the forms in which disfluencies in speech appear. Others include repetition of words, filled pauses or spurious words, and they are excluded from this work.

2.3 Related work

2.3.1 Psychological and paralinguistic research

P. Ekman investigated systematically the relation of the speech and eyebrow movements in [24]. According to his findings, eyebrow movements occur during word searching pauses, as punctuators or to emphasize certain words or parts of the sentence. Lowered eyebrows might indicate difficulty, or doubt, while eyebrows are often raised to show a question being asked. During thinking pauses, i.e. searching for word, raised eyebrows occur accompanied by an upward gaze direction and, typically, people tend to look at a still object to reduce visual input. Eyebrows also may be lowered in this situation, especially in conjunction with a filled pause like "errr", "mmm" or similar.

Chovil in [14] concentrated on the role of facial displays in conversation. Her research results showed that syntactic displays, i.e. emphasized words or punctuators, are the most frequent facial gestures accompanying the speech. Among those facial gestures, raising or lowering eyebrows are the most relevant.

The strong interrelation between facial gestures and prosodic features was given in following works.

Cavé *et al.* in [25] investigated links between rapid rising and falling eyebrow movement and fundamental frequency changes, rising and falling. Correspondence in 71% of the examined cases was found. Another finding was that typically 38% of overall eyebrow movements occur during pauses, in particular during hesitation pauses, or while listening. This suggests that eyebrow movements and fundamental frequency are not automatically linked, i.e. not the result of muscular energy, but are consequences of linguistic and communicational choices. They serve to indicate turn taking in dialogues, assure the speaker of the listener's attention, and mirror the listener's degree of understanding, serving as back-channel.

Kuratate *et al.* in [26] presented a preliminary evidence for a close relation between head motion and acoustic prosody, specifically the pitch contour. They concluded that production systems of the speech and head motion are internally linked. These results were further elaborated in [27, 28, 29] within the same group, adding the head motion correlation with the amplitude of the subject's voice and showing that nonverbal gestures such as head movements play significant role in the perception of speech. These results were used to animate a natural face and head motion.

Granström *et al.* investigated in [30] contribution of eyebrow movement to the perception of prominence. Further on, head movements were added, indicating that combined head and eyebrow movements are effective cues to prominence when synchronized with the stressed vowel [31] stating that the perceptual sensitivity to timing is around 100 ms to 200 ms, which is about the average length of a syllable. Similarly, Krahmer *et al.* in [32] report that both pitch accents and eyebrow movements can have a significant effect on the perception of a focus.

2.3.2 Related systems

In this subsection systems which include facial gestures and, more specifically, visual prosody are described. Systems are divided according to the type of the input data used for generating facial gestures, whereas the focus is on speech and text driven systems.

As it is stated in the introductory section, valuable knowledge for building virtual humans comes from research on Embodied Conversational Agents (ECAs), thus some included systems are from ECA field.

ECAs are computer generated human-like characters that interact with human users in face-to-face conversations [33]. ECA based systems express modalities that are natural to human conversation, both verbal and nonverbal. Examples are words, voice, facial expressions, gestures, body movements or body posture. If carefully designed for a specific task, ECAs give a social component to an interaction with computers, improve user experience and enhance effectiveness. A work done in [34] showed the importance of virtual agent's design in accordance with its task. Earlier investigations found out that agents produce a feeling of social presence for users [35], and this feeling increases when the agent looks like a person instead of having an animal or abstract form [36]. However, people generally expect an ECA to behave in accordance with its appearance – people will be more critical to behaviour of highly photo-realistic agents than to cartoon characters. A work in [37] showed that user's experience of ECA is positive when perceived believable behaviour and vice versa.

Systems for facial gesturing using annotation

DeCarlo *et al.* in [17], in their exploratory data analysis, have found that small head movements related to discourse structure and interpretation are among the most common nonverbal cues people provide. In that work, a real-time facial animation system is described. It accepts input as a text with open-ended annotations specifying head motions (e.g. tilting nod on a phrase, downward nod accompanying a single syllable and upward nod accompanying a single syllable), and other facial actions (e.g. eyebrow movements following work in [24]). In that work facial gestures are annotated manually and given as input.

Text driven systems for facial gesturing

There are numerous visual text-to-speech (VTTS) and dialogue systems based on the speech recognition and text-to-speech (TTS) engine that incorporate speech-related facial gestures in some way. Those works are generally concerned with the semantic structure of the speech, i.e. meaning, and/or are related to specific scenarios, or incorporate only some aspects of speech-related facial gestures.

Takeuchi *et al.* in [6] analysed the conversation between users and the speech dialogue system with added communicative facial displays – syntactic, speaker and listener comment. In this case, the speech dialogue subsystem recognizes a number of typical conversational situations (e.g. recognition failure with "non-confident" facial display – eyebrow lowering; beginning of dialogue with listener comment display "indication of attendance" – eyebrow raising, mouth corners turned down; shift to another topic with syntactic display "end of story" – eye contact and "beginning of story" – eyebrow raising) and associate them with facial displays.

Cassel *et al.* in [15] automatically generate and animate conversations between multiple human-like agents. Conversations are created by a dialogue planner that produces the text as well as the intonation of the utterances which than together with the speaker/listener relationship drive hand gestures, lip motions, eye gaze, head motion (nod) and facial expressions – conversational signals, punctuators, manipulators and emblems following principles of a synchrony. A conversational signal starts and ends with the accented word, punctuator happens on the pause, blink is synchronized at the phoneme level, while emblems are performed consciously and must be specified by the user.

Cassell *et al.* in [7] experiment with the autonomous conversational agent in a specific scenario. Their results confirmed that envelope feedback, defined as nonverbal behaviour related to the process of conversation, is more important in interaction than emotional feedback, i.e. emotional expression, thus playing a crucial role in supporting the dialogue. The envelope feedback that was included consists of: turning head and eyes towards user when listening, averting gaze and lifting eyebrows when taking turn, gazing back at person when giving turn, tapping fingers to show that it is "alive" and beat gesture when providing a verbal content.

Lundeberg *et al.* in [38] developed a new talking head with the purpose of acting as an interactive agent in a dialogue system. To add to the realism and believability of the dialogue system, emotional cues, turn-taking signals and prosodic cues such as punctuators and emphasizers were given to the agent. The main rules they followed for creating the prosodic gestures were to use a combination of head movements and eyebrow motion and maintain a high level of variation between various utterances.

Pelechaud *et al.* in [4] reports results from a program that produces an animation of facial expressions and head movements, conveying information correlated with the intonation of the voice. Facial expressions are divided by its function called determinant. The algorithm for each determinant is described, with special attention given to the to lip synchronization and coarticulation problems. Determinants considered are: conversational signals, punctuators, manipulators and regulators. Only functions related directly to pattern of voice are included. The computation of each determinant of facial expressions is done by a set of rules. Facial actions included are various head and eye movements, and intonation. The input to the program is a file containing phonological representation of an utterance, with its accents marked.

The BEAT system [39] controls movements of hands, arms and the face and the intonation of the voice, relying on rules derived from extensive research in the human conversational behaviour. It takes a plain text as an input, annotates it with contextual and linguistic information, based on which various gestures are suggested. The language tags that are currently implemented are clause, theme and rheme, word newness, contrast, and objects and actions. Examples of the use of facial gestures within BEAT system include raising eyebrows to signal introduction of new material, and gaze away/towards for theme/rheme.

Graf *et al.* in [40] analyse head and facial movements that accompany speech and investigate how they relate to the prosodic structure of the text. Prosody analysis is mainly consisted of identified phrase boundaries and pitch accents which are accompanying pitch movements, indicating whether the frequency of the F0 is rising or falling. Head movements in this work are described with its type (nod, overshoot nod and swing), amplitude and duration. They concluded that despite large variations from person to person, when considered direction and strength of head movements, their timing is typically well synchronized with the prosodic structure of the text.

Pelechaud *et al.* in [41] give ECA some aspects of nonverbal communication using taxonomy of communicative behaviour as proposed by Poggi [42]. To each of those functions corresponds a signal, i.e. a facial expression, gaze behaviour and head movement. The system takes a text and a set of communication functions as an input.

Bui *et al.* in [43] propose a scheme of combining facial movements concentrating on the dynamic aspects of facial movements and the combination of facial expressions in various channels using marked up text as an input. Included channels of the facial movement are: manipulators, lip movements, conversational signals, emotion displays and emblems, gaze movements and head movements.

The Autonomous Speaker Agent in [1] performs dynamically correct gestures that correspond to the underlying text by using lexical analysis and statistical models of facial gestures in order to generate gestures related to the spoken text. The focus is on conversational signals, punctuators and manipulators with three main classes of facial gestures: head movements (nod, overshoot nod, swing, reset), eye movements (movement in various directions and blinking) and eyebrow movements (raise or frown).

Lee *et al.* in [44] have created rules for generating nonverbal behaviour by extracting information from the lexical, syntactic and semantic structure of the surface text. The behaviour included is: head movement (nods, shakes, a head moved to the side, head tilt, pulled back, pulled down), eyebrow movement (a brow raised, a brow lowered, a brow flashes), eye/gaze movement (look up, look down, look away, eyes squinted, eyes squeezed, eyes rolled), shoulder shrug and mouth pulled on one side. Each rule has associated nonverbal behaviours and a set of words that are usually spoken with it (e.g. for contrast – head moved to the side and brow raise co-occurring with words but). As an input an affect state and emphasis are also given to additionally shape generated nonverbal behaviour.

Speech driven systems for facial gesturing

State of the art lacks methods which would be able to automatically generate a complete set of facial gestures by only analysing the speech signal. Existing systems in this field are capable of automatic lip synchronization producing quite correct lip movements (e.g. [45, 46, 47, 48, 49]), but often missing natural experience of the whole face.

For synthesizing general dynamics of the face, a data-driven approaches have been widely used. For example, Brand in [50] learns dynamics of real human faces during the speech by applying a Hidden Markov Model (HHM) to both the audio signal and the facial motion. Given a novel audio, the algorithm generates mouth movements, including coarticulation as well as speech-related facial gestures such as eyebrow movements. Similarly, a system in [51] learns speech-based orofacial dynamics from the video, generating facial animation with realistic dynamics.

Many works generate only head movements.

An automatic data-driven system for head motion synthesis is developed in [52], taking pitch, the lowest five formants, MFCC (Mel-Frequency Cepstral Coefficients) and LPC (Linear Prediction Coefficient) as audio features. A K-Nearest Neighbors (KNN) based dynamic programming algorithm is used to synthesize a novel head motion given a new audio input.

Chuang and Bregler in [53] generate a head motion in addition to expressive facial animation from the speech. They use a database of examples which relate audio pitch to the motion. A new audio stream is matched against segments in the database. A head motion is synthesized by finding a smooth path through the matching segments.

Sargin *et al.* in [54, 55] propose a two-stage data-driven method for synthesizing head gestures from speech prosody for a particular speaker. In the first stage, Hidden Markov Model (HMM) is used for unsupervised temporal segmentation of head gesture and speech prosody (pitch frequency and speech intensity) features separately, while, in the second stage, multistream HMMs are used for joint analysis of correlations between these elementary head gesture and prosody patterns. In the synthesis stage, the resulting audio-visual mapping model is used to predict head gestures from arbitrary input speech given a head model for the speaker. Similarly, a work in [56] generates a sequence of head motion units using Hidden Markov Models given the speech based on the assumption that temporal properties should be taken into the account and therefore the data has to be segmented into longer parts.

A rule-based method for automatic generation of several nonverbal facial expressions from the speech is introduced by Albrecht *et al.* in [57]. Gestures included are:

head and eyebrow raising and lowering dependent on pitch; gaze direction, movement of eyelids and eyebrows, and frowning during thinking and word search pauses; eye blinks and lip moistening as punctuators and manipulators; random eye movement during normal speech. Intensity of facial expressions is additionally controlled by the power spectrum of the speech signal, which corresponds to loudness of the utterance.

Speech driven systems described so far need a preprocessing step. A real-time speech driven facial animation is addressed in [58]. Several facial animation components are differentiated based on the statistical model: head and eyebrow movements and blinking as punctuators, head and eyebrow movements during thinking and word-search pauses and blinking as manipulator. In [59] a system for generating expressive body language, including head movements, in real-time is presented. The system selects segments from motion capture data directly from the speech signal. The selection is driven by Hidden Markov Model and speech features used are pitch, intensity and syllable duration.

Other related systems

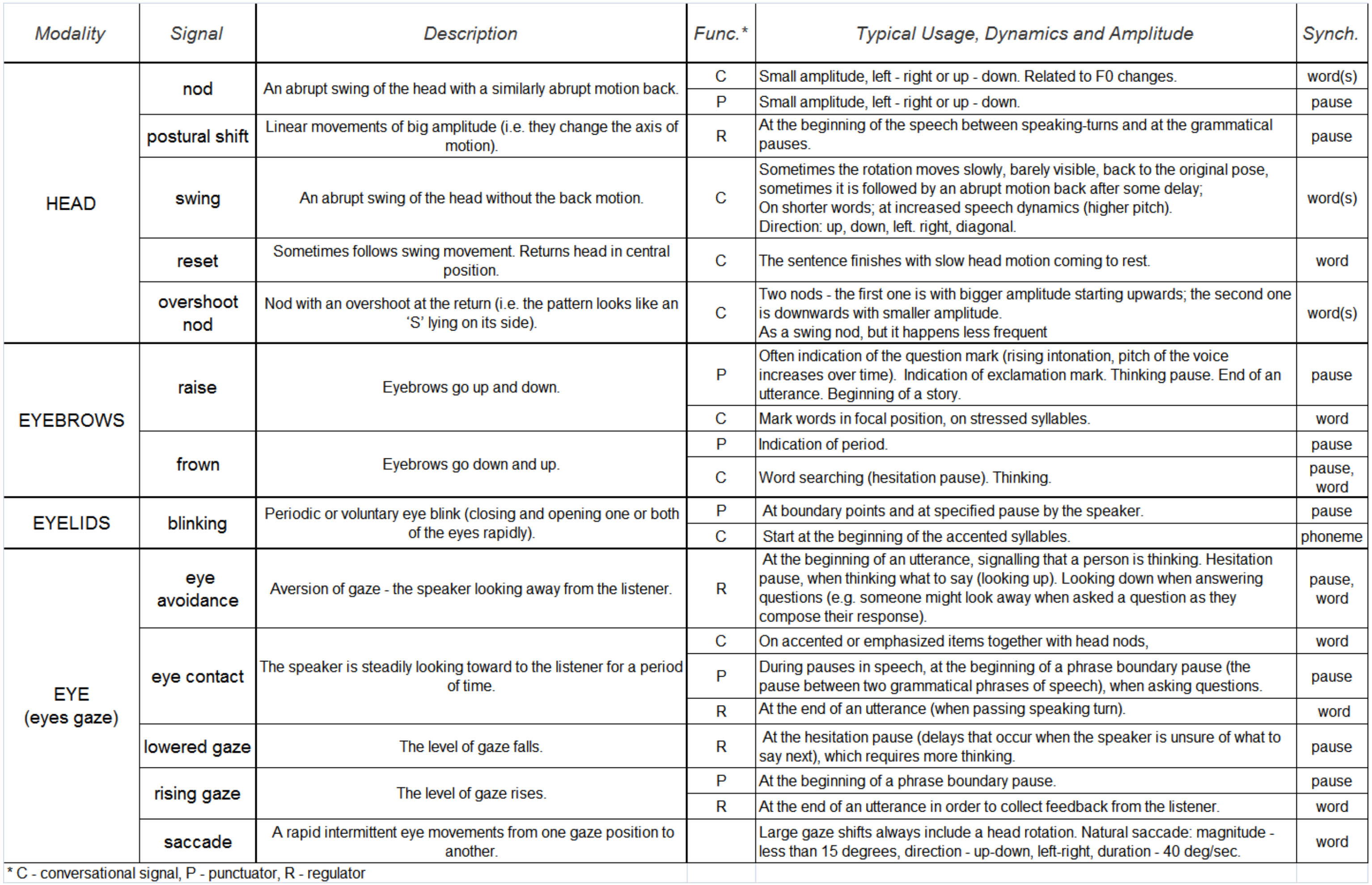
The Eyes Alive system [60] reproduces eye movements that are dynamically correct at the level of each movement, and that are also globally statistically correct in terms of the frequency of movements, intervals between them and their amplitudes (based on the statistical analysis of the eye-tracking video). Speaking and listening modes are distinguished, head rotation is monitored, duration of mutual gaze and gaze away is included, and eye movement is modelled based on parameters which are magnitude, direction, duration and velocity.

2.4 Facial gestures overview

The knowledge needed for generating visual prosody presented so far, combined with some additional information on facial gestures from literature, is systematically organized in this section. A set of usual facial gestures is described with its functions, usage and typical dynamics, including other valuable knowledge, and those data is summarized and presented in two tables.

Table 2.1 is organized as follows.

The first column, named *Modality*, contains the anatomic part of the face that takes part in facial gesture creation, since facial gestures are differentiated by the facial



Modality	*Signal*	*Description*	*Func.**	*Typical Usage, Dynamics and Amplitude*	*Synch.*
HEAD	nod	An abrupt swing of the head with a similarly abrupt motion back.	C	Small amplitude, left - right or up - down. Related to F0 changes.	word(s)
			P	Small amplitude, left - right or up - down.	pause
	postural shift	Linear movements of big amplitude (i.e. they change the axis of motion).	R	At the beginning of the speech between speaking-turns and at the grammatical pauses.	pause
	swing	An abrupt swing of the head without the back motion.	C	Sometimes the rotation moves slowly, barely visible, back to the original pose, sometimes it is followed by an abrupt motion back after some delay; On shorter words; at increased speech dynamics (higher pitch). Direction: up, down, left. right, diagonal.	word(s)
	reset	Sometimes follows swing movement. Returns head in central position.	C	The sentence finishes with slow head motion coming to rest.	word
	overshoot nod	Nod with an overshoot at the return (i.e. the pattern looks like an 'S' lying on its side).	C	Two nods - the first one is with bigger amplitude starting upwards; the second one is downwards with smaller amplitude. As a swing nod, but it happens less frequent	word(s)
EYEBROWS	raise	Eyebrows go up and down.	P	Often indication of the question mark (rising intonation, pitch of the voice increases over time). Indication of exclamation mark. Thinking pause. End of an utterance. Beginning of a story.	pause
			C	Mark words in focal position, on stressed syllables.	word
	frown	Eyebrows go down and up.	P	Indication of period.	pause
			C	Word searching (hesitation pause). Thinking.	pause, word
EYELIDS	blinking	Periodic or voluntary eye blink (closing and opening one or both of the eyes rapidly).	P	At boundary points and at specified pause by the speaker.	pause
			C	Start at the beginning of the accented syllables.	phoneme
EYE (eyes gaze)	eye avoidance	Aversion of gaze - the speaker looking away from the listener.	R	At the beginning of an utterance, signalling that a person is thinking. Hesitation pause, when thinking what to say (looking up). Looking down when answering questions (e.g. someone might look away when asked a question as they compose their response).	pause, word
	eye contact	The speaker is steadily looking toward to the listener for a period of time.	C	On accented or emphasized items together with head nods,	word
			P	During pauses in speech, at the beginning of a phrase boundary pause (the pause between two grammatical phrases of speech), when asking questions.	pause
			R	At the end of an utterance (when passing speaking turn).	word
	lowered gaze	The level of gaze falls.	R	At the hesitation pause (delays that occur when the speaker is unsure of what to say next), which requires more thinking.	pause
	rising gaze	The level of gaze rises.	P	At the beginning of a phrase boundary pause.	pause
			R	At the end of an utterance in order to collect feedback from the listener.	word
	saccade	A rapid intermittent eye movements from one gaze position to another.		Large gaze shifts always include a head rotation. Natural saccade: magnitude - less than 15 degrees, direction - up-down, left-right, duration - 40 deg/sec.	word

* C - conversational signal, P - punctuator, R - regulator

Table 2.1: Set of usual facial gestures described with its functions, typical usage, dynamics and other valuable data

regions involved, i.e. modality used. Included are following modalities: head, eyebrow, eyelids and eye (representing the eye gaze).

Various signals (e.g. eyebrow – raising, frowning) that might be produced by a single modality are given in the second column, named *Signals*. For some modalities, for example the head there are several known signals such as various kinds of nods, while some modalities are characterized with only one signal (e.g. eyelids with the blinking).

In the column named *Description* each facial gesture is briefly defined.

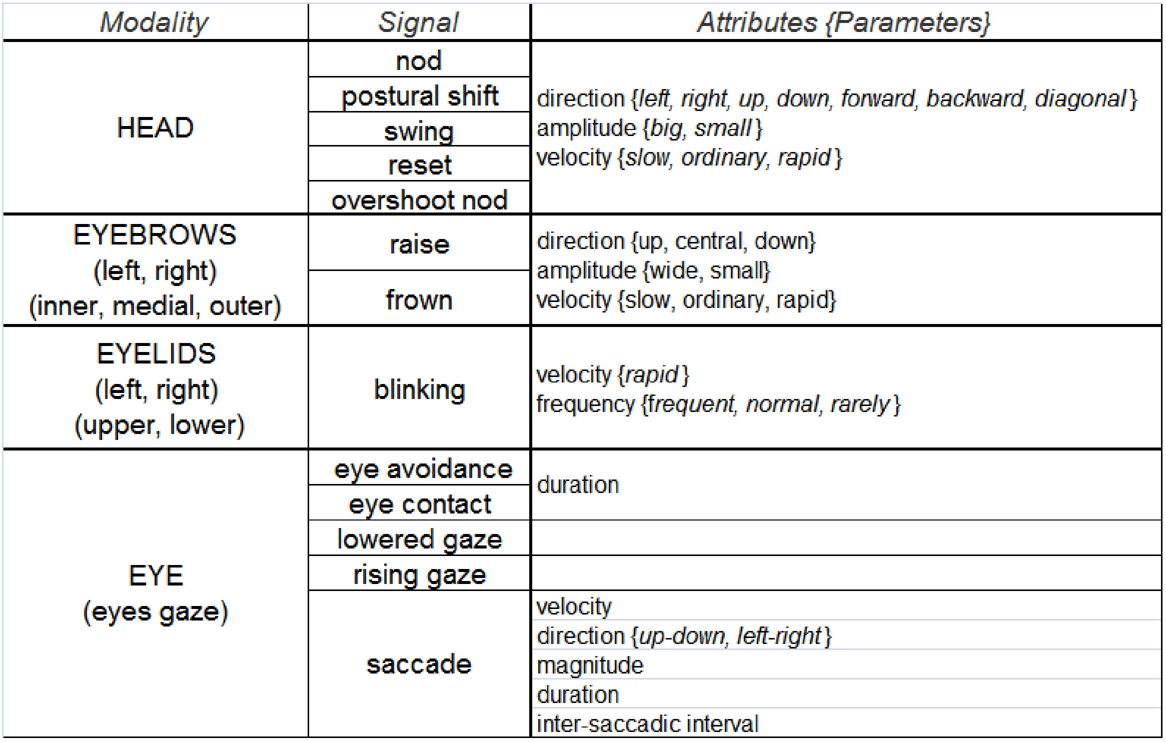
A function that a facial gesture can have is given in the column *Function*. Functions supported are: conversational signal (C), punctuator (P) and regulator (R). Some facial gestures can have more than one function (e.g. frowning serves as the conversational signal as well as the punctuator).

According to the function signal can have, it is additionally described by its usage, typical dynamics or amplitudes, and level of synchronization. The fifth column, named *Typical Usage, Dynamics and Amplitude* provides information about each gesture, according to the function it has and concerning a typical usage scenario. It gives information about situations in which certain facial gesture might appear and its meaning. Also, it gives the available knowledge on typical dynamics and amplitudes of the gesture including known rules for speech related facial gesturing. For example head nods are characterized with the small amplitude and the direction up-down or left-right. Another example are saccadic eye movements which are often accompanied by a head rotation.

The last column, named *Synch.*, i.e. Synchronization, specifies on which level is the gesture synchronized with the underlying speech. Verbal and nonverbal signals are synchronized, and synchrony occurs at all levels of speech: phonemic segment, word, syllable or long utterance, as well as at pauses.

Additionally, all facial gestures are described with its known attributes and parameters as given in Table 2.2.

Attributes column contains characteristics important for either the single modality or the single gesture (aspects, actions, and presence/absence). Some examples are: direction, amplitude, duration etc.. If the facial gesture is not characterized with any attribute, the cell is left empty. For the given attributes, parameters providing values that certain attribute can achieve (e.g. a direction of the eyebrows raise might be up, central or down) are given.



Modality	*Signal*	*Attributes {Parameters}*

HEAD	nod	direction {*left, right, up, down, forward, backward, diagonal*}
		amplitude {*big, small*}
		velocity {*slow, ordinary, rapid*}

	postural shift	
	swing	
	reset	
	overshoot nod	

EYEBROWS
(left, right)
(inner, medial, outer)
	raise	direction {up, central, down}
		amplitude {wide, small}
		velocity {slow, ordinary, rapid}

	frown	

EYELIDS
(left, right)
(upper, lower)
	blinking	velocity {*rapid*}
		frequency {*frequent, normal, rarely*}

EYE (eyes gaze)	eye avoidance	duration
	eye contact	
	lowered gaze	
	rising gaze	
	saccade	velocity direction {*up-down, left-right*} magnitude duration inter-saccadic interval

Table 2.2: Facial gestures described with attributes and parameters

2.4.1 Head

Various head movements are among the most frequently used facial gestures. Attributes and parameters that characterize head movements are:

- direction: left, right, up, down, forward, backward and diagonal,
- amplitude: wide and small,
- velocity: slow, ordinary and rapid.

Amplitude and velocity are in inverted proportion. A movement with the big amplitude is rather slow and vice versa. Various combinations of these parameters, define several head movement types [4, 1]:

- **Nod.** That is an abrupt swing of the head with a similarly abrupt motion back [40]. The nod can be used as a conversational signal (e.g. to accentuate what is being said), synchronized at the word level or as a punctuation mark. Typically, the nod is described as the rapid movement of the small amplitude with four directions: left and right, right and left, up and down and down and up.

- **Postural shifts.** That are linear movements of wide amplitude often used as a regulator [4]. Postural shifts occur at the beginning of the speech between speaking-turns [61] and at grammatical pauses [62] maintaining the flow of conversation. The synchronization with the verbal cues is generally achieved at the pauses of the speech.

- **Overshoot nod.** That is a nod with an overshoot at the return. The pattern looks like an "S" lying on its side [40]. It is composed of two nods – the first one is with bigger amplitude starting upwards, while the second one is downwards with smaller amplitude.

- **Swing.** That is an abrupt swing of the head without the back motion. Sometimes the rotation moves slowly, barely visible, back to the original pose, and sometimes it is followed by an abrupt motion back after some delay [40]. Possible directions are up, down, left, right and diagonal. It occurs at increased speech dynamics, when pitch is also higher, and on shorter words.

- **Reset.** It sometimes follows swing movement; and returns head in central position. The reset is a slow head movement. It can be noticed at the end of the sentence – the sentence finishes with the slow head motion coming to neutral position.

Another issue that can be observed within this modality is the base head position, or the orientation that can be towards or away from a listener, up or down etc., and which generally follows the gaze direction. For example, if the utterance is a statement, the head is positioned to look down as the speaker reaches the end of the sentence. The head direction may depend on a speaker-listener relationship, or can be used to point at something, but that type of gesture is out of the scope of this work.

2.4.2 Eyebrows

Eyebrow movements appear frequently as conversational signals or punctuators. When serving as punctuators, they are used to mark prolonged (hesitation, thinking and word search) pauses, both as eyebrow raise (eyebrows go up and down) and frown (eyebrows go down and up). Eyebrow raise is often used to accentuate a word or a sequence of words. Besides the direction, eyebrow movements are described with the amplitude and the velocity, and are closely related to pitch changes.

2.4.3 Eyelids

Eyelids determine the openness of eyes. Temporal closures of eyes happen quite frequently due to eye blinking, described as rapid closing and opening of one or both eyes which might happen in frequent, normal or rare periods. Apart from periodic blinks, which serve the physical need to keep the eyes wet, there are voluntary blinks. They appear in two roles, as punctuators to mark a pause synchronized with a pause, or as conversational signals to emphasize speech or to accentuate a word synchronized with a word or syllable [63, 24].

2.4.4 Eyes

Eyes play an essential role as a major channel of nonverbal communicative behaviour [60]. Various expressions can be reflected in eyes. Eyes can be in tears, red or dry, open or closed, showing clearly the state of the mind. For example, slightly narrow eyes during the talk when adding more precise information, or wide open eyes when asking for speaking turn [64].

Eyes interact in the face-to-face communication through the gaze direction or intensity, and the saccade. The saccade is a rapid intermittent eye movement from one gaze position to another executed voluntary by a human. It is characterized with several attributes [60]:

- Direction,
- Velocity,
- Magnitude or amplitude – the angle through which the eyeball rotates as it changes fixation from one position to another,
- Duration – the amount of time that the movement takes to execute, typically determined using a velocity threshold, and
- Inter-saccadic interval – the amount of time which elapses between the termination of one saccade and the beginning of the next one.

The natural saccade movement, usually in the direction up-down or left-right, rarely have a magnitude greater than 15 degrees, while the duration and the velocity are functions of its magnitude [60].

Among others, the eye gaze is used to signal the search for feedback: during an interaction, look for information, or help regulate the flow of conversation [15]. It follows the same rules as head movements for speaking turns. However, some cultural differences are found in the amount of gaze allowed [4]. Gaze can be classified into four primary categories depending on its role in the conversation [65, 66]:

- **Planning.** Corresponds to the first phase of a turn when the speaker organizes thoughts,
- **Comment.** Accompanies and comments speech, by occurring in parallel with the accent and emphasis,
- **Control.** Controls the communication channel and functions as a synchronization signal and
- **Feedback.** Used to collect and seek feedback.

Aversion of gaze (the speaker looking away from the listener) happens at the beginning of an utterance, signaling that a person is thinking while speaking as opposed to listening, or at the hesitation pause, when thinking what to say [67].

Eye contact (the speaker is steadily looking toward to the listener for a period of time) occurs at the end of an utterance (when passing speaking turn), during pauses in the speech or at the beginning of a phrase boundary pause (the pause between two grammatical phrases of the speech).

Eye avoidance and eye contact follow the same rules as head movements for speaking turns. The level of gaze falls at the hesitation pause, while it rises at the end of an utterance in order to collect feedback from the listener or at the beginning of a phrase boundary pause.

Chapter 3

Proposed method for generating visual prosody

A key issue in speech driven facial gesturing is to find a mapping between speech signal and facial gestures. As explained earlier, it is a complex task, since there is no clear connection between the speech signal and occurrence of gestures. Additionally, with the real-time requirement the challenge arises, not only because of processing constraints, but also because all calculations need to be done using only the preceding speech signal (offline implementations often use both, preceding and subsequent information).

In this chapter, a method for generating speech related facial gestures in real-time is proposed. It is applicable to virtual characters, and based on available knowledge on nonverbal facial communication. The idea is to include various issues that are considered important for visual prosody in such a way as to obtain statistically correct facial movements in real-time. In this work, facial gestures correlated with prominent parts of speech are selected using a data-driven approach (explained in Section 3.3), while punctuation and prolonged pauses are covered using a rule-based approach (Section 3.4). Additionally, the results of both approaches are fine-tuned and further shaped using database statistics (Section 3.5). Hence, this method presents a hybrid approach to speech driven facial gesturing. Before explaining the method, Section 3.1 briefly lists included facial gestures and their properties as used in the video database with annotated facial gestures since both of them are important for method understanding, and Section 3.2 gives method overview.

3.1 Included facial gestures

Based on the information found in literature, e.g. [19, 40, 4, 24], presented in *Section 2.3*, and summarized in *Section 2.4* the following set of gestures are included in this work:

- **Nod,** used as a conversational signal to emphasise what is being said, synchronized at the word level, or synchronized with the pause when thinking or hesitating.
- **Swing,** used in a similar way to the nod, to emphasise what is being said.
- **Reset,** used sometimes after the swing movement to return the head in the neutral position, as well as at the end of the sentence.
- **Eyebrow movements,** used to accentuate a word or a sequence of words (e.g. eyebrow raise), or when hesitating or thinking (e.g. eyebrow frown or raise).
- **Eye blinking,** used as punctuator synchronized with a pause, as a conversational signal to emphasize the speech, often in combination with head and eyebrow movements, or as manipulator.
- **Gaze away,** used at the beginning of a talk, at the hesitation pause or when thinking what to say.
- **Eye contact,** used at the end of an utterance in order to collect feedback from the listener and as punctuator synchronized with the pause.

Facial gestures are characterized by various properties. Except for the gaze and the reset, those properties are extracted from a database with annotated facial gestures which was available as a starting point.

The database was built by Karlo Smid as a part of his Master Thesis [1]. The database consists of news presentations only, thus virtual humans built using it are suitable for the use as virtual news presenters, which generally do not take part in the conversation, but are acting only as presenters.

Annotations include following gesture properties:

- **Type:** blink, eyebrow movement, nod and swing,
- **Subtype:** eyebrow raise or frown; nod up, down, right or diagonal; swing up, down, right or diagonal,

- **Start and end time,** in milliseconds (ms),
- **Amplitude,** in Mouth-Nose Separation units (MNS0).

Used information on the gaze and its aversion is based on the work in [60].

Examples of gestures, the eyebrow movement occurring on the emphasized word and an eye blink appearing as punctuator are shown in Figure 3.1 and Figure 3.2 respectively.

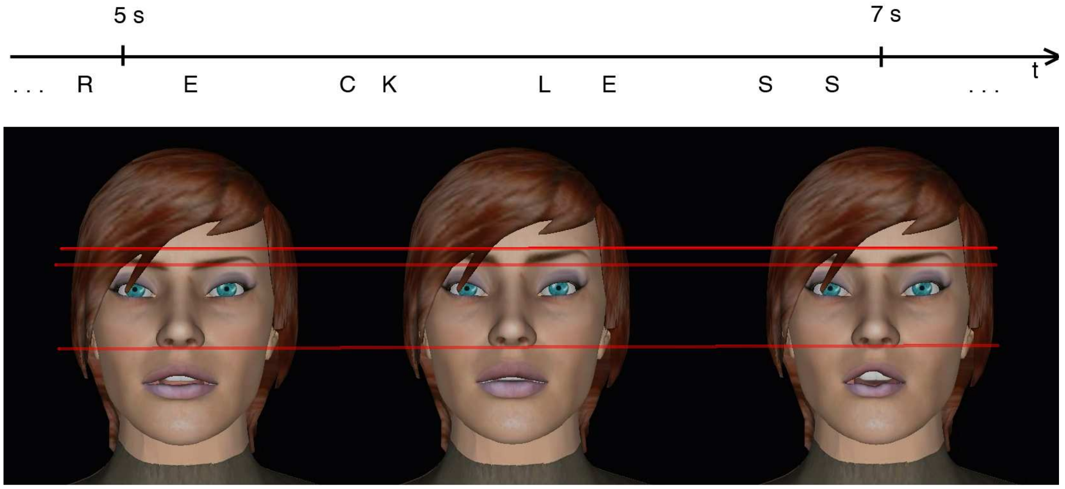


Figure 3.1: Eyebrow movement appearing on emphasized word

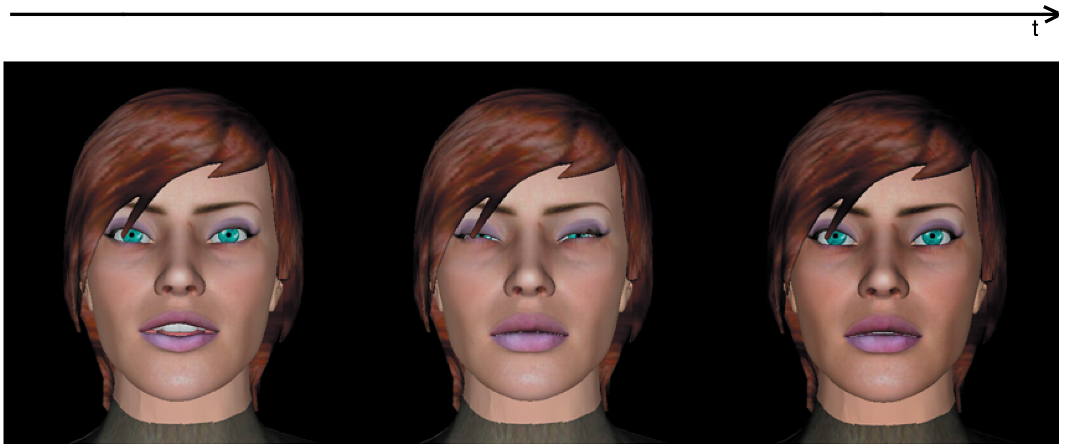


Figure 3.2: Eye blink appearing as punctuator

3.2 Method overview

The method for the mapping of speech signal to facial gestures is performed in real-time using a hybrid data-driven and rule-based approach. It runs in two parallel steps, data-driven and rule-based, whereas the output of each step is further shaped and fine-tuned with statistics derived from the database (Figure 3.3). Data-driven, Rule-based and Statistics modules are described in Sections 3.3, 3.4 and 3.5 respectively.

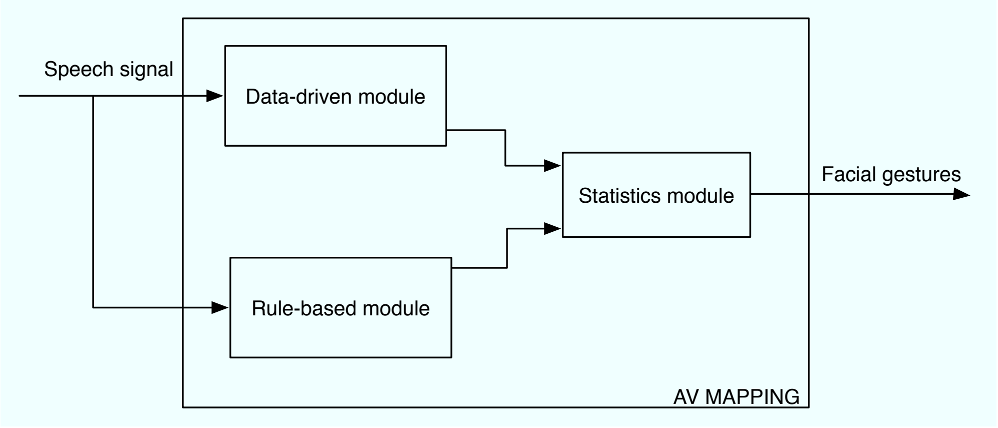


Figure 3.3: Simplified view of the proposed hybrid method

To use only a data-driven approach, a lot of training data is needed to properly generate various gesture types or subtypes from the speech, whilst, a rule-based approach might give predictable results in a long run.

In this work, the first approach to this problem was using only a statistical method followed by the work [68] as described in [58], which is actually an advanced case of the rule-based method. Rules and the statistical model for punctuators and thinking and word-search pauses were implemented. Next, whilst working on extending the work to include prominent parts of the speech, for which temporal changes of fundamental frequency and intensity are important, Hidden Markov Models (HMMs) were tried out as a possible solution. HMM is a statistical framework that takes into the account the temporal relationship, whereas the context information is represented by state-transition probabilities, and it is often used to model similar problems like in this work. There are several works, e.g. [54, 55, 56, 59], that used HMMs for alike audio to visual mapping, and that obtained promising results. In this work, HMMs were first used to classify gestures into four basic types: blink, eyebrow movement, nod and swing. Those gestures were additionally fine-tuned by the rules generated using the information found in literature on nonverbal communication. Additionally, statistics from the training database was used to characterize each facial gesture with its subtype, amplitude and duration as described in [69]. Based on the promising results that were obtained, the proposed method uses similar hybrid approach.

The hybrid approach has been chosen in this thesis as a possible optimal solution and a balance between severe real-time requirements and a complex timing relation between the speech features and the occurrence of gestures. The key point when designing the solution was to create statistically correct facial gesturing, which generally follows underlying speech signal within given real-time requirements.

With the real-time requirement and possibility to analyse only the preceding speech, a set of the rules to apply was limited. For the same reason, semantic analysis of the speech without having the whole utterance is not trivial either, hence was not used in this work. Non-semantic speech features that are available in real-time are generally connected to loudness (intensity) and pitch (F0) calculations, as well as to detecting disfluencies in the speech.

Acoustic features used are fundamental frequency and intensity (RMS, root mean square, amplitude), and corresponding delta and acceleration coefficients.

3.3 Data-driven module

The goal of this module is to determine the emphasized segments of the input speech. As it is known from literature (*Section 2.3.1*), facial gestures used to emphasize what is being said are mainly various head and eyebrow movements connected to changes in fundamental frequency and intensity. Based upon that, this module includes nods and swings in various directions, and eyebrow movements, in combination with eye blinking.

The core of this module are HMMs. The idea is to model the observed speech signal, and that is five consecutive frames or 80 ms of speech prosodic parameters. The number of frames is determined, and empirically tested, with two constraints: not too big in order not to lose the real-time performance, and not too small in order to be able to capture the relationship between facial gesture and changes in pitch and intensity, as it is shown in [25, 26, 27, 28, 29].

Having that in mind, HMMs are modelled with four states and left-right transition probabilities, where states represent speech frames. Transition probabilities are set in such a way that the most probable move is from one speech frame to the following one, with the small probability for staying in the same state, i.e. for spending more time in one of the frames.

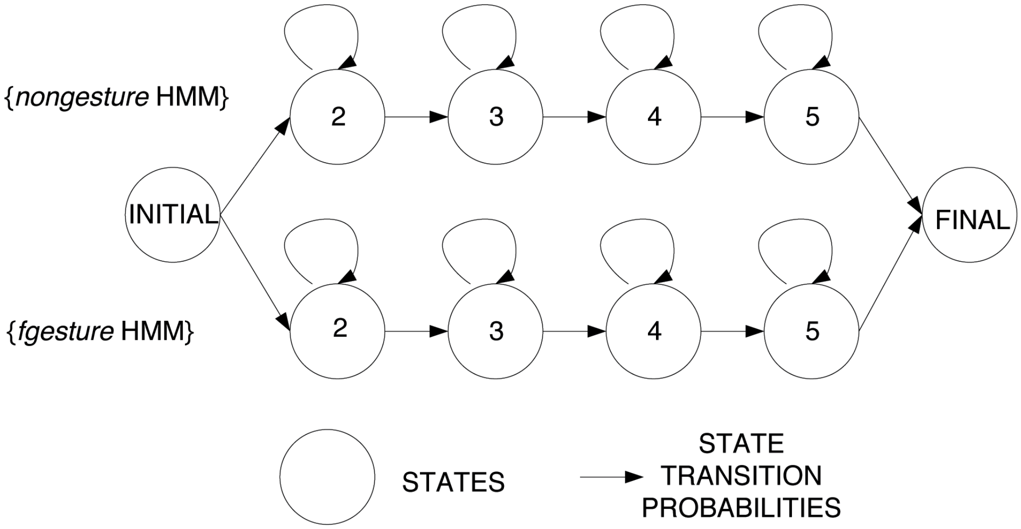


Figure 3.4: Two HMMs used

Since this module, i.e. HMMs are used to distinguish between emphasized segments of the input speech and the neutral speech, two HMMs are modelled – one for the neutral speech (*nongesture HMM*), and one connected to emphasized speech (*fgesture HMM*), as shown in Figure 3.4. Those HMMs share initial and final states, and that will be explained in *Section 4.1.2*, together with other details.

Figure 3.5 shows the basic idea of HMM training and recognition process. Used HMMs are indicated on the figure with M_G for the *fgesture HMM* (emphasized speech), and with M_N for the *nongesture HMM* (neutral speech). They are generated in the training process based on the segments of five frames, indicated in the figure as rectangles. Those segments consist of five frames of acoustic features and are called observations. HMMs generated in the training process are used in the recognition process – given a novel speech, i.e. observation in the form of the acoustic features from five frames, the HMM with the maximum probability is chosen. It means that each five frames of a novel speech is categorized either as the emphasized or as the neutral speech. More details on how HMMs were used in this thesis will be given in the *Chapter 4*, while next subsection gives short overview of Hidden Markov Models and the way they are used in this work in order to help understanding of HMM connected issues.

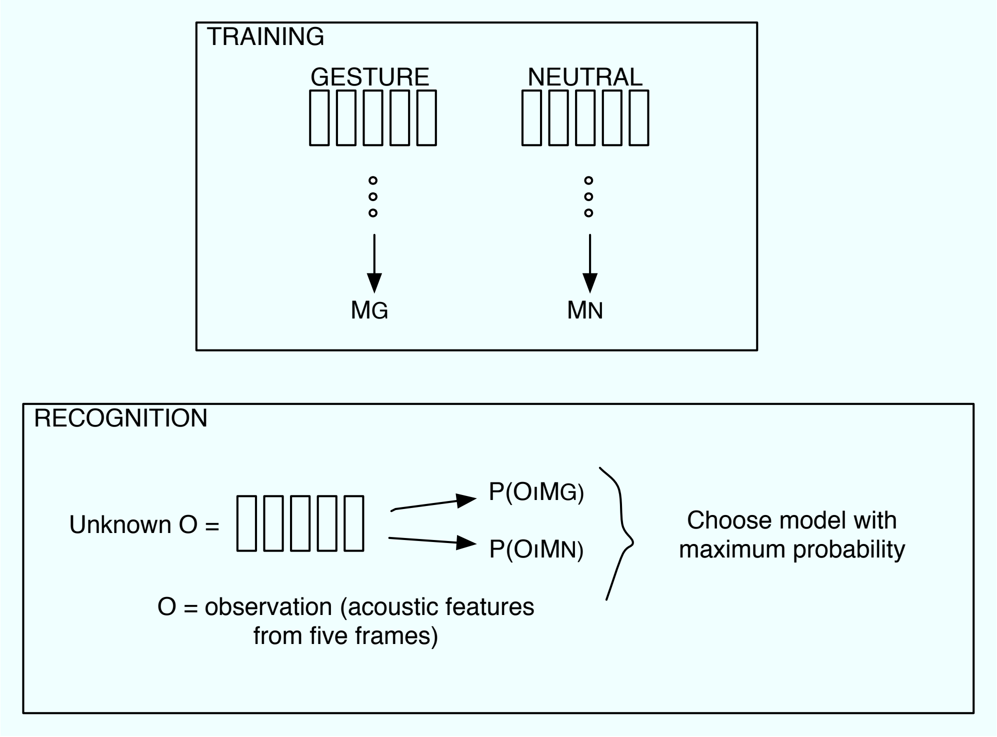

TRAINING
GESTURE
NEUTRAL
MG
MN
RECOGNITION
Unknown O =
P(OıMG)
P(OıMN)
Choose model with maximum probability
O = observation (acoustic features from five frames)

Figure 3.5: Facial gesture HMM, training and recognition process

3.3.1 Hidden Markov Model basics

A Hidden Markov model [70] is a statistical model where the system being modelled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters, from the observable parameters, based on this assumption. The extracted model parameters can then be used to perform further analysis, for example for speech recognition applications. HMMs are widely used in speech based applications, since they take into the consideration context cues that are inherent in the speech signal, unlike e.g. neural networks.

In a *Hidden* Markov model, the state is not directly visible, but output, dependent on the state, is visible. Although the states are hidden, there is the physical meaning attached to the states. Though, it is important to determine what does the HMM state represents, i.e. what is the basic unit for recognition. In this work it is generally assumed that the speech signal contains information about facial gesture encoded as a sequence of one or more symbols. Therefore, the continuous speech waveform is first converted to a sequence of equally spaced discrete parameter vectors. This sequence of parameter vectors is assumed to form an exact representation of the speech waveform on the basis that for the duration covered by a single vector, the speech waveform can be regarded as being stationary.

Elements of an HMM

An HMM is characterized by the following, as given in [70]:

- **N,** the number of hidden states in the model, $S = \{S_1, S_2, ..., S_N\}$. The state at time t is denoted as q_t.
- **M,** the number of distinct observation symbols per state, $V = \{v_1, v_2, ..., v_M\}$.
- $\mathbf{A} = \{a_{ij}\}$ the state transition probability distribution,
  $aij = P[q_{t+1} = S_j | q_t = S_i], i \leq i, j \leq N.$
- $\mathbf{B} = \{B_j(k)\}$, The observation symbol probability distribution in state j,
  $b_j(k) = P[v_k \text{ at t} | q_t = S_j], 1 \leq j \leq N, 1 \leq k \leq M.$
- $\Pi = \{\pi_i\}$**,** The initial state distribution where
  $\pi_i = P[q_1 = S_i], 1 \leq i \leq N.$

Having N, M, A, B and Π, the HMM can be used to generate the observation sequence $O = O_1, O_2, ..., O_T$, where O_t is one of the symbols from V, and T is the number of observations in the sequence. The HMM specification includes parameter set denoted as $\lambda(A, B, \Pi)$.

Three basic problems for HMMs

There are three fundamental problems to be solved when designing HMMs as given in [70]:

- **1. problem:** An evaluation of the probability of a sequence of observations given a specific HMM. Given an observation O, and the HMM description λ, how to efficiently compute $P(O|\lambda)$? *Solution:* Forward-backward algorithm.
- **2. problem:** A determination of the best sequence of model states. Given an observation O, and the HMM description λ, how to choose a corresponding state sequence $Q = q_1, q_2, ..., q_T$, i.e. an optimal corresponding state sequence, best explaining the observation? *Solution:* Viterbi algorithm.

- **3. problem:** An adjustment of model parameters so as to best account for the observed signal. How to adjust model parameters λ to maximize $P(O|\lambda)$? *Solution:* Baum-Welch Expectation Maximization algorithm.

If those three problems are successfully solved, the HMM can be applied to selected problems, in speech recognition and similar areas. It is assumed that model parameters are estimated from the (large) collection of training data that is assumed to be representative of the recognition task.

Facial gesture recognition

In the speech recognition field, HMMs are often used for isolated word recognition as described in [71]. The idea is to recognize a word that belongs to a small vocabulary. Each word in the vocabulary can be modelled as a succession of phones, where each phone is modelled by a single HMM state. The time that stochastic process stays in one state is known as a state duration, and it is measured as the number of observations emitted from the process. The process is said to "stay" in the state j if a transition is made from j to itself.

For every word of a W word vocabulary, it is needed to design a separate N state HMM. A speech signal of a given word is represented by the time sequence of coded spectral vectors. It is done by using spectral codebook with M unique spectral vectors, whereas each observation is an index of the spectral vector closest to the original speech signal. For every vocabulary word, a training sequence consists of the number of repetitious of sequences of codebook indices of the word.

The procedure is following:

1. **HMM designed.** Build word models using solution to problem 3. It will give optimally estimate parameters for every word model.

2. **HMM optimized and studied.** Develop understanding of the physical meaning of model states using solution to problem 2. It includes segmenting training sequences of every word into states, studying properties of spectral vectors in order to make refinements of the model so as to improve its capability of modelling the spoken word sequences.

3. **Word recognition.** Recognition of unknown word using solution to problem 1. Score each word model based upon the given test observation sequence and select the word whose model score is highest.

In this work, isolated word recognition problem is applied to gesture recognition, whereas a vocabulary consists of gestures instead of words. For every gesture in vocabulary, an HMM is designed. However, here are only used two gestures - neutral gesture denoting neutral speech, modelled by *nongesture HMM*, and gesture denoting emphasized speech modelled by *fgesture HMM*. To recognise some unknown gesture, the likelihood of each model generating that gesture is calculated and the most likely model identifies the gesture.

3.3.2 From emphasized speech to facial gesture using statistics

In the case when the *fgesture HMM* is activated, the database statistics is used to determine the type of gestures to appear as well as other properties – subtype, amplitude and duration. Since on the accentuate speech often appear not only head or eyebrow movements, but also both of them in combination, and eye blinking, it is taken into the consideration when building statistics. Accordingly, the possible gesture output might be: eyebrow movement, swing, nod, eyebrow movement and swing, or eyebrow movement and nod, with or without the eye blinking. *Section 3.5* gives more details on building the statistic and its values.

3.4 Rule-based module

The main idea of this module is to apply a set of rules on a novel speech in order to generate corresponding facial gestures.

Based on knowledge on nonverbal facial communication found in literature, presented in *Section 2.3* and summarized in *Section 2.4*, the set of rules is derived:

- **Gaze away, eyebrow movements and the nod during prolonged pauses.** Gaze up or down, and correspondent head and eyebrow movement are added when a long pause is detected. As prolonged pauses thinking, word-search and hesitation pauses are observed.

- **Eye blinking and eye contact as punctuator.** Eye blinking is added, and mutual gaze is kept when a short pause is detected.

- **Gaze away at the beginning of a talk.** An eye aversion, i.e. the saccadic eye movement as it is referred in [60], is added when starting talking and organizing thoughts.

- **Eye contact and head reset at the end of a talk.** When possible, the end of the talk is detected, eye contact is established, and the head is returned to the neutral position.

Last two rules regulate the flow of the talk. In this work, there are no more similar rules included since the focus is on virtual news presenters. This means that they incorporate only a talking mode and not listening, when regulating the flow of the talk is applicable only when both modes exist. Similarly, considering a news presenter, not many hesitation pauses are present. However, the rule for such pauses is included in the case of the method adaptation for different purposes – prolonged pauses have an important role in generating visual prosody.

A pause detection algorithm is based on the root mean square amplitude value, and long and short pauses are differentiated, as well as possible end of the talk – since working in real-time, it cannot be detected with certainty when the talk ends. Each frame of the novel speech can be marked as a potential pause in the speech using the obtained amplitude value for the frame and thresholds set in the initial frames. If the pause is identified in 16 consecutive frames, i.e. 256 ms, it is classified as a short one. When the pause in the speech is longer than 32 frames, i.e. 512 ms, it is classified as a long pause. The possible end of the talk is set after having 64 silent frames, i.e. 1024 ms. These values are set based on the trial and error method since the length of pauses is individual, both in terms of the speaker characteristics and the nature of the talk. Still, in literature some empirical values exist for various kinds of pauses that occur during the speech, as briefly presented *Section 2.2.1.*

Additionally, this module includes several other rules which add to the naturalness of the facial gesturing:

- **Gaze direction in accordance with head movements.** In gaze following animation model, the eyes of the virtual human are moving in the opposite direction of a head movement, whereas an amplitude of the head movement needs to be smaller than the defined threshold.

- **Eye blinking as manipulator.** In addition to voluntary eye blinks, implemented are periodic eye blinks based on the values given in [4]. If an eye blink does not happen according to other rules within 5 s, it will be manually added.

- **Small periodic head movements in periods when there are no nods or swings.** Since the head is never still, random head movements are added in accordance with the database statistics to fill "empty" periods.

- **Periodic gaze away and mutual gaze.** Based on the work in [60], mutual gaze and gaze away interchange with the timer set on approximately 3100 ms and 900 ms respectively. In the case when gaze away is added while head or eyebrow movement is on, its direction is copied to the gaze away movement.

For each chosen gesture, the subtype, amplitude and duration are determined using the database statistics.

Addition of a new gesture to the set of chosen gestures for a specific moment implies that no other gesture incorporating the same facial part, i.e. modality is still running. The same is valid for adding gestures using the data-driven approach. For example, if the application of rules results with a nod to be included, and previously added swing has not yet finished, the nod is not added. Similarly, adding the eye blink while gazing away is not possible.

3.5 Statistics module

The statistics module consists of the statistical data extracted from the database. It is applied after the data-driven and rule-based modules to shape their results, and to add details needed to describe facial gestures detailed enough, so they could be rendered on the screen. Statistics used here is generated in a similar manner as in [68]. The way it is used, differs if it is applied after rule-based module or after data-driven module, as described next.

3.5.1 Statistics applied after rule-based module

After applying rules on the novel speech, a type of a gesture and its start time is known. Further properties, i.e. the gesture subtype, amplitude and duration are obtained using statistics. This process aims to produce the global statistics based on the existing database for frequency of occurrence of facial gesture, amplitude in a Mouth-Nose Separation unit, and duration of various gestures in milliseconds.

The statistical data are grouped according to the identified gesture type, i.e. eye blinks, eyebrow movements, nods and swings. Calculations are based on the information obtained from the gesture annotation of all training video clips. For every gesture type, its properties are characterized in the following way:

- **Subtype,** as statistical probability of occurrence.

- **Amplitude,** as cumulative frequency histogram distributions.
- **Duration,** as cumulative frequency histogram distributions.

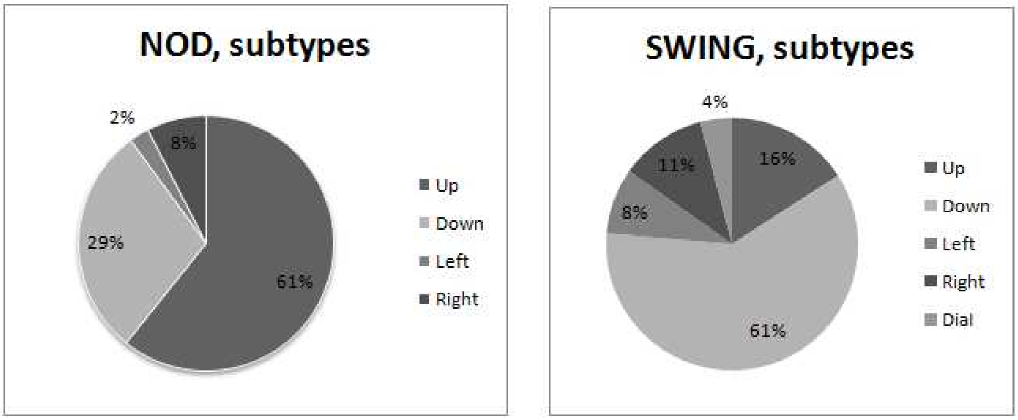


Figure 3.6: Probability of occurrence for nod and swing subtypes

Probability of occurrence for particular gesture subtype is simply calculated by dividing the number of particular gesture occurrences (e.g. the nod down) with the size of corresponding gesture group (e.g. nods) as found in the database. It means, that having a total of e.g. 358 annotated nods in the database, and 104 nods up, the probability of occurrence for the nod up is 29%. Figure 3.6 shows the probability of occurrence for nod and swing subtypes. In the database, for the nod type, no diagonal nods were found, although they were mentioned in literature. As eyebrow movements, generally are differentiated eyebrow raises and frowns. However, in this specific case there are no frowns presented. It is because the database deals with news presentations, and presenters generally do not use frowns. On the other hand, for eye blinking, there are no subtypes specified.

Duration and amplitude data for for the gesture type are grouped in intervals in order to produce their frequency distributions of occurrence, as in [68]. Interval values for duration and the amplitude depend on the amount of the training corpus – the more training data there is, the interval is narrower, since certain amount of data examples is needed for each interval. Based on cumulative frequency distribution values, a simple linear approximation of probability functions for duration and amplitude values is calculated for each interval. A uniformly distributed random number between 0 and 100 is generated for the amplitude or duration value in one of the non-uniform intervals.

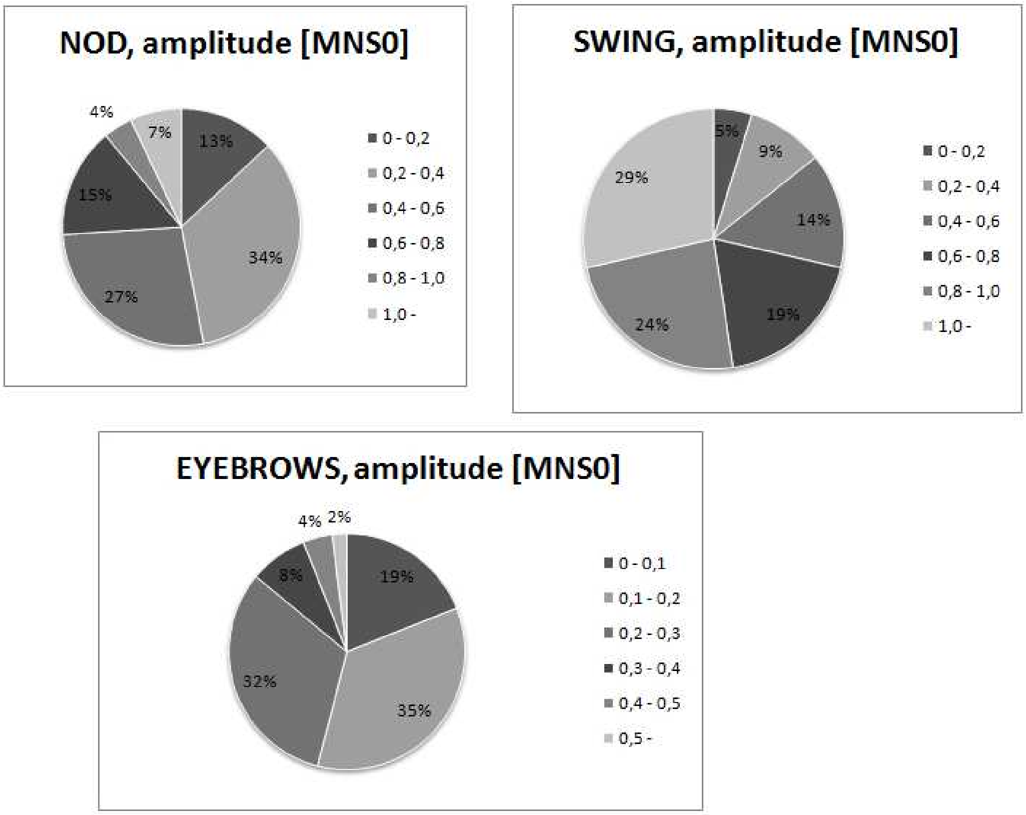


Figure 3.7: Frequency of occurrence for facial gesture amplitude intervals

In this thesis, the range of possible values for the amplitude and duration calculation is divided into 6 intervals. For each interval, gestures with duration and amplitude fitting into that interval are counted, and frequency of occurrence for each group is calculated based on the total number of specific gesture, and the number of the gestures in the interval. On the previous nod example, if there is 40 nods with duration between 0 and 300 ms, frequency of occurrence for that first group is 11%. Figure 3.7 shows frequency of occurrence for gesture amplitude intervals, while figure 3.8 shows frequency of occurrence for gesture duration intervals. The only exception is the eye blinking gesture, for which there is no amplitude property specified.

For other used gestures, i.e. the reset and gaze, which are not covered by the database statistics, the following is done.

Reset, i.e. the head movement that returns the head back to the central position, is applied after the swing in the case there is no other head movement going on, and when end of a talk is detected.

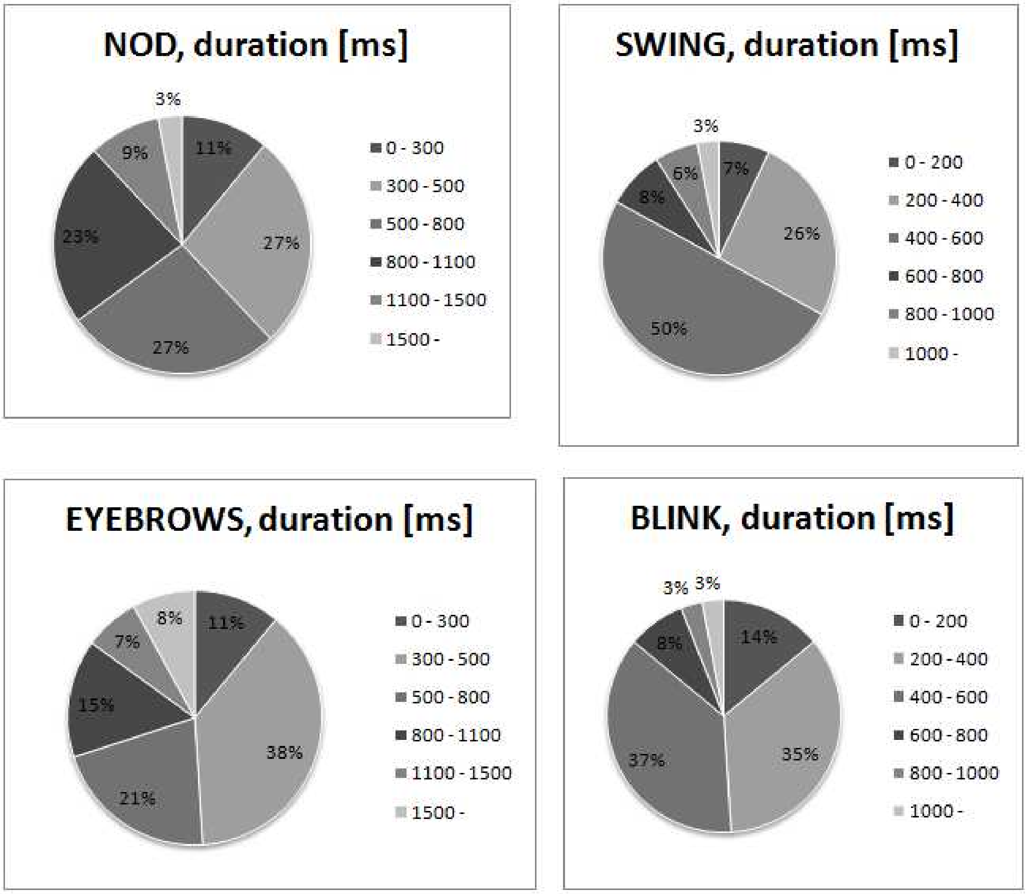


Figure 3.8: Frequency of occurrence for facial gesture duration intervals

For gaze, the information on a statistical eye movement model based on both empirical studies of saccades and acquired eye movement data is used as found in [60]. Mutual gaze and gaze away interchange with the timer set on approximately 3100 ms and 900 ms respectively. In [60], possible direction values are up, down, left, right and diagonal, whereas up-down and left-right movements happened more than twice as often as diagonal movements. However, the used animation model does not support diagonal movements, so up-down and left-right movements are equally probable. Still, if there is a head movement happening at the same time, the saccadic eye movement follows its direction. The magnitude value is adjusted to the model used, and does not exceed 15 degrees.

3.5.2 Statistics applied after data-driven module

The data-driven module provides only the possible position of gestures to be applied to emphasize the speech. Hence, the statistical data is used to determine the type of gestures to appear as well as its subtype, amplitude and duration.

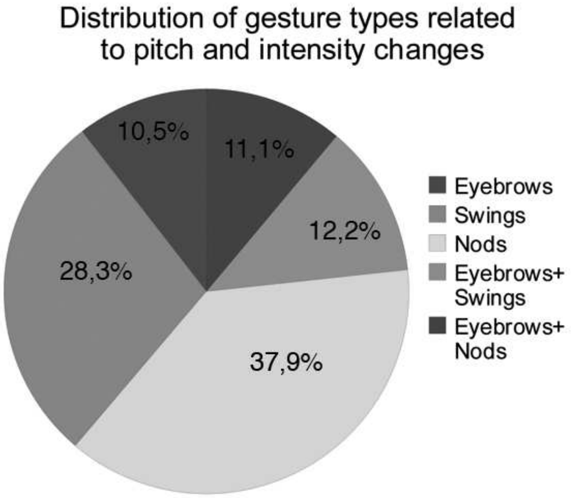


Figure 3.9: Probability of occurrence for gesture types connected to changes in pitch and intensity

On the accentuated speech, eye blinking, head and eyebrow movements happen often in combination with each other, so the possible gesture output might be: eyebrow movement, swing, nod, eyebrow movement and swing, or eyebrow movement and nod, with or without the eye blinking. Figure 3.9 shows the probability of occurrence for gesture types connected to changes in pitch and intensity.

Eye blinking occurs together with these gestures approximately in one third of examples in the database, so it is accordingly added to the set of gestures that will be potentially applied to the virtual human. In this context, potential denotes that some gestures might not be rendered in the case there is already the gesture being animated that "occupies" the same part of the body.

Next, the statistical data for the subtype, amplitude and duration, as described in the previous subsection, is applied.

Chapter 4

System for speech driven facial gesturing for virtual characters

This chapter describes a system for automatic speech driven facial gesturing based on the previously presented hybrid method. It takes as input a novel speech, analyses it, and produces facial gestures which are used together with lip movements to animate the MPEG-4 compatible face of the virtual character. This chapter includes basic implementation information, as well as details on the method itself, relevant for complete understanding and building of the method. As the Figure 4.1 shows, the system consists of two phases: training and runtime, described in Section 4.1 and 4.2 respectively. All components are described with the level of detail suitable to the component's novelty and complexity. An example application is given in Section 4.2.2, and for better understanding, MPEG-4 facial animation is briefly introduced.

4.1 Training phase

A main purpose of the training phase is to generate resources needed by AV mapping method in the runtime phase. It consists of four main components, described later in this section. As input, the training database collected by Karlo Smid as a part of his Master Thesis [1] was used. It consists of 57 news video clips of 3 female and 2 male speakers, 13 minutes of total duration, and a frame rate of 15 frames/s. From the database, following data was extracted:

- **Audio clips,** corresponding to training video clips in PCM format (.wav file), with 16 kHz sample rate and 16 bit sample size, and

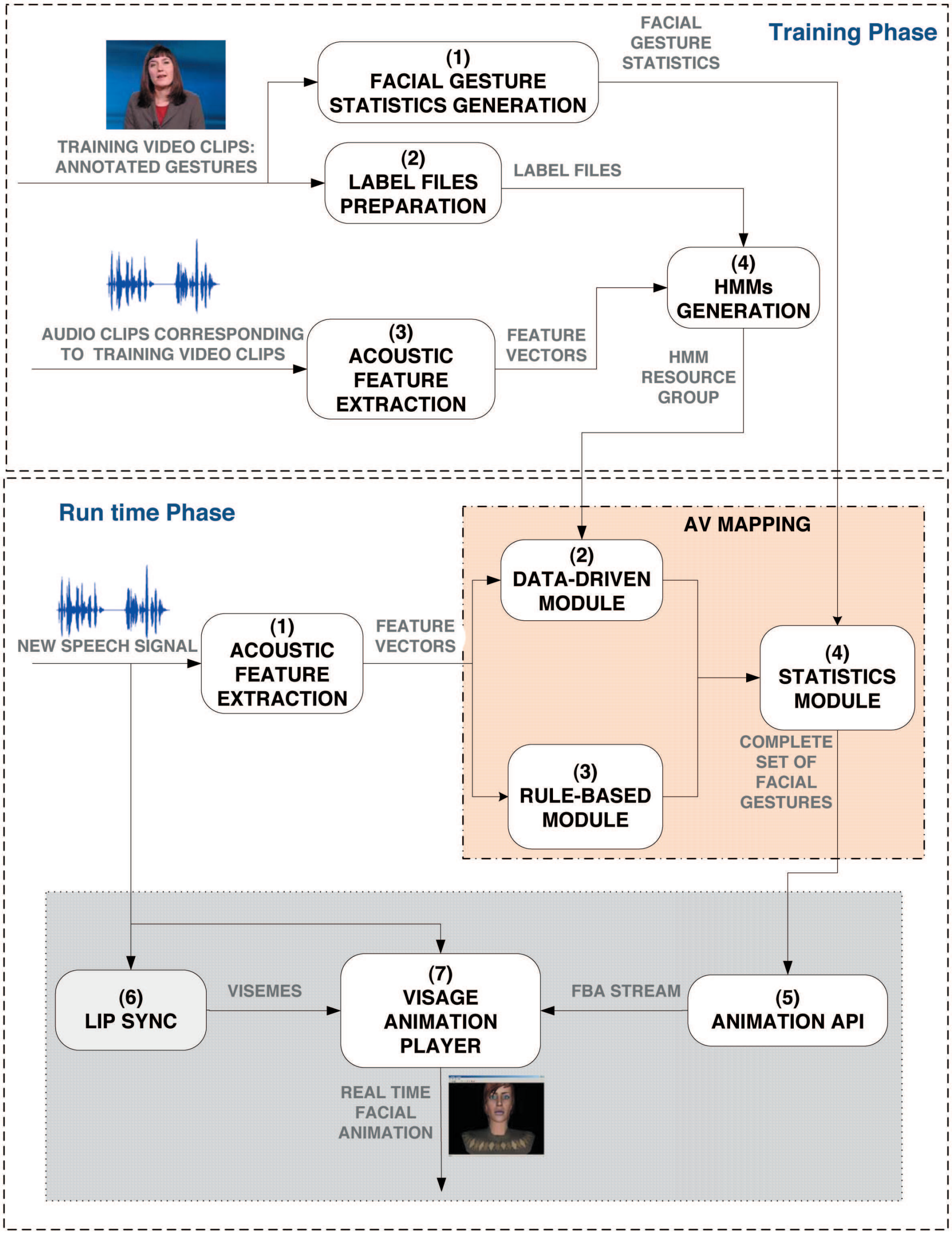
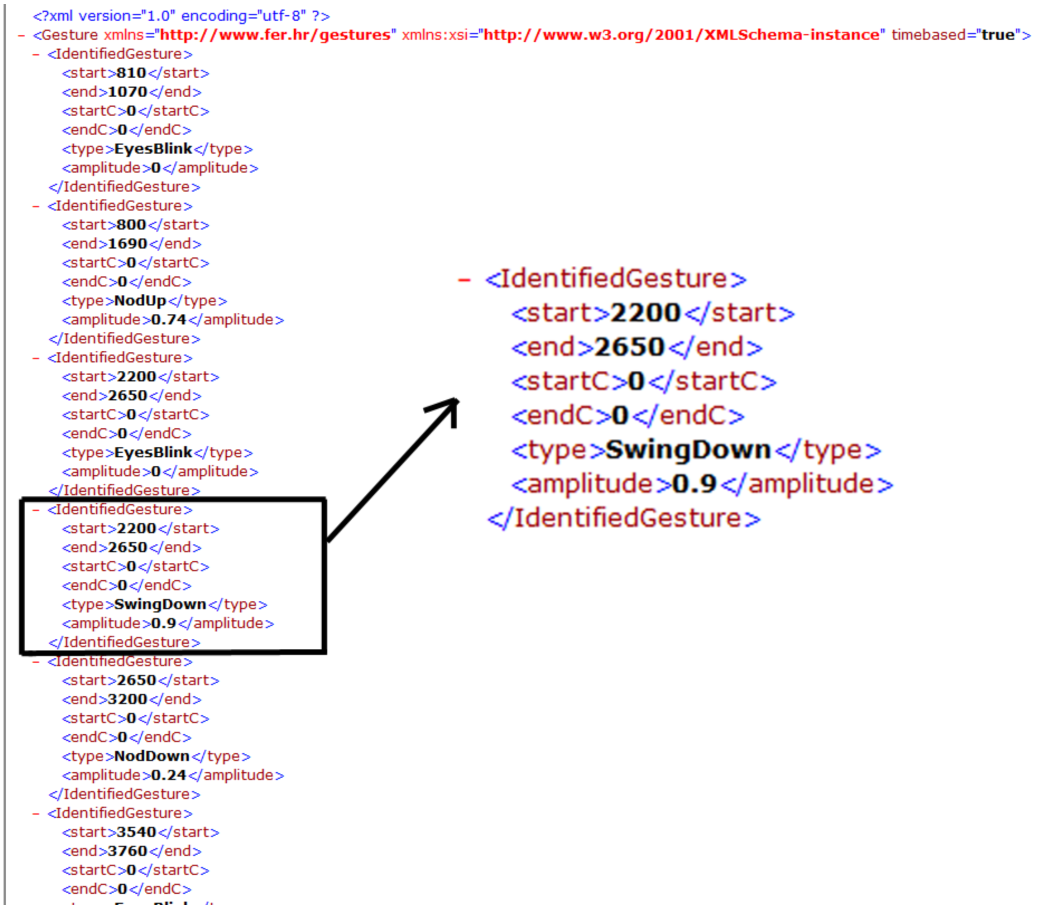

Training Phase
TRAINING VIDEO CLIPS: ANNOTATED GESTURES
(1) FACIAL GESTURE STATISTICS GENERATION
FACIAL GESTURE STATISTICS
(2) LABEL FILES PREPARATION
LABEL FILES
AUDIO CLIPS CORRESPONDING TO TRAINING VIDEO CLIPS
(3) ACOUSTIC FEATURE EXTRACTION
FEATURE VECTORS
(4) HMMs GENERATION
HMM RESOURCE GROUP
Run time Phase
AV MAPPING
NEW SPEECH SIGNAL
(1) ACOUSTIC FEATURE EXTRACTION
FEATURE VECTORS
(2) DATA-DRIVEN MODULE
(3) RULE-BASED MODULE
(4) STATISTICS MODULE
COMPLETE SET OF FACIAL GESTURES
(6) LIP SYNC
VISEMES
(7) VISAGE ANIMATION PLAYER
FBA STREAM
(5) ANIMATION API
REAL TIME FACIAL ANIMATION

Figure 4.1: Speech driven facial gesturing - a system overview

- **Annotated facial gestures,** corresponding to training video clips in XML format. Annotations were done manually by Karlo Smid, and include the gesture type, the subtype, start and end time in ms, and the amplitude in MNS0 units, as explained in *Section 3.1*, and shown in Figure 4.2..



```
  <?xml version="1.0" encoding="utf-8" ?>
- <Gesture xmlns="http://www.fer.hr/gestures" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" timebased="true">
  - <IdentifiedGesture>
      <start>810</start>
      <end>1070</end>
      <startC>0</startC>
      <endC>0</endC>
      <type>EyesBlink</type>
      <amplitude>0</amplitude>
    </IdentifiedGesture>
  - <IdentifiedGesture>
      <start>800</start>
      <end>1690</end>
      <startC>0</startC>
      <endC>0</endC>
      <type>NodUp</type>
      <amplitude>0.74</amplitude>
    </IdentifiedGesture>
  - <IdentifiedGesture>
      <start>2200</start>
      <end>2650</end>
      <startC>0</startC>
      <endC>0</endC>
      <type>EyesBlink</type>
      <amplitude>0</amplitude>
    </IdentifiedGesture>
  - <IdentifiedGesture>
      <start>2200</start>
      <end>2650</end>
      <startC>0</startC>
      <endC>0</endC>
      <type>SwingDown</type>
      <amplitude>0.9</amplitude>
    </IdentifiedGesture>
  - <IdentifiedGesture>
      <start>2650</start>
      <end>3200</end>
      <startC>0</startC>
      <endC>0</endC>
      <type>NodDown</type>
      <amplitude>0.24</amplitude>
    </IdentifiedGesture>
  - <IdentifiedGesture>
      <start>3540</start>
      <end>3760</end>
      <startC>0</startC>
      <endC>0</endC>
```

```
- <IdentifiedGesture>
    <start>2200</start>
    <end>2650</end>
    <startC>0</startC>
    <endC>0</endC>
    <type>SwingDown</type>
    <amplitude>0.9</amplitude>
  </IdentifiedGesture>
```

Figure 4.2: Example of the used facial gesture annotation in XML format

This phase results in:

- **Facial gesture statistics,** used in the runtime, i.e. in the statistics module,
- **HMM resource group,** used in the runtime, i.e. in the data-driven module.

Those outputs depend greatly on the database used – its size and the type of video clips in it.

Training of the system is generally time-consuming. However, for a typical usage, as in this work it is for news presenters, it needs to be done only once.

4.1.1 Description of system components

The training phase consists of four main components:

1. Facial gesture statistics generation,
2. Label files preparation,
3. Acoustic feature extraction,
4. HMMs generation.

Only the first component is used by the statistics module, while the other three are needed for generation of HMM resource group, used by the data-driven module based on HMMs.

For HMM training, HTK, *HMM toolkit* [72], was used. HTK is a set of library modules and tools available in C source form for speech analysis, HMM training, testing and results analysis. Although it is specifically developed for speech recognition, it can be used for numerous purposes that can be mapped to it. In this work, HTK is used for isolated gesture recognition, as explained in *Section 3.3.1*.

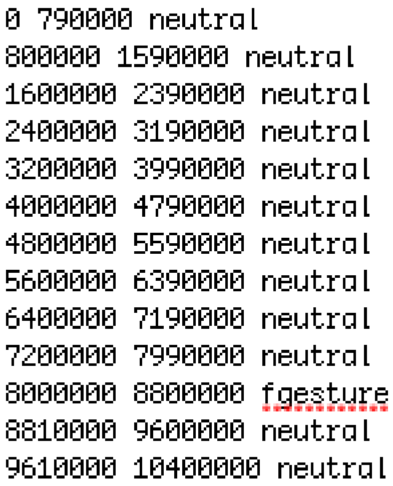
Facial gesture statistics generation

Based on the information found in the database, this component outputs facial gesture statistics. Details on how the statistics is calculated can be found in *Section 3.5*.

Label files preparation

Label files are used for Hidden Markov Models production and they are tightly connected to the HMMs used, thus the structure of label files is in accordance with HTK rules.

The speech underlying video clips is divided into segments, and each segment is given a name or a label. In this work, possible labels are *neutral* and *fgesture*, and the segments are chosen to be five frames long, which is 5 x 16 ms = 80 ms. HTK toolkit uses HTK time units for measuring time, and it is 100 ns. According to this, 16 ms equals to 10000 in 100 ns units. The set of labels associated with a speech file constitute a transcription, and each transcription is saved in a separate label file. The label file is actually the transcription of what is going on in a data sequence.



```
0 790000 neutral
800000 1590000 neutral
1600000 2390000 neutral
2400000 3190000 neutral
3200000 3990000 neutral
4000000 4790000 neutral
4800000 5590000 neutral
5600000 6390000 neutral
6400000 7190000 neutral
7200000 7990000 neutral
8000000 8800000 fgesture
8810000 9600000 neutral
9610000 10400000 neutral
```

Figure 4.3: Snippet of the label file in HTK format

This component takes as input a XML gesture annotation file corresponding to the speech file, and as the output generates the label file in HTK format (e.g. Alan1.xml → Alan1.lab). Figure 4.3 shows a snippet of the label file. It consists of:

- **Beginning time,** in HTK time units,
- **End time,** in HTK time units. It is equivalent to the beginning time with summed time for the length of five frames, i.e. $t_{end} = (t_{start} + 80)$ x 10000,
- **Label name,** typically corresponds to the name of HMM chosen. *fgesture* is written whenever in the XML annotation file, a mark for the swing, the nod or the eyebrow movement is found. As the beginning time, the time for the beginning of the corresponding gesture is taken. Otherwise *neutral* is written in steps of 80 ms.

Acoustic feature extraction

This component takes as the input speech files from the database and outputs acoustic feature vectors consisting of two prosodic features, fundamental frequency (pitch) and intensity (loudness). For speech segmentation, windowing and pitch calculation SPTK, *Speech Processing ToolKit* [73], is used.

The speech signal is first framed into 16 ms long frames, i.e. 256 samples, without the overlapping, using the following SPTK synopsis:
$frame - l256 - p256$,

where *-l* denotes the frame length, and *-p* the frame period. Next windowing is done:
$window - l256 - w1$,
with *-w 1* denoting a Hamming window.

A cepstrum method is used to calculate the pitch period values corresponding to the frames of input data. For unvoiced frames, the output value is zero, and for voiced frames, the output value is proportional to the pitch period. SPTK synopsis used is:
$pitch - s16 - l256 - t7.5 - L70 - H300 - e0.1$,
where the meaning of the arguments is the following:

-*s* — sampling frequency (kHz),

-*l* — frame length (samples),

-*t* — voiced/unvoiced threshold,

-*L* — minimal fundamental frequency to search for (Hz),

-*H* — maximal fundamental frequency to search for (Hz),

-*e* — small value for calculate log-spectral envelope.

Values for *-t, -L, -H* and *-e* are empirically determined and they suit the data used.

Intensity is calculated as root mean square (RMS) amplitude. The RMS value of a single speech frame is calculated as the square root of the arithmetic mean of the squares of the values for each sample, $x_1, x_2, ..., x_n$. Having 16 ms long frames, $n = 256$ samples, and the RMS value is given by:
$x_{rms} = \sqrt{\frac{x_1^2+x_2^2+...+x_n^2}{n}}$.

Once F0 and intensity values for the frame are known, they are written into HTK format parameter file. It consists of a contiguous sequence of samples written as 4-byte floating points, and preceded by a header which is 12 bytes long, and which contains the following data:

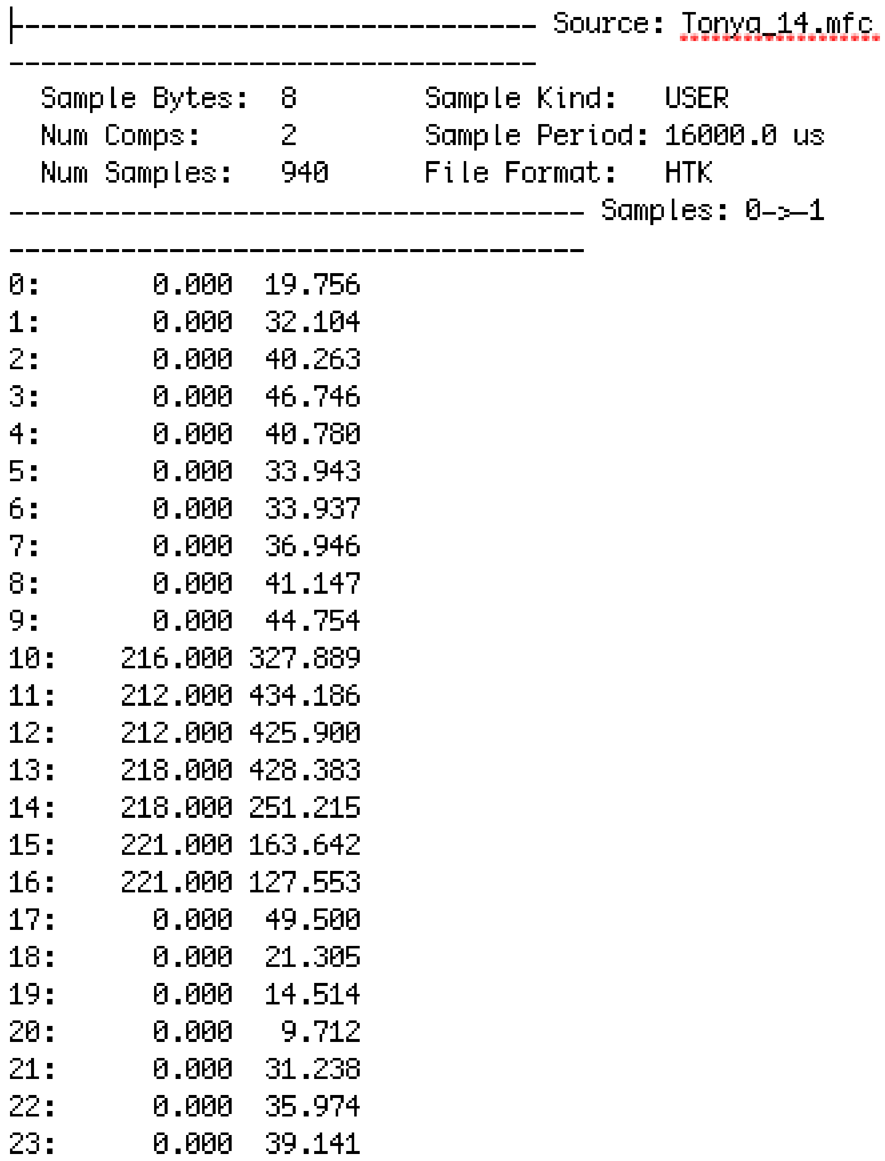
nSamples — the number of samples in the file (4-byte integer),

sampPeriod — the sample period in 100 ns units (4-byte integer),

sampSize — the number of bytes per sample (2-byte integer),

parmKind — a code indicating the sample kind (2-byte integer).

The parameter kind used is USER, which is user defined sample kind, and it consists of two acoustic features, fundamental frequency and intensity. Figure 4.4 shows a snippet of the HTK format parameter file with the speech signal written in the defined USER format (with F0 values in the first column, and intensity values in the second one).



```
|-------------------------------- Source: Tonya_14.mfc
--------------------------------
  Sample Bytes:  8       Sample Kind:   USER
  Num Comps:     2       Sample Period: 16000.0 us
  Num Samples:   940     File Format:   HTK
----------------------------------- Samples: 0->-1
-----------------------------------
0:       0.000  19.756
1:       0.000  32.104
2:       0.000  40.263
3:       0.000  46.746
4:       0.000  40.780
5:       0.000  33.943
6:       0.000  33.937
7:       0.000  36.946
8:       0.000  41.147
9:       0.000  44.754
10:    216.000 327.889
11:    212.000 434.186
12:    212.000 425.900
13:    218.000 428.383
14:    218.000 251.215
15:    221.000 163.642
16:    221.000 127.553
17:      0.000  49.500
18:      0.000  21.305
19:      0.000  14.514
20:      0.000   9.712
21:      0.000  31.238
22:      0.000  35.974
23:      0.000  39.141
```

Figure 4.4: Snippet of the parameter file in HTK format

HMMs generation

For modelling of HMMs, label files and corresponding acoustic feature vectors are needed. As the output, besides HMM models, a dictionary and a language model are created, and all together are called *HMM Resource Group*.

Section 3.3, and specifically *Subsection 3.3.1* paragraph on *Isolated word speech recognizer applied to facial gesture recognition*, describes the idea behind the use of HMMs in this work.

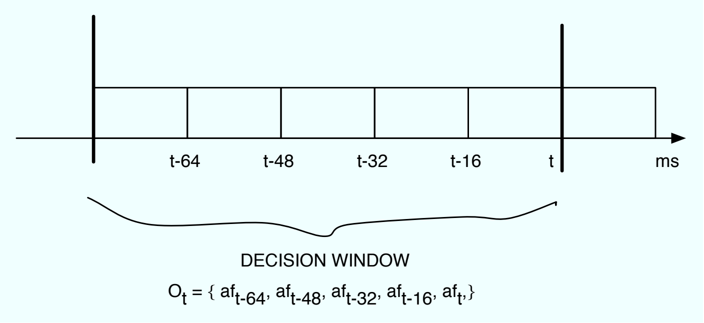


Figure 4.5: Decision window and observation

Hence, two gesture HMMs need to be modelled - *fgesture*, and *neutral*. A speech signal of a given gesture is represented by the time sequence of coded vectors. The parametric representations that are used are the fundamental frequency and intensity values, and their delta and acceleration coefficients, as they are connected with visual prosody (explained in *Section 2.3.1*). A sequence of speech vectors representing a gesture is actually an observation O, defined as

$O = o_1, o_2, ..., o_T$,

where o_t is the speech vector observed at time t. In the proposed solution, the observation at time t,

$O_t = \{af_{t-4}, af_{t-3}, af_{t-2}, af_{t-1}, af_t\}$

consists of five acoustic feature vectors, where each of them is composed of fundamental frequency, RMS amplitude, i.e. intensity, and corresponding delta and acceleration coefficients shown in Figure 4.5:

$af_t = \{F0_t, RMS_t, \delta F0_t, \delta RMS_t, \Delta F0_t, \Delta RMS_t\}$.

F0 and RMS are calculated in the Acoustic Feature Extraction component, while theirs delta and acceleration components are calculated automatically in HTK by giving _D mark for the former, and _A mark for the later when defining the HMM prototype model, which will be explained later.

Used HMM topology for both models is left-right where only transitions in forward direction between adjacent states are allowed. HMMs are modelled as continuous density models with four emitting states and left-right transition probabilities. In doing so, the observed speech signal is modelled, and that is 5 frames of speech prosodic parameters with states representing speech frames, without the overlapping. Since the frames are ordered in time, it is only possible to move from one speech frame to the following one, whereas self transitions are only possible within one frame.

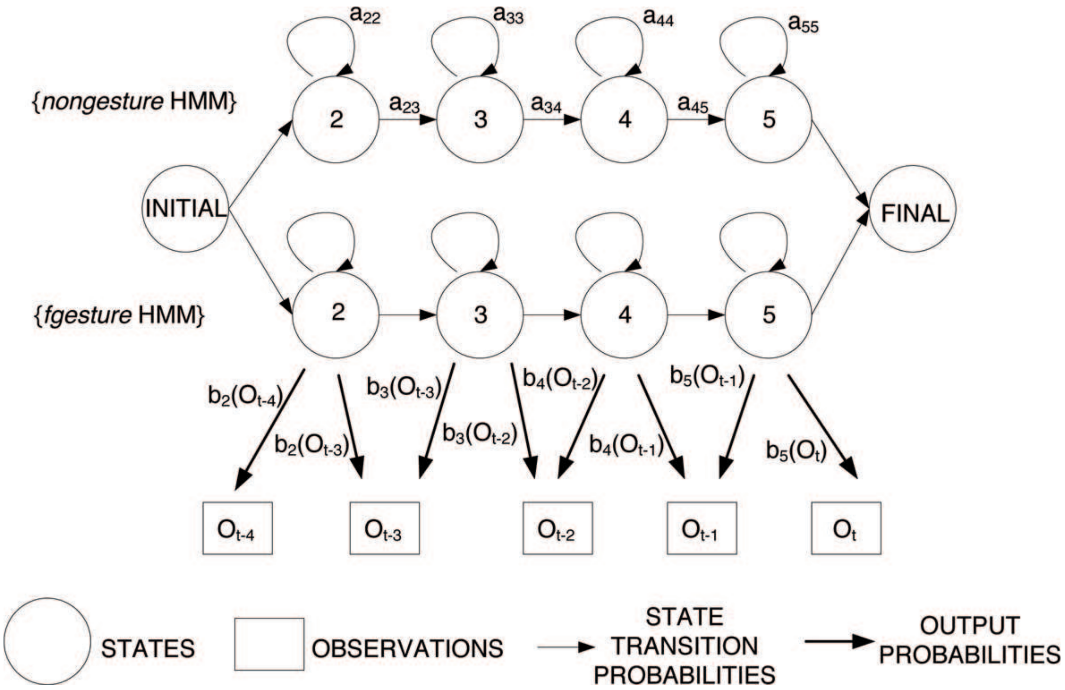
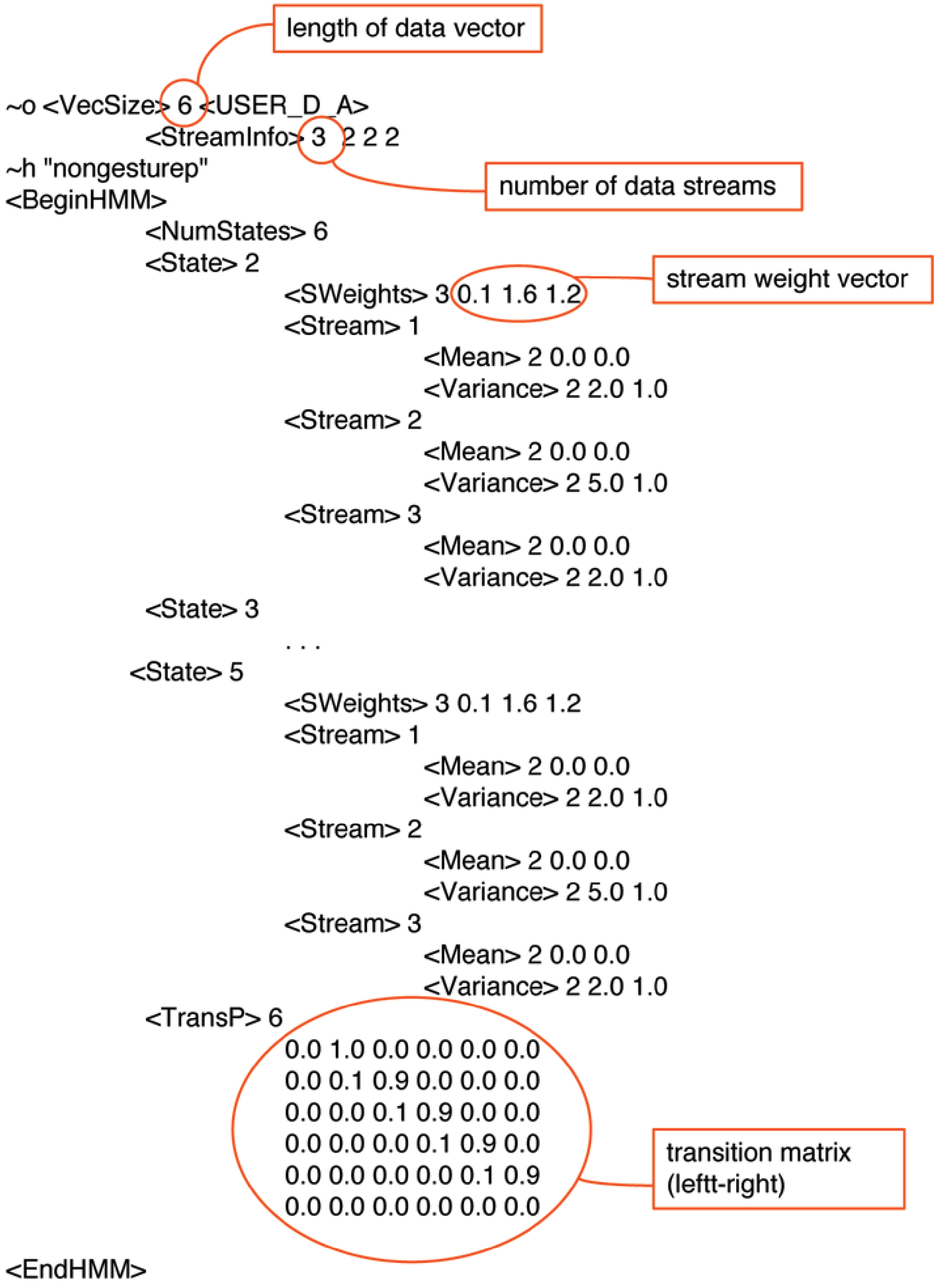


Figure 4.6: HMMs used with transition and output probabilities

Figure 4.6, which is an extension of Figure 3.4, shows used HMMs with states, transition and output probabilities. Two HMMs used, *nongesture* and *fgesture* HMM, share initial and final states, which are non emitting states. Emitting states are individual for each HMM, and they are marked on the figure with numbers 2,3,4, and 5 in circles. For each emitting state, the transition probability $a_{x,y}$ is known, as well as the output probability $b_x(O_t)$.

To train a set of HMMs, an associated transcription for every file and a dictionary is needed. 47 out of 57 videos from the database are used for the training, and the rest for testing of HMM models.

Following are the steps to follow when generating HMMs, assuming that the set of speech data files is converted into appropriate parametric form, i.e. that parameter



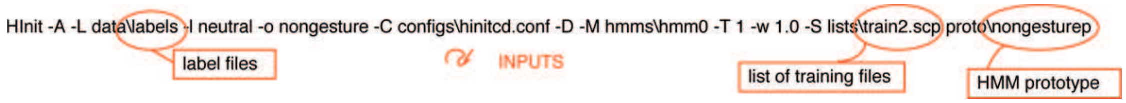
```
~o <VecSize> 6 <USER_D_A>
            <StreamInfo> 3  2 2 2
~h "nongesturep"
<BeginHMM>
            <NumStates> 6
            <State> 2
                        <SWeights> 3 0.1 1.6 1.2
                        <Stream> 1
                                    <Mean> 2 0.0 0.0
                                    <Variance> 2 2.0 1.0
                        <Stream> 2
                                    <Mean> 2 0.0 0.0
                                    <Variance> 2 5.0 1.0
                        <Stream> 3
                                    <Mean> 2 0.0 0.0
                                    <Variance> 2 2.0 1.0
            <State> 3
                        . . .
           <State> 5
                        <SWeights> 3 0.1 1.6 1.2
                        <Stream> 1
                                    <Mean> 2 0.0 0.0
                                    <Variance> 2 2.0 1.0
                        <Stream> 2
                                    <Mean> 2 0.0 0.0
                                    <Variance> 2 5.0 1.0
                        <Stream> 3
                                    <Mean> 2 0.0 0.0
                                    <Variance> 2 2.0 1.0
            <TransP> 6
                        0.0 1.0 0.0 0.0 0.0 0.0
                        0.0 0.1 0.9 0.0 0.0 0.0
                        0.0 0.0 0.1 0.9 0.0 0.0
                        0.0 0.0 0.0 0.1 0.9 0.0
                        0.0 0.0 0.0 0.0 0.1 0.9
                        0.0 0.0 0.0 0.0 0.0 0.0

<EndHMM>
```

Figure 4.7: HTK HMM prototype model

vectors are in the form of observations, and that associated transcriptions are in the correct format:

1. **Defining a prototype model.** The parameters of the model are not important – its purpose is to define a model topology. The topology used is left-right with 6 states (4 emitting) with no skips. Three data streams are used where parameter vector is split with static coefficients in stream 1, delta coefficients in stream 2 and acceleration coefficients in stream 3, with different weights for each stream. Since changes in prosodic features, specifically pitch, are the ones that are connected to facial gestures, the data stream consisting of delta coefficients is given the biggest weight. Figure 4.7 shows the prototype model for the neutral HMM (*nongesturep*) given with basic explanations. The similar prototype model is created for the gesture HMM (*fgesturep*).

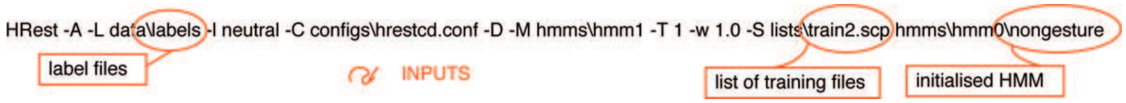
2. **Initialization.** The parameters of a new HMM are computed using a Viterbi style of estimation in HTK isolated-unit initialization (*HInit*). Figure 4.8 explains the use of HInit, where marked arguments are used as inputs.



```
HInit -A -L data\labels -I neutral -o nongesture -C configs\hinitcd.conf -D -M hmms\hmm0 -T 1 -w 1.0 -S lists\train2.scp proto\nongesturep
```

Figure 4.8: HTK HInit isolated-unit initialization

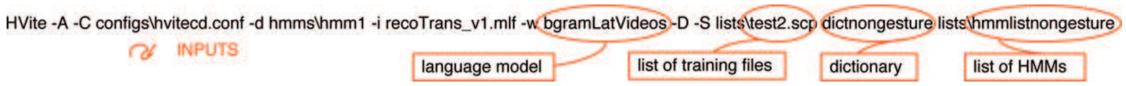
3. **Training.** The parameters of existing HMMs are refined using a Baum-Welch re-estimation in HTK isolated-unit training (*HRest*). Figure 4.9 explains the use of HRest.



```
HRest -A -L data\labels -I neutral -C configs\hrestcd.conf -D -M hmms\hmm1 -T 1 -w 1.0 -S lists\train2.scp hmms\hmm0\nongesture
```

Figure 4.9: HTK HRest isolated-unit training

4. **Testing.** It is used for monitoring of the recognition performance, i.e. it is the basic evaluation done on the testing part of the database. HTK *HVite* calculates labels using a language model, a dictionary and HMMs. Figure 4.10 explains the use of *HVite*. HMMs are generated in the training step. The dictionary is



```
HVite -A -C configs\hvitecd.conf -d hmms\hmm1 -i recoTrans_v1.mlf -w bgramLatVideos -D -S lists\test2.scp dictnongesture lists\hmmlistnongesture
```

Figure 4.10: HTK HVite

prepared so that it contains a set of gestures and HMMs used to recognize them, as shown in Figure 4.11. In order to observe an appearance of the gesture in

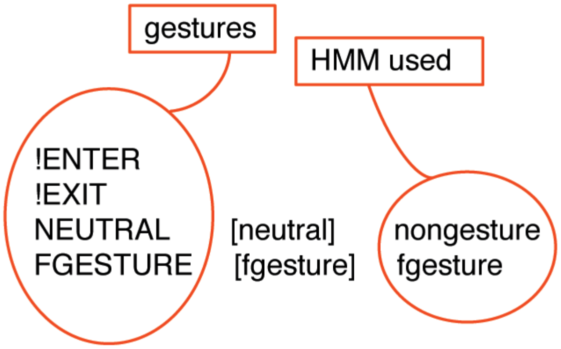
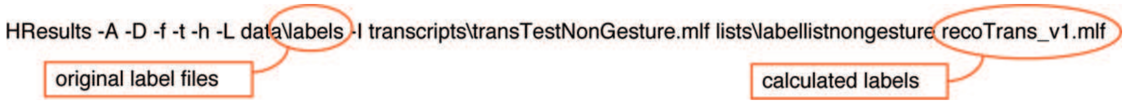


Figure 4.11: HTK dictionary

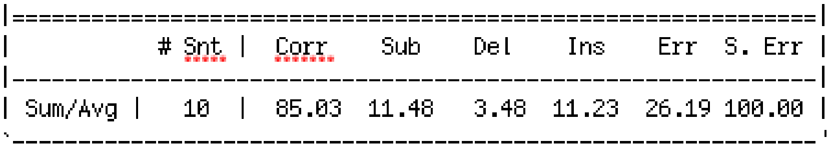
the context of the previous ones, and in such a way obtain statistically more correct models, n-grams language models are used to predict each symbol in the sequence given its n-1 predecessors, whereas n-grams are sequences of n gestures or symbols. Explanations, and the use of the language model can be found in HTK Book [71]. In this work, 2-gram and 3-gram language models with gestures *neutral* and *fgesture* are used for testing and in runtime respectively. Number of symbols in the language model is determined by the limitation of the used tools. HResults compares calculated labels with the original ones as shown in



```
HResults -A -D -f -t -h -L data\labels -I transcripts\transTestNonGesture.mlf lists\labellistnongesture recoTrans_v1.mlf
```

Figure 4.12: HTK HResults

Figure 4.12 by matching each of the recognized and reference label sequences by performing an optimal string match using dynamic programming. Once the optimal alignment has been found, the number of substitution errors (S), deletion errors (D) and insertion errors (I) can be calculated: $Corr = (N - D - S) \times 100/N$, $Acc = (N - D \check{} S - I) \times 100/N$ in percentage. Figure 4.13 shows computed gesture accuracy and other related statistics for testing videos.



```
|===========================================================|
|           # Snt |  Corr    Sub    Del    Ins    Err  S. Err |
|-----------------------------------------------------------|
| Sum/Avg |   10   |  85.03  11.48   3.48  11.23  26.19 100.00 |
`-----------------------------------------------------------'
```

Figure 4.13: HTK HResults output

As output of HMMs generation component, HMM resource group is created consisting of 3-gram language model, the dictionary and 2 HMM models.

4.2 Runtime phase

The runtime phase runs in real-time and is fully automatic. This phase takes as input a new speech signal, extracts acoustic features, triggers the method for AV mapping consisting of data-driven, rule-based and statistical modules, and produces a set of facial gestures with all needed details. Those gestures are then turned into the MPEG-4 FBA (Facial and Body Animation) bitstream. Input speech is also used in the Lip sync component to produce lip movements, i.e. visemes which are, together with facial gestures, used in an animation player to animate the face of virtual human, as shown in Figure 4.1 in the shaded part with dotted frame. An implementation of the runtime phase is done using Microsoft Visual C++ .net.

4.2.1 Description of system components

The runtime phase consists of the following components:

1. Acoustic feature extraction,
2. Data-driven module,
3. Rule-based module,
4. Statistics module,
5. Anim API,

6. Lip-sync,
7. Visage animation player.

The last three components, Anim API, Lip-sync and Visage animation player, were available for use, they were not developed in the scope of this work, thus they are only briefly introduced here.

Acoustic feature extraction

Prosody features, fundamental frequency and RMS amplitude, are extracted for every incoming audio frame as described in the *Section 4.1.1*, subsection on *Acoustic feature extraction*. In the data-driven module, both of them are needed for creating observations, and in the rule-based module only RMS amplitude is used for the detection of pauses.

Data-driven module

Input of this component are acoustic feature vectors from a novel speech signal and the HMM resource group created in the training phase. An observation vector is created every five frames, and for each vector, the most probable gesture is chosen. This means that the observed segment is classified either as a possible gesture or not. Figure 4.14 shows input speech signal in the form of observation O. HMMs used are denoted on the figure as M_{FG} for the model *fgesture*, and M_N for the *neutral* model. The probability of each model is calculated given the observation, the values are compared, and the model with highest value is chosen.

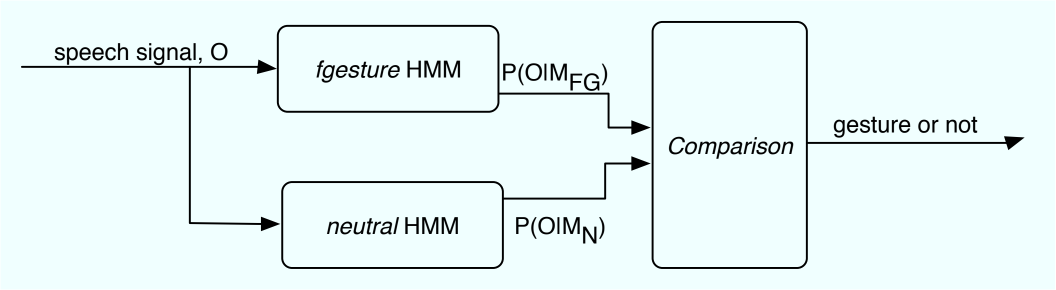


Figure 4.14: Comparison between two HMM models based on the underlying observation and HMM parameters

In the runtime, *ATK Toolkit* [74] is used for handling of HMMs. ATK is a real-time API for HTK. It consists of a C++ layer above the standard HTK libraries. By using it, HMMs built with HTK Toolkit can be used in the system. ATK consists of several library modules like *AHTK, AGram, ANGram, ADict, AHMMs* which enable handling HTK models, dictionaries, n-grams and other resources, and some basic modules like *APacket, ABuffer, AComponent*. More information on ATK components can be found in its manual [75]. When building an application using the ATK, following steps are needed:

- Instating the needed resources and components,
- Connecting them together,
- Providing an application-level code to control components and to process incoming and out-going packets.

At runtime, the recogniser can be configured to run in various modes, which includes when the recogniser should start and stop, and the method of returning results. In this system, it starts on the beginning of a file, ends at the end of a file, and results are returned every five frames.

Figure 4.15 gives an overview of the used ATK functionalities. The roles of basic components are:

- **ABuffer.** Supports a data communication, including the waveform data, speech features and the recognition output, between different components.
- **ASource.** Speech frames are grouped into wave packets and sent to the output buffer.
- **ACode.** Converts the speech waveform data into feature vectors. Information about the number of streams and the parameter kind is given to this component.
- **ARec.** Has similar functionality as *HVite* in HTK. When supplied with the resource group, it decodes incoming feature vectors every five frames, as given with the trace back frequency value (*trbakFreq* = 5). The recognition result is sent in the form of phrase packets.
- **AHMMs.** Supports HMM models to be used, given in the form of a list.
- **ADict.** Supports the use of the dictionary.

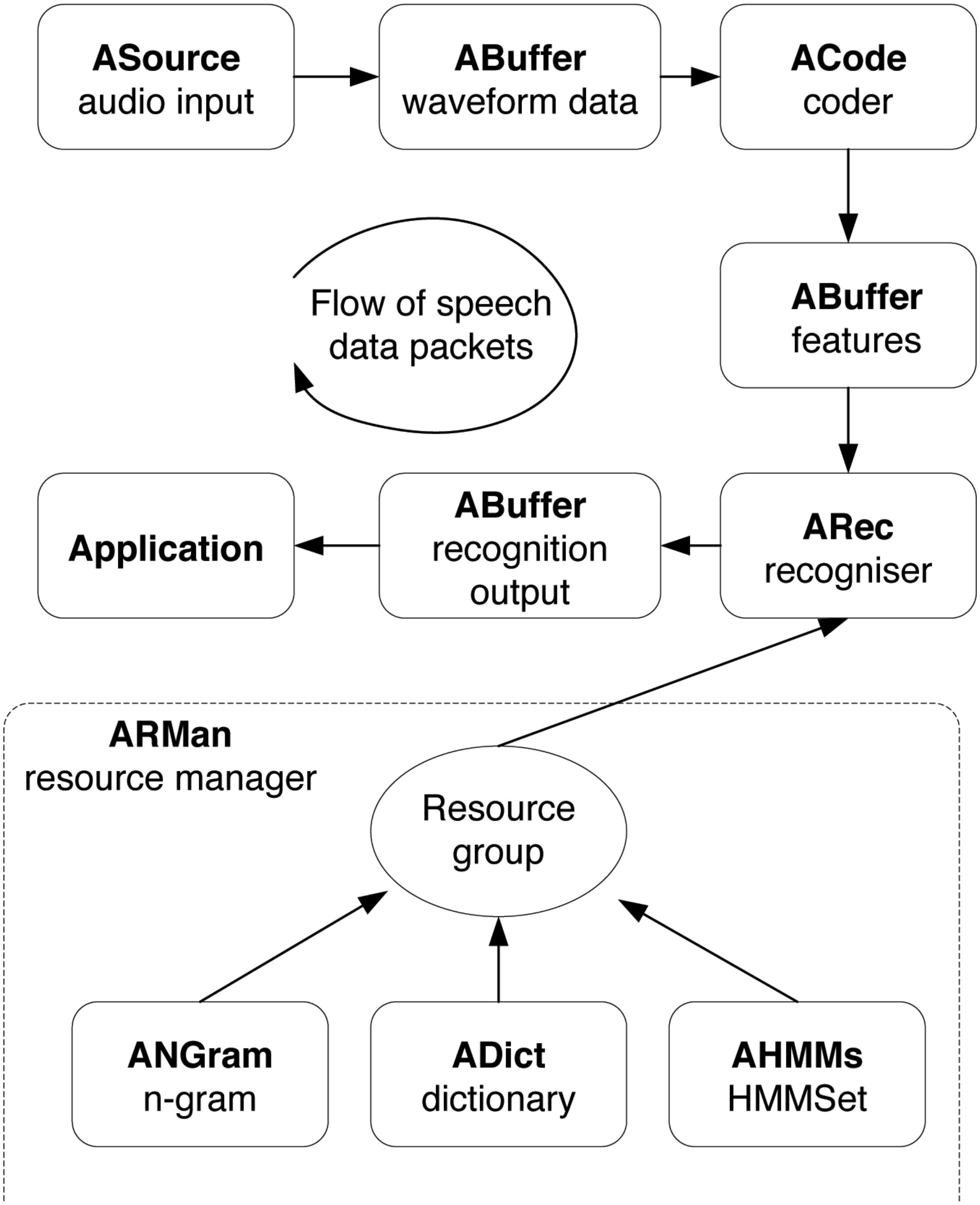


Figure 4.15: Overview of ATK usage

- **ANGram.** Supports the use of the n-gram language model.
- **ARMan.** Is a resource manager object storing HMM models, the dictionary and the language model.

Items, *ASource* and *ACode* actually belong to the Acoustic feature extraction component, but they are stated here to have a complete view of the used ATK functionality.

Rule-based module

Rules described in the previous chapter are applied based on the incoming speech, as well as on the information about facial gestures that are currently being rendered, and on the information about last appearance of certain gesture. The implementation is done in C++.

Statistics module

Statistics generated in the training phase is applied on outputs from data-driven and rule-based modules as described in the previous chapter. The implementation is done in C++.

Anim API

An animation API creates animation FBA stream for generated facial gestures including:

- nod, swing and eyebrow movements in all feasible directions using a sine trigonometric function, where eyebrows frowns are considered as eyebrows raises with a negative amplitude,
- simple animation models for eye blinking,
- gaze following,
- gaze model, for left, right, up, down, including the eyelid movement.

Anim API relies on Visage|SDK [76] which is C++ Software Development Kit for virtual character animation. It has been implemented by Karlo Smid [68].

In the AnimAPI, constants for different movements are set. In current implementation, their values are:

- EYELIDS(801) – determines the maximum amplitude for the opening of eyelids,
- EYEBROWS(350) – determines the maximum amplitude for the eyebrow raise or frown,
- NOD(6319) – determines the maximum amplitude for the nod or swing movement,
- GAZE_ V(5000) – determines the maximum amplitude for the vertical gaze,
- GAZE_ H(10000) – determines the maximum amplitude for the horizontal gaze.

These values can be adjusted for different use of the system – with bigger values the movements are more exaggerated, and with smaller they are more moderate.

Lip-sync

This component is used for a lip synchronization, i.e. for generating visemes from the input speech. Visemes are visual representatives of phonemes. For every speech frame, it calculates the correspondent viseme. The results for four consecutive frames are summed up, and the viseme with the best score is chosen. The classification of the speech in viseme classes is done by Neural Networks (NNs), one for each viseme, based on the input, which is a 12-dimensional MFCC vector representing the speech signal.

More details on Lip sync can be found in [49], and [77], and it is a part of Visage|SDK.

Visage animation player

As an animation player the FAPlayer from Visage|SDK is used. Based on the MPEG-4 standard, it merges two tracks – calculated visemes and the FBA stream containing facial gestures, and renders it on the screen. Once the required information is extracted from the speech, any parametrized face model can be animated.

4.2.2 Example application

The runtime usage and functionality described in the previous section is demonstrated in an example Windows application developed in Microsoft Visual Studio .net using ATK Toolkit and Visage|SDK. It is used to validate the system for speech driven facial gesturing for virtual characters, as it will be described in the next chapter.

The application takes the speech file and the animatable face model as inputs, generates facial gestures and visemes by analysing the speech signal, and applies them on the face model. Facial gestures in the form of FBA stream, and visemes are used as two separated animation tracks and they are both applied on the face model. It is done using Visage|SDK, and animation player (*FAPlayer*). The synchronization between these two streams is assured through the internal Lip-sync clock.

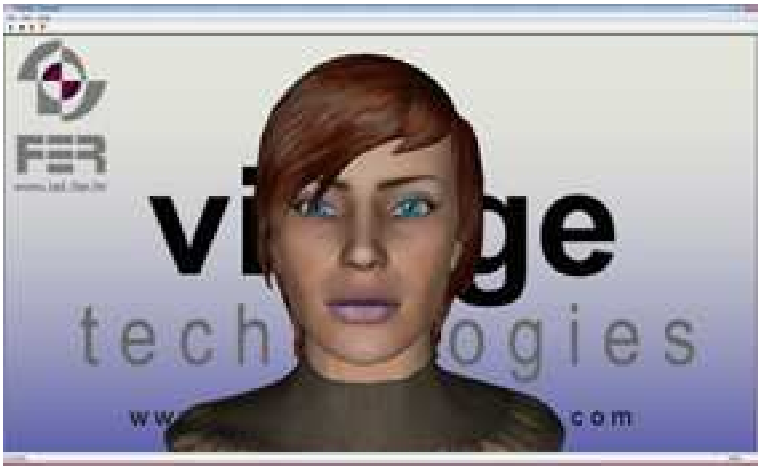


Figure 4.16: Application interface

Figure 4.16 shows an interface of this application.

Since the used face model and the generated facial animation are based on the MPEG-4 standard for facial animation, the next paragraph gives some basic information about it.

MPEG-4 facial animation

The MPEG-4 standard on Face and Body Animation (FBA) [78] specifies a face model in its neutral state, a number of feature points (FPs) on this neutral face as reference points and a set of Face Animation Parameters (FAPs) used to control the animation of a face model.

There are 84 FPs on the neutral face serving as reference points for defining FAPs. The 68 FAPs are categorized into 10 groups, whereas each FAP corresponds to a particular facial action deforming a face model in its neutral state. The first group contains high-level parameters, visemes and expressions. There are 15 static visemes included in the standard set, including the neutral face. Deforming the model according to some specified FAP values at each time instant generates a facial animation sequence.

The FAP value for a particular FAP indicates the magnitude of the corresponding action, e.g. amplitude of the eyebrow raise.

In order to define FAPs for arbitrary face models, MPEG-4 defines Facial Animation Parameter Units (FAPUs) that scale FAPs for any face model. FAPUs are defined as distances between key feature points.

The FAPs can be compressed and included in an FBA bitstream for low bitrate storage or transmission. An FBA bitstream can be decoded and interpreted by any MPEG-4 compliant face animation system.

More details on the MPEG-4 standard (i.e. MPEG-4 compatible 3D faces, MPEG-4 FA player) can be found in [78].

Chapter 5

Perceptual evaluation of visual prosody

This chapter evaluates the system and the previously set up hypothesis: *Adding facial gestures driven by the speech signal to the animation of virtual characters using the proposed hybrid approach in real-time will result in more coherent and appropriate facial movements than random or no movements.*

Section 5.1 briefly describes a system evaluation process, while Section 5.2 explains a subjective evaluation used to validate the system. The purpose is to compare facial animation produced using the hybrid method for speech driven facial gesturing with two other animations – the first one containing facial gestures directly copied from the video clip from which the used audio was extracted, and the second one with the facial gestures produced randomly. A description of each model is given in Section 5.2.1, a stimulus is presented in Section 5.2.2, results in Section 5.2.3, and a discussion is provided in Section 5.2.4.

5.1 System evaluation process

The development of the AV mapping method and the whole system for speech driven facial gesturing was done in several iterations. A design process was started studying the related work which was spread out in several research fields. In order to correlate the speech signal with facial movements, it was important to understand how humans communicate to each other. Those findings could then be applied on virtual humans and on HCI. Besides human communication rules, a quality of digital speech processing and limitations of the real-time system played an important role in designing the method and the system. Several versions of the system were developed based on different methods for AV mapping and tested with several subjects. Implied conclusions were used in the next iteration. The process is roughly demonstrated on Figure 5.1.

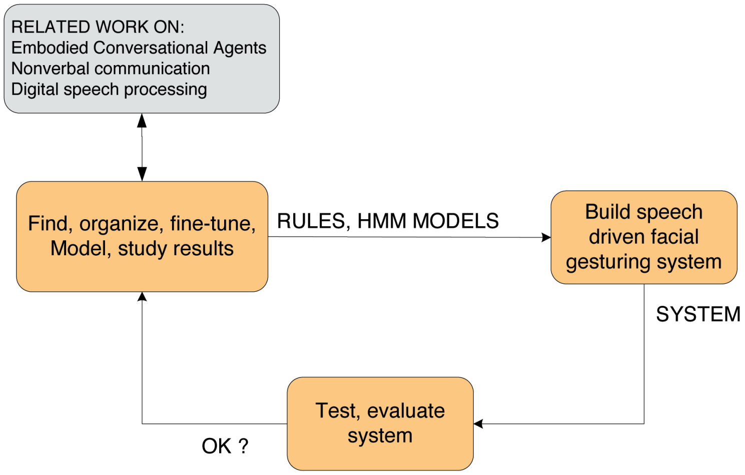

RELATED WORK ON:
Embodied Conversational Agents
Nonverbal communication
Digital speech processing
Find, organize, fine-tune, Model, study results
RULES, HMM MODELS
Build speech driven facial gesturing system
SYSTEM
Test, evaluate system
OK ?

Figure 5.1: Method and system design process and testing cycle

The first visual results obtained using the hybrid approach looked promising. However, they still needed improvement. Considering the data-driven module and HMMs, following different parameters, configurations and approaches were tried out:

- Number of states and HMMs,
- Only the pitch acoustic feature used, compared to pitch and amplitude,
- One data stream, compared to two or three streams with delta and acceleration coefficients and their associated weights,
- A transition matrix with different values,
- A simple (grammar) network compared to a language model used,
- Different labels for non gesture HMM, with difference in the length and the position in the training speech file,
- Different labels for gesture HMM, covering the whole gesture duration or certain number of frames,

In each iteration, the test results produced in the training phase were checked with HTK *HVite* and *HResults* and based on this the method was further evaluated in the runtime phase on one or several subjects, depending on how good the solution was. In each iteration where it was applicable, literature and existing solutions were further consulted for better parameter definition.

Considering the rule-based module, various values were tried out for the duration of short and long pauses, the threshold for determining silent parts of the speech, as well as amplitudes of gestures in Anim API. The choice of gestures for the particular rule was considered, constantly checking the visual output. If the visual output was unsatisfactory, literature was additionally reviewed looking for new directions. Consequently, the calculation of the statistics data has followed the design of data-driven and rule-based approaches.

Once visual impression of the obtained visual prosody was satisfactory, a perceptual evaluation as described in the next section was done.

5.2 Perceptual evaluation

For virtual human applications, in addition to the system performance and capabilities, visual impression is an important indicator of the quality, since the observers are the ones to whom virtual characters talk. Moreover, when generating visual prosody, there is no only one possible output, but rather certain number (if not even unlimited) of correct combinations of gestures to use. Therefore, a subjective test similar to [59] was conducted.

The main goal of the perceptual evaluation experiment was to rate the system for speech driven facial gesturing compared to the systems with randomly added facial gestures, or with no gestures at all.

For this purpose, four facial animation models like in [79] were created. Facial animations produced were captured using Fraps [80], a real-time video capture software, and saved as videos in xvid format for the later presentation. The models used are described next.

5.2.1 Models used for perceptual evaluation

In this experiment, all produced models use the same audio file as the input and the same animatable face model Reana provided with Visage|SDK [76] and generate four different visual outputs. Each of them uses the same lip-synchronization, and only added facial gestures were observed (lip movements are not in the scope of this work).

No visual prosody

The production of the animation without visual prosody involved only the Lip-sync component as described in *Section 4.2.1*. For the chosen audio file, lip movements were generated while the rest of the face was still.

Random visual prosody

For this model, random movements for eyebrows and head, as well as for the gaze and eye blinking were employed. Each of these gestures is randomly generated within the period of 10 seconds using the uniform distribution. Swing and nod movements are equally probable. Values for the amplitude and the duration were randomly generated, but still within the range that is acceptable in order not to get obviously unnatural facial movements. Also, in the case when gaze away is happening at the same time as the eyebrow or head movement, gaze away follows eyebrow or head movement direction. When choosing the gesture subtype, a random number is used, whereas each subtype was equally probable.

In doing so, an animation which generally do not correspond to the speech is produced.

Speech driven visual prosody

Speech driven visual prosody model is produced using the hybrid method and the example application described in *Chapter 3* and *Section 4.2.2* respectively.

Ground truth visual prosody

This model contains facial gestures as they appear in the video corresponding to the used audio file. For testing purposes, one of the videos from database which was not used for training of HMMs was chosen. From available XML file with gesture annotations, information about facial gestures, including gesture type, timing and amplitude, was manually incorporated in the example application, and used to produce the animation model.

5.2.2 Stimulus presentation

In the earlier work on lip-synchronization [49] one of the main negative comments on the results was on unnatural impression of the whole face since only the lips were moving. To confirm this, an introductory experiment was done with a small number of participants to compare that model with the speech driven visual prosody model. As expected, the animation without visual prosody was graded notably worse than animation which included visual prosody. Because of this, it was not included in the main experiment.

In this experiment, two testing utterances were used.

One of them is extracted from the testing pool of the training database, and belongs to the female speaker whose utterances were used, among others, for the training of HMMs. It is used as an audio input, and three animations were generated using models with random, speech driven and ground truth visual prosody. Those models are presented in a random order to 47 testing subjects. Recruited participants had different backgrounds and levels of knowledge on the subject, whereas most of them were unfamiliar with the details of the system. A transcription of the 18 seconds long utterance is:

In a case of apparent mobile rage in America a seventy seven old man is charged with reckless endangerment. He saw a woman in another car dial her phone in a red light and forced her of the road once the light was green. He haven't seen her hang up.

In this survey, a quality of the animation driven with the proposed method is compared to the random and the original, i.e. ground-truth generated visual prosody.

Another used utterance is pronounced by a novel female speaker in Croatian language. Two animations are generated, using random and speech driven visual prosody models, and presented in random order to 37 testing subjects. A transcription of the 17 seconds long utterance is:

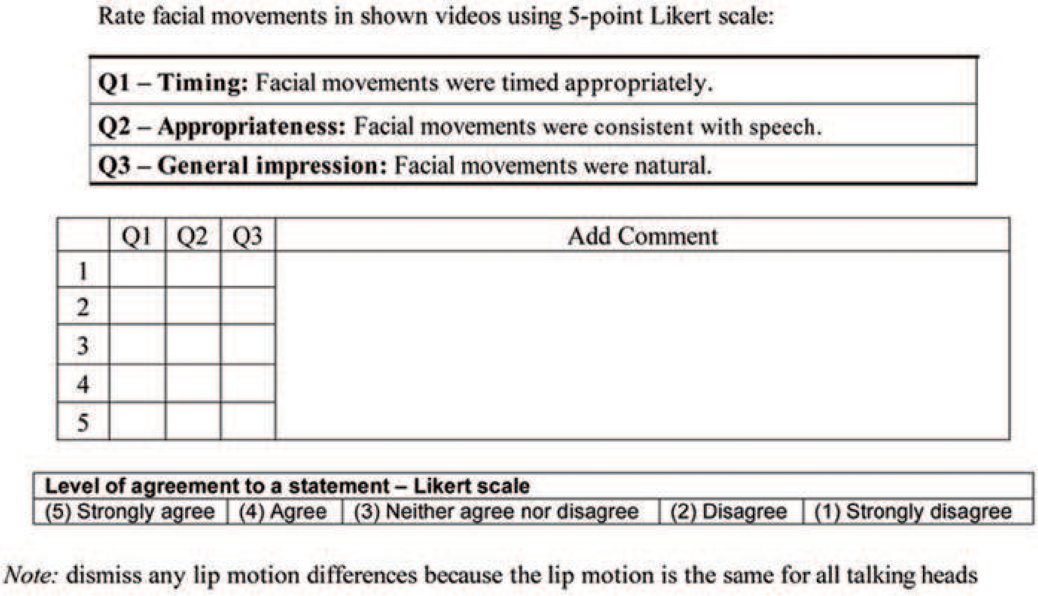
Danas je virtualna animacija toliko napredovala pa se već koristi u filmovima i dugometražnim crtićima. No stvaranja lica virtualnih ljudi još uvijek je problematično. Najteže za postići kod animacije virtualnih ljudi upravo je animacija lica.

Along with comparing the speech driven system to random visual prosody, the aim of this survey was to check how well it works with different speakers.

In both cases, participants were asked to rate each animation on a five-point Likert scale, which specifies a level of agreement to a statement, where grades 5, 4, 3, 2 and 1 denote strongly agree, agree, neither agree nor disagree, disagree and strongly disagree respectively. Used statements are:

- **Timing:** Facial movements were timed appropriately.
- **Appropriateness:** Facial movements were consistent with speech.
- **General impression:** Facial movements were natural.

Additionally, participants were asked to optionally enter a comment on seen videos. They were also instructed to dismiss any lip motion inaccuracy because the lip movements were the same for all generated videos of the same utterance. Figure 5.2 shows used evaluation form.



Rate facial movements in shown videos using 5-point Likert scale:

Q1 – Timing: Facial movements were timed appropriately.
Q2 – Appropriateness: Facial movements were consistent with speech.
Q3 – General impression: Facial movements were natural.

	Q1	Q2	Q3	Add Comment
1				
2				
3				
4				
5				

Level of agreement to a statement – Likert scale				
(5) Strongly agree	(4) Agree	(3) Neither agree nor disagree	(2) Disagree	(1) Strongly disagree

Note: dismiss any lip motion differences because the lip motion is the same for all talking heads

Figure 5.2: Evaluation form used in the perceptual evaluation

5.2.3 Results

Questions and average scores for the survey with the first utterance are presented in Figure 5.3, and for the second utterance in Figure 5.4.

In both cases, speech driven generated animation outperformed the random sequence. A statistical confidence of the obtained results is checked with a pair-wise two-tailed t-test using online calculator [81] as shown in the table 5.1, where a *t value* is the estimated number of standard errors between the two mean scores and a *p value* determines the strength of the evidence. More information about *p value* can be found in [82].

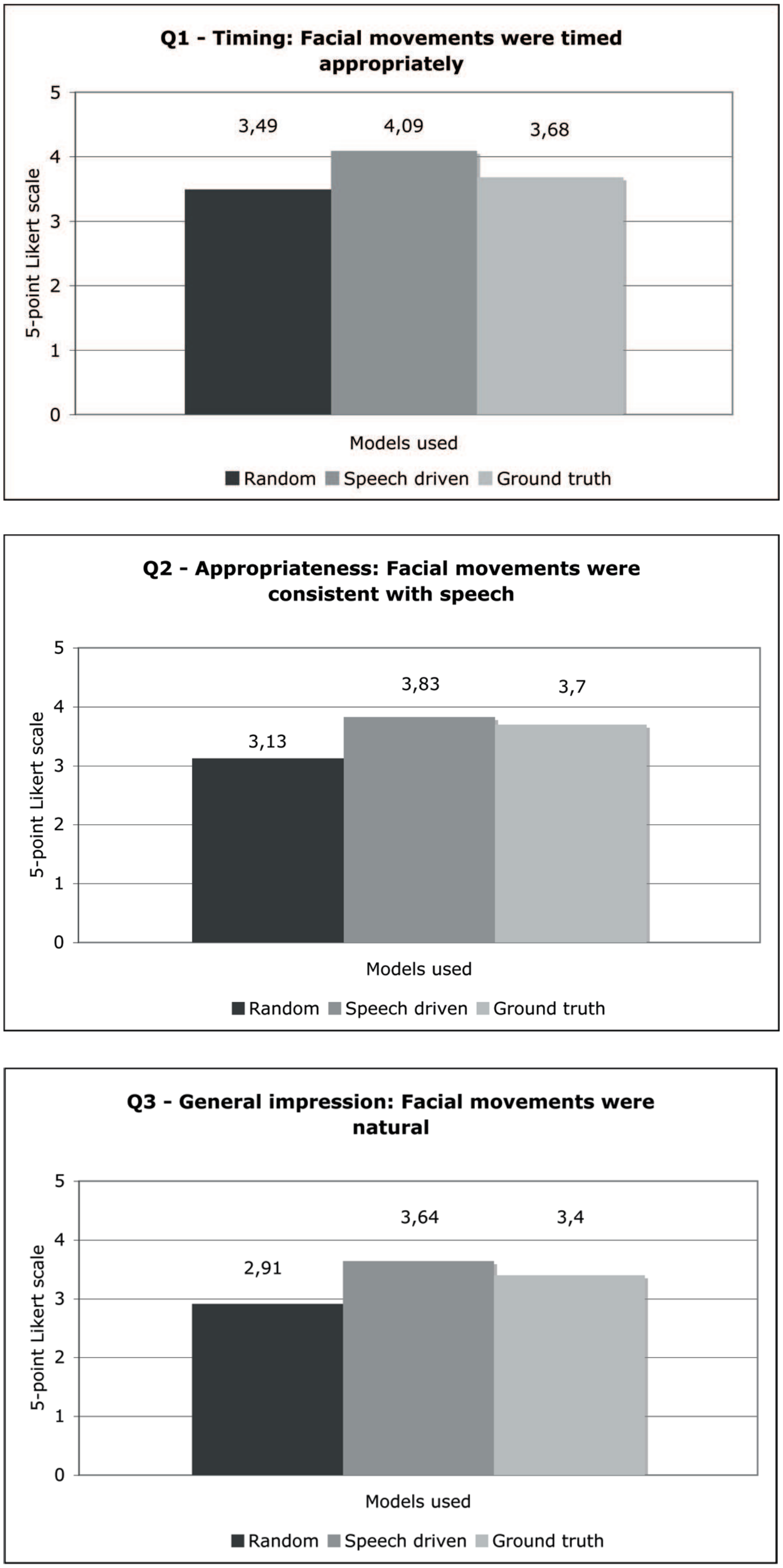


Figure 5.3: Results obtained with the first testing utterance

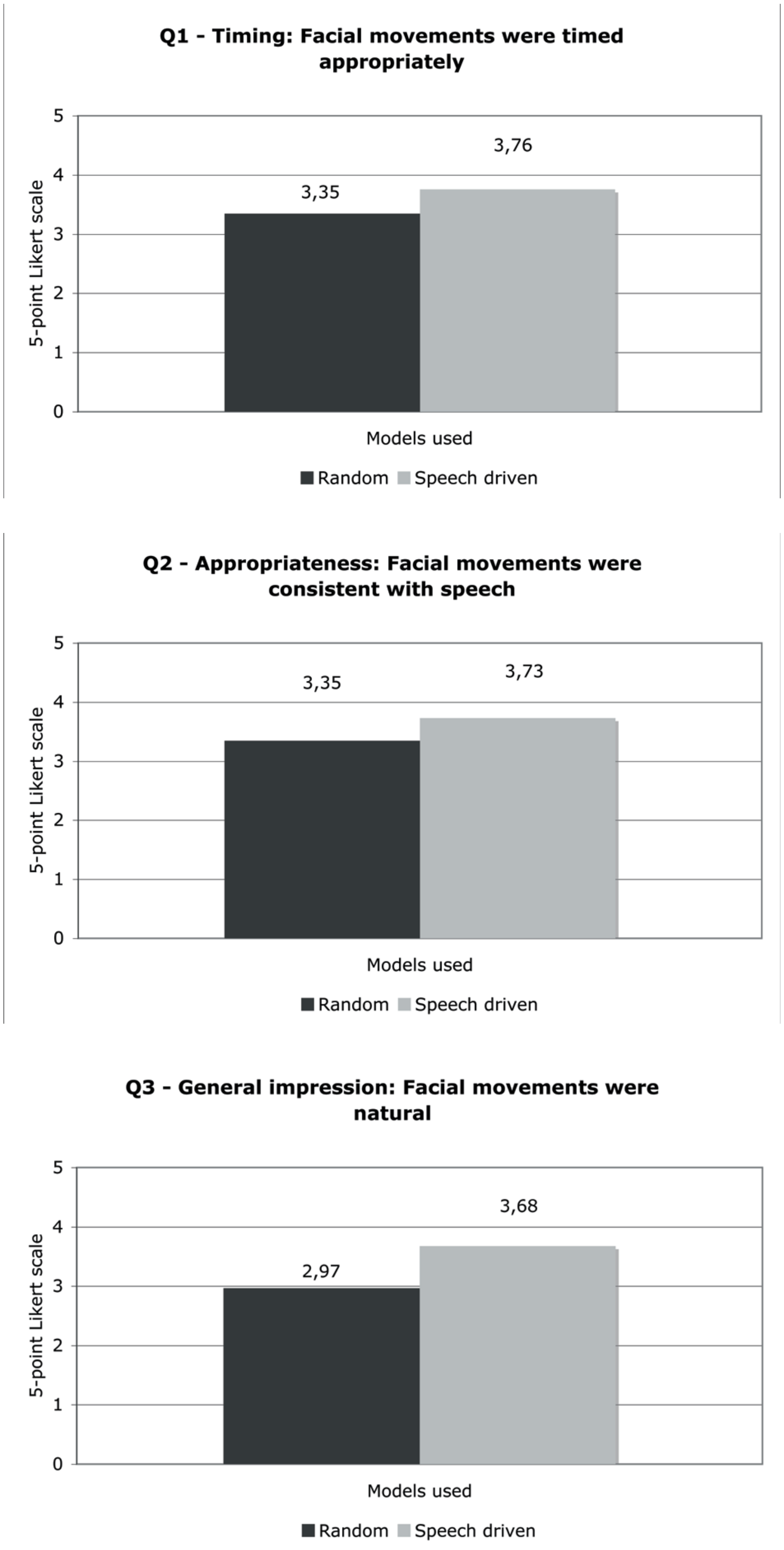
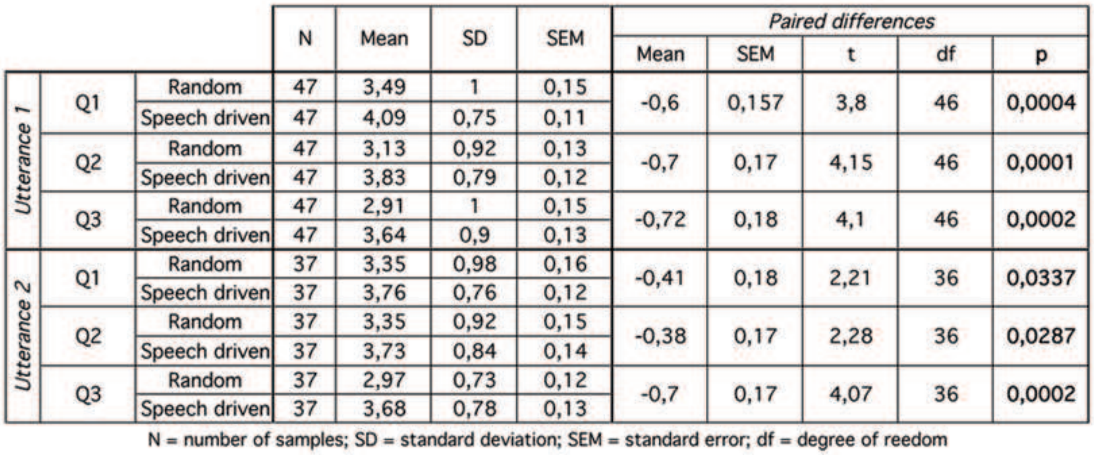


Figure 5.4: Results obtained with the second testing utterance



			N	Mean	SD	SEM	Paired differences				
							Mean	SEM	t	df	p
Utterance 1	Q1	Random	47	3,49	1	0,15	-0,6	0,157	3,8	46	0,0004
		Speech driven	47	4,09	0,75	0,11					
	Q2	Random	47	3,13	0,92	0,13	-0,7	0,17	4,15	46	0,0001
		Speech driven	47	3,83	0,79	0,12					
	Q3	Random	47	2,91	1	0,15	-0,72	0,18	4,1	46	0,0002
		Speech driven	47	3,64	0,9	0,13					
Utterance 2	Q1	Random	37	3,35	0,98	0,16	-0,41	0,18	2,21	36	0,0337
		Speech driven	37	3,76	0,76	0,12					
	Q2	Random	37	3,35	0,92	0,15	-0,38	0,17	2,28	36	0,0287
		Speech driven	37	3,73	0,84	0,14					
	Q3	Random	37	2,97	0,73	0,12	-0,7	0,17	4,07	36	0,0002
		Speech driven	37	3,68	0,78	0,13					

N = number of samples; SD = standard deviation; SEM = standard error; df = degree of reedom

Table 5.1: Pair-wise t-test comparison of random and speech driven models

According to this test, in all three statements, for both utterances, speech driven visual prosody was better graded than random visual prosody, and the difference is considered to be statistically significant, with p < 0,05, which proves the hypothesis.

5.2.4 Discussion

In the previous subsection results from the perceptual evaluation are presented. They are elaborated here, including the comments from the participants.

The most frequent comment was that video representing speech driven visual prosody looked natural compared to others, both random and ground truth visual prosody. Results, as shown on figures 5.3 and 5.4 support this comment – an average score of the speech driven model for all statements and both utterances was higher than for random and ground truth models (applicable only for the first utterance). Although a result like this was expected in the case of the random and speech driven model relationship, it was unforeseen in the speech driven model to ground truth model relationship. Still, it can be explained with the imperfect virtual head model and animations of facial gestures. Perfectly natural head and facial movements when applied to the virtual head do not necessary look as good as the original. Moreover, in building the method for speech driven facial gesturing, a special care was given to fine-tune the visual output, to be both correct in terms of visual prosody and its applicability to the facial model and gesture animation used. The following comment reflexes this in a good way: *I like the second model (i.e. speech driven) best, it is the most pleasant one, although I can* *not tell that it looks natural. On the other hand, the third model (i.e. ground truth) looks very natural, but still some movements are somehow wrong.* Nevertheless, the ground truth model outperformed the random one.

Another frequent comment was on excessively exaggerating facial movements, especially eyebrow movements in random visual prosody, which leads to the perception of an unnatural look. The reason for this was not only due to inappropriate amplitudes or duration of the movements, but also because the movements were not always coherent with the speech, giving the impression that something was wrong with the face. Nonetheless, a certain number of subjects preferred the random model because of these too affective movements and at the same time considering that the other two models did not have enough visible gestures. Similar conclusion was stated in [79]: *subjects used the amount of head movements as a strategy to select the preferred facial animation, rather than coherence between these movements and the speech track.*

As a final comment to single out, several testing subjects remarked on an appropriateness of speech driven model having in mind a special target use – the news presenter.

When comparing results between the first utterance, pronounced by the speaker that exists in the database, and the second utterance, pronounced by the novel speaker, in most cases the mean score for the first utterance is slightly higher than for the second utterance. Because of the hybrid nature of the proposed approach, a result like this was actually expected. The training database plays a role in data-driven and statistics modules, but not in the rule-based, which is mostly responsible for the final touch of the visual output.

Comparison among three statements, in both cases, gives approximately the same mean score for the statements on timing and appropriateness, and lower score for the third statement on general impression and naturalness of the facial movements. Having that in mind, the next research question might be raised: Besides well timed and speech consistent facial movements what else is needed to achieve a more natural impression?

All these findings led us to the conclusion that the facial movements synthesized using the hybrid method are generally consistent and time aligned with the underlying speech and a virtual human applying them is perceived as pleasant and suitable for the use as news presenter. More on that will be given in the next chapter.

Chapter 6

Speech driven virtual news presenters in networked environment

The work on real-time speech driven facial gesturing for virtual characters that aims to reproduce the behaviour of real news presenters was described so far. In this chapter, more details on the system from the application perspective will be given through a virtual news presentation use case. In this, a specific interest lies in real-time services based on the speech driven virtual characters used in networked environments. Thus, in Section 6.1 a basic idea is briefly explained, and the related work on virtual news presenters is given, and then Section 6.2 focuses on issues that arise when using animated virtual characters in networked environments.

6.1 Virtual news presenter

For already more than a decade, virtual characters are used as news presenters. A pioneering work was done in 1995. [83] – a virtual actor acting as a television presenter using previously prepared animation sequences was created. Another example of a television speaker is work done in [1] where virtual presenter is automatically built using visual text-to-speech (VTTS). The result is demonstrated on the competition *Gathering of Animated Lifelike Agents*, Gala 2006 (Figure 6.1).

Besides as TV presenter, a virtual newscaster, which is another commonly used name, can be found on a web page, reading news over the Internet. It is typically driven by the visual text-to-speech, and in some scenarios, in the combination with the scripts describing the agents behaviour.

An example is *Ananova* [84], which is often called the first virtual newscaster.

Figure 6.1: News presenter driven by visual TTS [1]

Its main idea was to personalize a news service by collecting suitable news items and converting the text into speech, and at the same time creating the real-time facial animation used for news presentation. Similarly, *News At Seven* [85] combine different web content and automatically generates a virtual news presentation using avatars and text-to-speech (TTS) technology. However, although these works were generally accepted positively, the users did not like the artificial voice produced by the text-to-speech engine.

With the speech driven avatars, the problem of computer-generated voice is completely avoided. Instead, the voice of real human is used to drive a virtual character. Yet, in this work an idea of the use of virtual news presenters do not go in the direction of the automatic news context generation, as it is often the case in VTTS based systems. Instead, the emphasis is on the scenarios where it is not possible to stream news as a video, or TV broadcast due to the network limitation – for speech driven facial animation it is only needed to transmit the voice and animation parameters, assuming that the terminal is already equipped with the suitable face model, which requires significantly less bandwidth than the video stream.

However, in such scenarios the real-time animation generation from the novel speech is required, and that is not possible when using scripts, since actions describing an agent behaviour in them are manually noted. An example is a system in [86] where as the input a speech text and commands related to the gestural movement are used.

To summarize, a good virtual news presenter in networked environment needs to have following qualities:

- Real-time animation generation synchronized with the speech,

- An automatic behaviour extraction,
- Verbal and correspondent nonverbal communication channel leading to the natural and convincing presenter-like behaviour, necessary for users to trust the agent.

A virtual presenter designed in that manner can effectively be used to inform on the latest news in various networked environments. Figure 6.2 shows symbolically how the system for speech driven facial gesturing can be used for that purpose. As an information source, a news item in the form of a speech signal is used. Further more, animation data is generated using the system for speech driven facial gesturing and the lip synchronization, which is then together with the source speech streamed through the IP (Internet Protocol) network to different terminals equipped with the MPEG-4 compatible face model.

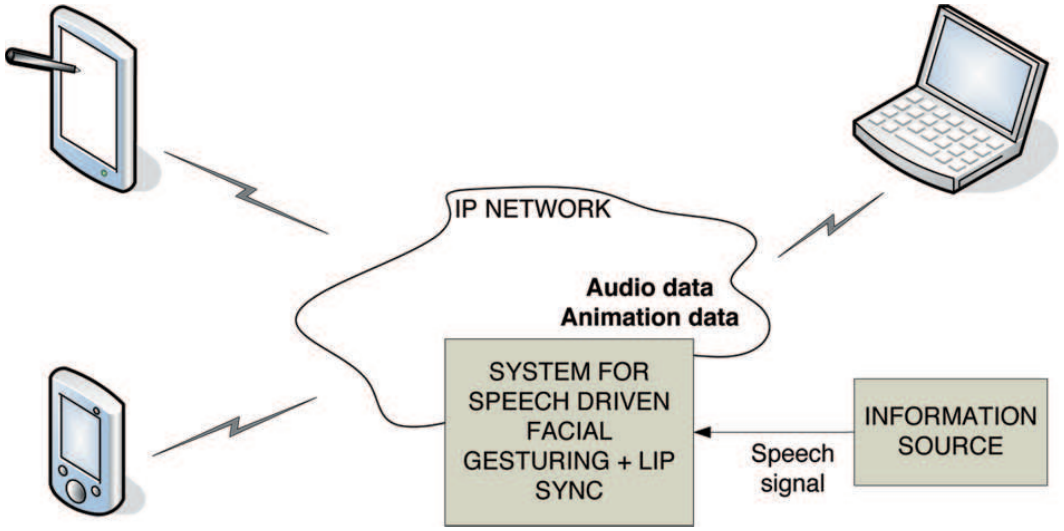


Figure 6.2: Symbolic view of the speech driven system use for virtual news presentation in networked environment

6.2 Issues in networked environment

When building services based on virtual characters not only a visual appearance matters, but a special care needs to be given also to the perceived audio and visual quality. In networked environments, a real-time animation of virtual humans includes streaming of media, therefore its design must address the challenges inherent in the real-time delivery over a best-effort network. If an application is not able to adjust its behaviour to the varying network conditions, a streaming media session can exhibit degraded performance, which can translate into a disappointing user experience. Whereas, network dynamics are caused by various factors, like the route changes, competing traffic, congestion etc.

Current state of the art includes several works on animation streaming that partially address different network issues. A framework for the error-resilient facial animation system is proposed in [87]. Some effective error concealment algorithms are applied to improve the reconstructed face animation in a synthesizer. A work in [88] proposes a new real-time transport protocol with error recovery capability to stream facial animation over the Internet. Similarly, a work in [89] proposes a solution for a network transmission of animation streams over RTP (Real time Transport Protocol). A paper in [90] investigates FAP coding of 3D facial animation at very low bitrate. The tests carried out using second generation mobile networks showed that FAP coded streams are reasonably robust to error when compared to normal MPEG-4 coded video.

Although, those works address the animation transmission, they do not take into consideration an user experience of the service performance. In [91], subjective experiments for the audio, video, and audiovisual quality using content and encoding parameters representative of a video for mobile applications are described. A used methodology is the absolute category rating and the focus was on the mobile video transmission. A work in [92] addresses the relationship between interactivity as the quality perceived by the user and standard QoS parameters for applications in the area of networked virtual reality, including virtual characters.

As it can be seen from the above mentioned related work, there is a set of issues that needs to be considered when streaming, i.e. transmitting and receiving audio and animation data in real-time in the Internet, wireless LANs, or third generation mobile networks. Some of the most important ones are: delay, bit rate, jitter, and loss. All these parameters influence achieved quality, including a synchronization between different streams, in this case audio and animation, and the stability of the stream flow.

Despite the network imperfection, it is important to keep a good audio visual quality. To do so, it is needed to build an error resilient and consilient system for real time virtual character animation in the Internet. MPEG-4 standard for facial animation supports this need through its mechanisms. Some of them are already mentioned in the previously stated related work. A brief description follows, first for the error resilience (transmitter side), and then for the error concealment (receiver side).

Error resilience algorithms applicable for MPEG-4 animation include:

- Facial Animation Parameters (FAPs) grouping and masking,
- Use of arithmetic compared to predictive coding,
- Use of MPEG-4 I (*Intra-coded picture*) and P (*Predicted picture*) frames through I frames distance and start codes,

More about FAPs coding and streaming can be found in [78].

Additionally, connected to Real time Transport Protocol (RTP) [93], a protocol that defines a standardized packet format for delivering audio and video over the Internet, it is possible to use a payload for FAPs containing recovery info, or the optimal number of frames per RTP packet compared to overhead generated by the header size.

On the other hand, error concealment algorithms are concerned with the following:

- Compensation of jitter and out of order packets using buffers,
- Detection of errors, i.e. packet losses by e.g. searching for next reference frame or recovery info,
- Synchronization of audio and animation streams using e.g. RTP timestamps and interpolation of FAPs.

Information given in this section provide an initial direction towards building the error resilient and consilient system for real-time virtual character animation in networked environment, and that is where the development of the system for speech driven facial gesturing strives next.

Conclusions and future work

This doctoral thesis has been motivated by the extensive use of various computer devices, as well as the growing importance of human-computer interaction design, where virtual humans already have an important role due to their multimodal nature. They incorporate modalities that humans naturally use in communication, thus the users are very sensitive to the behaviour of virtual humans. They expect a human-like behaviour for a human-like character, whereas a special care needs to be given to the face of virtual character, since through the face majority of communication messages are transmitted, both verbally and nonverbaly.

A research area of this thesis is automatic facial gesturing of virtual characters based only on a speech input. The main issue is a correlation of the speech signal with facial movements, which is not a trivial task, since the relationship is neither direct nor obvious, as it is in the lip synchronization. In order to perform the audio to visual mapping, it is important to understand how humans communicate to each other, and how to apply those findings on virtual humans, and on human-computer interaction. This implies linking the related knowledge from psychology and paralinguistic, the knowledge on digital speech processing, and the knowledge obtained by researchers from community working on embodied conversational agents.

In this thesis, facial gestures that has been included are various nods and head movements, eye blinks, eyebrow gestures and gaze, i.e., facial displays used in communication, including head movements, but excluding explicit verbal and emotional displays (e.g. visemes or expressions such as smile). Facial gestures are largely responsible for what we call natural behaviour of the face, and it is important to model them properly.

In this thesis, a hybrid method for real-time mapping from the speech signal to facial gestures has been proposed. After state of the art literature was studied, several different methods have been constructed, implemented and validated. Finally the method based on both a data-driven and rule-based approach has been chosen.

An idea behind the proposed method is to include various issues that are considered important for visual prosody through three different modules. A data-driven module is based on Hidden Markov Models, and through it, facial gestures correlated with prominent parts of the speech are included. A rule-based module consists of a set of rules, which among others include rules for punctuation and prolonged pauses. Additionally, in a statistics module, results of first two modules are fine-tuned and further shaped.

In order to check how believable virtual characters are if animated using facial gestures obtained by the proposed method, the system for speech driven facial gesturing has been implemented, and used for a perceptual evaluation. Testing subjects generally considered facial gestures consistent and time aligned with the underlying speech and a virtual human applying them is perceived as natural. Moreover, the set up hypothesis has been proven: adding facial gestures driven by the speech signal to the animation of virtual characters using the proposed hybrid approach in real-time will result in more coherent and appropriate facial movements than random or no movements.

The main contributions of this thesis are:

1. A systematic overview of facial movements and their connection with the speech signal. Movements included are those used in nonverbal communication that are connected to the syntactic and prosodic structure of the underlying speech rather than to semantic or emotions.

2. A hybrid method for mapping from the speech signal to statistically correct facial gestures correlated with the speech prosody in real-time.

3. A system for automatic facial gesturing for virtual characters in real-time based on the proposed hybrid method.

4. A validation of the method and the system through the perceptual evaluation.

The contributions presented in this thesis provide a step closer to building believable virtual humans, and moreover embodied conversational agents driven by the speech signal. Nonetheless, the method proposed here might be improved by including parameters reflecting individual, cultural and gender differences in gesturing since all these factors impact the intensity and the frequency of facial gestures. Another way to incorporate, at least in some extent, the individual characteristics would be through an automatic system calibration or a speaker adaptation during the initialization phase. Adding support for a microphone input is certainly an important issue in order to have a wider set of possible applications for the system. To do so, first a noise cancellation or similar algorithms should be added.

Further research interests strive towards the use of speech driven virtual characters in networked environments which implies the system adaptation for streaming of audio and animation data in changeable network conditions.

Literature

[1] K. Smid, I. Pandzic, and V. Radman, “Autonomous speaker agent,” in *Computer Animation and Social Agents Conference, CASA*, 2004.

[2] D. B. Givens, *Nonverbal dictionary of gestures, signs and body language cues*, http://center-for-nonverbal-studies.org/6101.html (16.12.2009.).

[3] M. Argyle and P. Trower, *Person to person: ways of communicating.* Thomas Nelson Australia, West Melbourne, Vic. :, 1979.

[4] C. Pelachaud, N. I. Badler, and M. Steedman, “Generating facial expressions for speech,” *Cognitive Science*, vol. 20, pp. 1–46, 1996.

[5] D. I. Perrett, P. A. Smith, D. D. Potter, A. J. Mistlin, A. S. Head, A. D. Milner, and M. A. Jeeves, “Neurones responsive to faces in the temporal cortex: studies of functional organization sensitivity and relation to perception,” *Human Neurobiology*, pp. 197–208, 1984.

[6] A. Takeuchi and K. Nagao, “Communicative facial displays as a new conversational modality,” in *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM Press, 1993, pp. 187–193.

[7] J. Cassell and K. R. Thórisson, “The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents,” *Applied Artificial Intelligence*, vol. 13, no. 4-5, pp. 519–538, 1999.

[8] P. Ekman and W. V. Friesen, “The repertoire of nonverbal behavior - categories, origins, usage, and coding,” *Semiotics 1*, pp. 49–98, 1969.

[9] A. J. Fridlund and A. N. Gilbert, “Emotions and facial expression,” *Science*, vol. 230, pp. 607–608, 1985.

[10] N. Chovil, "Communicative functions of facial displays in conversation," Ph.D. dissertation, University of Victoria, 1989.

[11] P. Ekman, "The argument and evidence about universals in facial expressions of emotion," in *Handbook of Social Psychophysiology.* John Wiley & Sons, 1989.

[12] C. Pelachaud, "Communication and coarticulation in facial animation," Ph.D. dissertation, University of Pennsylvania, 1991.

[13] J. Cassell, "Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents," in *Embodied conversational agents*, J. Cassell, S. Prevos, J. Sullivan, and E. Churchill, Eds. MIT Press, 2000, pp. 1–27.

[14] N. Chovil, "Discourse-oriented facial displays in conversation," *Research on Language and Social Interaction*, vol. 25, pp. 163–194, 1991.

[15] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," *Computer Graphics*, vol. 28, pp. 413–420, 1994.

[16] J. B. Bavelas and N. Chovil, "Visible acts of meaning. an integrated message model of language use in face-to-face dialogue," *Journal of Language and Social Psychology*, vol. 19, pp. 163–95, 2000.

[17] D. Decarlo, C. Revilla, M. Stone, and J. J. Venditti, "Making discourse visible: Coding and animating conversational facial displays," in *In Proceedings of Computer Animation*, 2002, pp. 11–16.

[18] D. Heylen, "Facial expressions for conversational agents," in *CHI Workshop on Subtle Expressivity*, 2003.

[19] G. Zoric, K. Smid, and I. S. Pandzic, "Facial gestures: Taxonomy and application of nonverbal, nonemotional facial displays for embodied conversational agents," in *Conversational Informatics: An Engineering Approach*, ser. Wiley Series in Agent Technology, T. Nishida, Ed. John Wiley & Sons, 2007, pp. 161–182.

[20] K. R. Scherer, *The functions of nonverbal signs in conversation.*, ser. The social and the psychological contexts of language. Hillsdale: NJ: Erlbaum, 1980.

[21] B. Zellner, “Pauses and the temporal structure of speech,” in *In.* John Wiley, 1994, pp. 41–62.

[22] W. J. Hardcastle and J. Laver, Eds., *The Handbook of Phonetic Science*, ser. Blackwell Handbooks in Linguistics. Blackwell, 1999.

[23] F. Grosjean and A. Deschamps, “Analyse contrastive des variables temporelles de l’anglais et du francais,” *Phonetica*, pp. 144–184, 1975.

[24] P. Ekman, “About brows: Emotional and conversational signals,” *Human ethology*, pp. 169–202, 1979.

[25] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, “About the relationship between eyebrow movements and f0 variations,” in *Proceedings of ICSLP*, H. T. Bunnell and W. Idsardi, Eds., 1996, pp. 2175–2178.

[26] T. Kuratate, K. Munhall, P. Rubin, E. Vatikiotis-Bateson, and H. Yehia, “Audio-visual synthesis of talking faces from speech production correlates,” in *EuroSpeech99*, vol. 3, 1999, pp. 1279–1282.

[27] K. Honda, “Interactions between vowel articulation and f0 control,” in *Proceedings of Linguistics and Phonetics: Item Order in Language and Speech*, B. D. J. O. Fujimura and B. Palek, Eds., 2000.

[28] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, “Facial animation and head motion driven by speech acoustics,” in *5th Seminar on Speech Production: Models and Data*, P. Hoole, Ed., 2000.

[29] K. G. Munhall, J. A. Jones, D. E. Callan, and E. Kuratate, T.and Bateson, “Visual prosody and speech intelligibility: Head movement improves auditory speech perception,” *Psychological Science*, vol. 15, pp. 133–137, 2004.

[30] B. Granström, D. House, and M. Lundeberg, “Eyebrow movements as a cue to prominence,” in *The Third Swedish Symposium on Multimodal Communication*, 1999.

[31] D. House, J. Beskow, and B. Granström, “Timing and interaction of visual cues for prominence in audiovisual speech perception,” in *Proceedings of Eurospeech*, 2001, pp. 387–390.

[32] E. Krahmer, Z. Ruttkay, M. Swerts, and W. Wesselink, “Pitch, eyebrows and the perception of focus,” in *In Symposium on Speech Prosody*, 2002.

[33] J. Cassell, *Embodied Conversational Agents.* MIT Press: Cambridge, 2000.

[34] J. Forlizzi, J. Zimmerman, V. Mancuso, and S. Kwak, “How interface agents affect interaction between humans and computers,” in *DPPI ’07: Proceedings of the 2007 conference on Designing pleasurable products and interfaces*, 2007, pp. 209–221.

[35] J. H. Walker, L. Sproull, and R. Subramani, “Using a human face in an interface,” in *CHI ’94: Conference companion on Human factors in computing systems*, 1994, p. 205.

[36] W. J. King and J. Ohya, “The representation of agents: anthropomorphism, agency, and intelligence,” in *CHI ’96: Conference companion on Human factors in computing systems*, 1996, pp. 289–290.

[37] V. Vinayagamoorthy, A. Brogni, M. Gillies, M. Slater, and A. Steed, “An investigation of presence response across variations in visual realism,” in *The 7th Annual International Presence Workshop*, 2004, pp. 148–155.

[38] M. Lundeberg and J. Beskow, “Developing a 3d-agent for the august dialogue system,” in *Proceedings of Audio Visual Speech Processing (AVSP)*, 1999, pp. 151–154.

[39] H. H. V. Justine Cassell and T. Bickmore, “Beat: the behavior expression animation toolkit,” in *Proceedings of the conference on Computer graphics and interactive techniques, SIGGRAPH*, 2001, pp. 477–486.

[40] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, “Visual prosody: facial movements accompanying speech,” in *In Proceedings of AFGR*, 2002, pp. 381–386.

[41] C. Pelachaud and M. Bilvi, “Computational model of believable conversational agents,” *Communications in Multiagent Systems*, vol. 20, pp. 300–317, 1993.

[42] I. Poggi, “Mind markers,” in *Gestures. Meaning and use*, M. R. N. Trigo and I. Poggi, Eds., 2002.

[43] T. D. Bui, D. Heylen, and A. Nijholt, “Combination of facial movements on a 3d talking head,” in *CGI ’04: Proceedings of the Computer Graphics International*, 2004, pp. 284–291.

[44] J. Lee and S. Marsella, “Nonverbal behavior generator for embodied conversational agents,” in *Intelligent Virtual Agents*, 2006, pp. 243–255.

[45] J. Lewis, “Automated lip-sync: Background and techniques,” *Visualization and Computer Animation 2*, 1991.

[46] D. McAllister, R. Rodman, D. Bitzer, and A. Freeman, “Lip synchronization of speech,” in *Proceedings of Audio Visual Speech Processing (AVSP)*, 1997.

[47] F. J. Huang and T. Chen, “Real-time lip-synch face animation driven by human voice,” in *IEEE Workshop on Multimedia Signal Processing*, 1998.

[48] S. Kshirsagar and N. Magnenat-Thalmann, “Lip synchronization using linear predictive analysis,” in *IEEE International Conference on Multimedia and Expo*, 2000.

[49] G. Zoric, “Automatic lip synchronization by speech signal analysis,” Master’s thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, 2005.

[50] M. Brand, “Voice puppetry,” in *Proceedings of the conference on Computer graphics and interactive techniques, SIGGRAPH*, 1999.

[51] R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, A. Bojorquez, J. Castillo, and I. Rudomin, “Speech-driven facial animation with realistic dynamics,” *IEEE Transactions on Multimedia*, 2005.

[52] Z. Deng, S. Narayanan, C. Busso, and U. Neumann, “Audio-based head motion synthesis for avatar-based telepresence systems,” in *Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence*, 2004, pp. 24–30.

[53] E. Chuang and C. Bregler, “Mood swings: Expressive speech animation,” *ACM Transactions on Graphics*, vol. 2, pp. 331–347, 2005.

[54] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, “Prosody-driven head-gesture animation,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[55] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Ozkan, “Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, pp. 1330–1345, 2008.

[56] G. Hofer and H. Shimodaira, "Automatic head motion prediction from speech data," in *In Proceedings of Interspeech*, 2007.

[57] I. Albrecht, J. Haber, and H. peter Seidel, "Automatic generation of non-verbal facial expressions from speech," in *In Proceedings of Computer Graphics International*, 2002, pp. 283–293.

[58] G. Zoric, K. Smid, and I. Pandzic, "Towards facial gestures generation by speech signal analysis using huge architecture," in *Multimodal Signals: Cognitive and Algorithmic Issues*, ser. Lecture Notes in Computer Science, LNCS, vol. 5398. Springer-Verlag, 2009, pp. 112–120.

[59] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," in *In proceedings of ACM SIGGRAPH Asia*, 2009.

[60] S. P. Lee, J. B. Badler, and N. I. Badler, "Eyes alive," in *Proceedings of the conference on Computer graphics and interactive techniques, SIGGRAPH*, 2002, pp. 637–644.

[61] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Nonverbal Communication*, 1974.

[62] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, "Kinematics of head movements accompanying speech during conversation," *Human Movement Science*, vol. 2, no. 1-2, p. 35–46, 1983.

[63] W. Condon and W. Ogston, "Speech and body motion synchrony of speaker-hearer," *Perception of language*, pp. 150–184, 1971.

[64] I. Poggi and C. Pelachaud, "Signals and meanings of gaze in animated faces," in *Language, Vision and Music: Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing*, 2002, pp. 133–144.

[65] M. Argyle and C. M., *Gaze and Mutual gaze.* Cambridge University Press, 1976.

[66] G. Collier, *Emotional expression.* Lawrence Erlbaum Associates, 1985.

[67] A. Kendon, *Some functions of gaze direction in social interaction.* Acta Psychologica 32, 1967.

[68] K. Smid, G. Zoric, and I. S. Pandzic, “[huge]: Universal architecture for statistically based human gesturing,” in *Proceedings of the 6th International Conference on Intelligent Virtual Agents IVA 2006*, 2006, pp. 256–269.

[69] G. Zoric, R. Forchheimer, and I. Pandzic, “On creating multimodal virtual humans - real time speech driven facial gesturing,” *accepted for publication in Multimedia tools and applications, Special issue on Multimodal interaction and Multimodal content Management*, 2010.

[70] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[71] “HTK Book,” http://htk.eng.cam.ac.uk/prot-docs/htk_ book.shtml/ (27.03.2010.).

[72] “HTK, The Hidden Markov Model Toolkit,” http://htk.eng.cam.ac.uk/ (27.03.2010.).

[73] “SPTK, The Speech Signal Processing Toolkit,” http://sp-tk.sourceforge.net/ (27.03.2010.).

[74] “ATK, Real-time API for HTK Toolkit,” http://htk.eng.cam.ac.uk/develop/atk.shtml (27.03.2010.).

[75] “ATK Manual,” http://mi.eng.cam.ac.uk/research/dialogue/atk_ manual.pdf (27.03.2010.).

[76] “Visage Technologies AB,” http://www.visagetechnologies.com/ (27.03.2010.).

[77] A. Axelsson and E. Bj/:orhall, “Real time speech driven face animation,” Master’s thesis, The Image Coding Group, Dept. of Electrical Engineering at Link/:oping University, 2003.

[78] I. S. Pandzic and R. Forchheimer, Eds., *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[79] M. E. Zavala Chmelicka, “Visual prosody in speech-driven facial animation: elicitation, prediction, and perceptual evaluation,” Master’s thesis, Texas A & M University, 2005.

[80] "Fraps, Realtime Video Capture Software," http://www.fraps.com/ (07.04.2010.).

[81] "Student t-test calculator," http://www.graphpad.com/quickcalcs/ttest1.cfm?Format=C (08.04.2010.).

[82] "Student t-test, p value," http://www.stattutorials.com/p-value-interpreting.html (08.04.2010.).

[83] N. Magnenat-Thalmann and P. Kalra, "The simulation of a virtual tv presentor," in *Proc. Pacific Graphics '95.* World Scientific Press, 1995, pp. 9–21.

[84] "Ananova," http://thescreamonline.com/technology/technology10-01/index.html (13.04.2010.).

[85] "News At Seven," http://infolab.northwestern.edu/projects/news-at-seven/ (13.04.2010.).

[86] T. Noma and N. I. Badler, "A virtual human presenter," in *Proceedings of the IJCAI Workshop on Animated Interface Agents*, 1997, pp. 45–51.

[87] F. Yun and Z. NanNing, "Advanced framework for an error-resilient parameter analysis-synthesis system of facial animation," in *IEEE International Conference on Systems, Man and Cybernetics*, 2003, pp. 4528– 4534.

[88] J. Ostermann, J. Rurainsky, and R. Civanlar, "Real-time streaming for the animation of talking faces in multiuser environments," *IEEE International Symposium on Circuits and Systems*, 2002.

[89] M. Fresia, C. Bonamico, F. Lavagetto, and R. Pockaj, "An encoding/packetization solution for mpeg-4 facial animations delivery over rtp," in *Proceedings of CNIT Advanced Multimedia Workshop*, 2002.

[90] S. T. Worrall, A. Sadka, and A. M. Kondoz, "3-d facial animation for very low bit rate mobile video," 2002.

[91] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video," 2005.

[92] M. Matijasevic and I. S. Pandzic, "A challenge for interactive virtual characters on the internet," in *Proceedings of the 10th International Conference on Telecommunication Systems, Modeling and Analysis*, 2002.

[93] “Real time Transport Protocol, RTP,” http://www.ietf.org/rfc/rfc3550.txt (14.04.2010.).

Glossary of abbreviations

3D	Three Dimensional
AV	Audio to Visual
ECA	Embodied Conversational Agent
F0	Fundamental frequency
FAPs	Face Animation Parameters
FAPUs	Facial Animation Parameter Units
FBA	Facial and Body Animation
FPs	Feature Points
HCI	Human-Computer Interaction
HHI	Human-Human Interaction
HMM	Hidden Markov Model
IP	Internet Protocol
KNN	K-Nearest Neighbors
LAN	Local Area Network
LPC	Linear Prediction Coefficient
PCM	Pulse-Code Modulation
QoS	Quality of Service
RMS	Root Mean Square
RTP	Real time Transport Protocol
MFCC	Mel-Frequency Cepstral Coefficients
MNS0	Mouth-Nose Separation units
MPEG	Moving Picture Experts Group
NN	Neural Network
TTS	Text-To-Speech
VCs	Virtual Characters
VTTS	Visual Text-To-Speech
XML	Extensible Markup Language

Summary

A hybrid approach to real-time speech driven facial gesturing of virtual characters

This thesis investigates automatic facial gesturing of virtual characters based only on a speech input. Facial gestures that have been included are various nods and head movements, eye blinks, eyebrow gestures and gaze, i.e. facial and head movements used in nonverbal communication that are connected to the syntactic and prosodic structure of the underlying speech rather than to semantic or emotions.

The main issue in speech driven facial gesturing is an audio to visual mapping. In this thesis, a hybrid method for the real-time correlation of the speech signal with facial gestures has been proposed. An idea behind the proposed method is to include different issues that are considered important for visual prosody through three different modules. In a data-driven module facial gestures are correlated with prominent parts of the speech. A rule-based module consists of a set of rules, which among others include rules for punctuation and prolonged pauses. Additionally, in a statistics module, results of first two modules are fine-tuned and further shaped.

In order to check how believable virtual characters are if animated using facial gestures obtained by the proposed method, a system for speech driven facial gesturing has been implemented, and used for a perceptual evaluation. Testing subjects generally considered facial gestures consistent and time aligned with the underlying speech, and a virtual human applying them is perceived as a natural. Moreover, the set up hypothesis has been proved: adding facial gestures driven by the speech signal to the animation of virtual characters using the proposed hybrid approach in real-time will result in more coherent and appropriate facial movements than random or no movements.

Further more, the system for speech driven facial gesturing of virtual characters has been put in a context of virtual news presenters used in networked environments, since that is where the future research of this system strives next.

Sažetak

Hibridni pristup stvaranju gesti lica virtualnog lika govornim signalom u stvarnom vremenu

Ova doktorska disertacija istražuje automatsko stvaranje gesti virtualnog lika govornim signalom u stvarnom vremenu. Pokreti lica uključeni u ovaj rad su različiti pokreti glave i obrva, treptanje i pogled, tj. svi pokreti lica i glave koji se koriste u neverbalnoj komunikaciji povezani sa sintaksom i prozodijom govora, a ne s emocijama ili značenjem riječi.

Kod generiranja animacije lica iz govora potrebno je naći korelaciju između govornog signala i pokretanja lica. Ta korelacija nije direktna, već je posljedica više elemenata, te postoje individualne i kulturološke specifičnosti. U disertaciji su istražene različite metode za preslikavanja iz audio informacije u vizualnu. Odabrano rješenje temelji se na hibridnom pristupu, pri čemu su statistički ispravni pokreti lica i glave usklađeni sa prozodijom govora u stvarnom vremenu dobiveni tehnikom strojnog učenja, korištenjem pravila te statistikom dobivenom iz baze podataka. Tehnikom strojnog učenja dobiveni su pokreti lica koji su povezani s naglašenim dijelovima govora, dok su pomoću pravila, između ostalog, dobiveni pokreti lica povezani s različitim pauzama u govoru. Ovi rezultati su dodatno oblikovani statistikom dobivenom iz baze podataka.

Kako bi provjerili pokrete lica virtualnog lika, izgrađen je sustav za automatsko pokretanje lica govornim signalom zasnovan na predloženoj hibridnoj metodi, te je iskorišten za subjektivnu evaluaciju. Sudionici testiranja su doživjeli pokrete lice uglavnom kao vremenski usklađene, te konzistentne s govorom, dok je općeniti dojam da virtualni lik djeluje prirodno. Također, potvrđena je postavljena hipoteza: pokreti lica iz govora dodani animaciji virtualnog lika korištenjem predložene hibridne metode u realnom vremenu daju bolje rezultate nego slučajni pokreti ili lice bez pokreta.

Tako izgrađen sustav stavljen je u kontekst virtualnih prezentatora vijesti koji se koriste u umreženim okruženjima, budući da daljnje istraživanje ove problematike ide u tom smjeru.

Keywords

- Facial gestures
- Visual prosody
- Virtual characters
- Embodied Conversational Agents
- Speech signal
- Facial animation
- Nonverbal communication
- Audio-visual speech processing
- Human-Computer interaction
- HMMs

Ključne riječi

- Geste lica
- Vizualna prozodija
- Virtualni likovi
- Utjelovljeni razgovorni agenti
- Govorni signal
- Animacija lica
- Neverbalna komunikacija
- Audio-vizualna obrada govora
- Interakcija čovjeka i računala
- HMM

Biography

Goranka Zorić was born in Zagreb in 1978. She finished mathematical and natural sciences high school (V. gymnasium in Zagreb) in 1997. The same year she continued her education at the Faculty of Electrical Engineering and Computing, University of Zagreb. She received her B.S. (Dipl.-Ing.) and M.S. degrees in electrical engineering with a major in telecommunications and information science from the University of Zagreb in 2002 and 2005, respectively. Her diploma thesis topic was "Mobility of agents in IPv6 Network", and her master thesis topic was "Automatic Lip Synchronization by Speech Signal Analysis". In 2006 she started doctoral program at the same faculty. She was employed at the Department of Telecommunications at Faculty of Electrical Engineering and Computing as research associate from 2002 till 2010, working on different projects focused on building believable virtual humans, including the project "Embodied Conversational Agents as interface for networked and mobile services" funded by the Ministry of Science, Education, and Sports of the Republic of Croatia. In 2010 she started working at the Mobile Life Centre at Stockholm University, Sweden as a senior researcher.

In 2005 she worked as a visiting scientist at the department of Signal Theory and Communications (TSC) at the Technical University of Catalonia, Barcelona, Spain. Two months research fellowship was provided by European project SIMILAR. During 2008/2009 she stayed for 15 months as a guest researcher at Linköping University, Department of Electrical Engineering, Division of Information Coding, Linköping, Sweden, working on the topic audio to visual mapping for facial gesturing. Her stay was supported by grants from The National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia, and The Swedish Institute, Sweden.

She has published 18 scientific papers with international review, out of which one book chapter and 7 journal papers.

Životopis

Goranka Zorić rođena je 1978. godine u Zagrebu. Prirodoslovno matematičku srednju školu (V. gimnazija u Zagrebu) završila je 1997. godine. Iste godine upisuje Fakultet elektrotehnike i računarstva, Sveučilišta u Zagrebu. Diplomirala je 2002. godine te magistrirala 2005. godine na smjeru telekomunikacije i informatika. Tema njenog diplomskog rada je "Pokretljivost agenata u IPv6 mreži", a magistarskog rada "Automatska sinkronizacija usana pomoću analize govornog signala". Doktorski studij na istom smjeru upisuje 2006. godine. Na Zavodu za telekomunikacije, Fakulteta elektrotehnike i računarstva bila je zaposlena kao zavodski suradnik od 2002. do 2010, radeći na različitim projektima vezanim uz izgradnju virtualnih ljudi, a između ostalog sudjelovala je i na znanstvenom projektu Ministarstva znanosti i tehnologije Republike Hrvatske pod nazivom "Utjelovljeni razgovorni agenti za usluge u umreženim i pokretljivim sustavima". Godine 2010. zaposlila se u *Mobile Life Centre*, na *Stockholm University*, Švedska.

Godine 2005. boravila je kao gostujući znanstvenik na zavodu *Signal Theory and Communications (TSC)*, na *Technical University of Catalonia*, Barcelona, Španjolska. Dvomjesečnu stipendiju dobila je preko evropskog projekta SIMILAR. Za vrijeme 2008./2009. godine boravila je 15 mjeseci kao gostujući znanstvenik na *Linköping University, Department of Electrical Engineering, Division of Information Coding*, Linköping, Švedska gdje je obavila dio doktorskog istraživanja. Usavršavanje je bilo omogućeno preko stipendije za doktorande Nacionalne zaklade za znanost, visoko školstvo i tehnologijski razvoj Republike Hrvatske, te stipendije instituta *The Swedish Institute*, Švedska.

Objavila je 18 radova s međunarodnom recenzijom od čega jedno poglavlje u knjizi i 7 radova u međunarodno priznatim časopisima.