Multimed Tools Appl
DOI 10.1007/s11042-010-0526-y

On creating multimodal virtual humans—real time speech driven facial gesturing

Goranka Zoric · Rober Forchheimer · Igor S. Pandzic

Abstract Because of extensive use of different computer devices, human-computer interaction design nowadays moves towards creating user centric interfaces. It assumes incorporating different modalities that humans use in everyday communication. Virtual humans, who look and behave believably, fit perfectly in the concept of designing interfaces in more natural, effective, as well as social oriented way. In this paper we present a novel method for automatic speech driven facial gesturing for virtual humans capable of real time performance. Facial gestures included are various nods and head movements, blinks, eyebrow gestures and gaze. A mapping from speech to facial gestures is based on the prosodic information obtained from the speech signal. It is realized using a hybrid approach—Hidden Markov Models, rules and global statistics. Further, we test the method using an application prototype—a system for speech driven facial gesturing suitable for virtual presenters. Subjective evaluation of the system confirmed that the synthesized facial movements are consistent and time aligned with the underlying speech, and thus provide natural behavior of the whole face.

Keywords Facial gestures · Visual prosody · Multimodal interfaces · Facial animation · Speech processing · Human-computer interaction

1 Introduction

The future applications are expected to be user-oriented and efficient. Having in mind how much time humans generally spend in interaction with different computing devices, the benefit is obvious. One way to approach this issue is through virtual humans, already researched in

G. Zoric (✉) · I. S. Pandzic
Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10 000 Zagreb, Croatia
e-mail: Goranka.Zoric@fer.hr

I. S. Pandzic
e-mail: Igor.Pandzic@fer.hr

R. Forchheimer
Department of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden
e-mail: robert@isy.liu.se

Published online: 29 April 2010

human-computer interaction for almost two decades. Interfaces designed based on virtual humans have potential to provide interaction with computers similarly to human-human interaction. However, it is important to properly map the way humans communicate to virtual humans. If not, the results might be just contrary—unpleasant or annoying.

When communicating, humans exchange information using different modalities to transmit messages (multimodal communication). Major modalities are seeing or vision and hearing or audition, but any human sense can be translated to modality (e.g. taste, smell, touch, pain etc.). Virtual human interfaces available today use multimodal output, primarily in the form of visual and auditory cues, while inputs usually used are speech, text or gestures. In the conditions when one communicative channel experience interference (such as noisy environment) or in cases of human disability (such as hearing impairment), human-computer interaction benefits from multimodal output due to information redundancy.

In this paper we present a multimodal interface, based on virtual human, which uses speech as input and speech and facial gestures as output. Output facial gestures are obtained using speech signal analysis and correspondent prosodic information. Visual output includes lip synchronization and facial gestures—different head and eyebrow movements, blinks and gaze.

Our main contribution is a method for automatic audio to visual mapping which uses a hybrid data-driven and rule-based approach to produce wide set of the facial gestures only from the speech signal in the real time.

The paper is organized as follows: Section 2 presents background of speech driven facial gesturing and related work. Section 3 is about the method used for mapping from speech signal to facial gestures. Section 4 describes a system for speech driven facial animation suitable for virtual presenters, including results of subjective evaluation and Section 5 concludes the paper and suggests future works.

2 Background and related work

The human face is an important communication channel in face-to-face communication. Through the face, diversity of signals is displayed—verbal, emotional or conversational. All of them should be carefully modeled in order to have natural looking facial animation. Speech driven facial animation has been researched much during the past decade. For human-like behavior of talking head, all facial displays need to be accurately simulated in synchrony with an underlying speech signal. Most efforts, so far, have focused on articulation of lip and tongue movements (lip synchronization), since those movements are necessary in order to have speech intelligibility. Also, emotional displays were investigated in great extent.

However, during speech articulation, we continuously use different facial gestures (also known as visual prosody), a form of nonverbal communication made with the face or head, instead of, or in combination with verbal communication. In everyday communication humans use facial gestures consciously or unconsciously to regulate flow of speech, accentuate words or segments, or punctuate speech pauses. Facial gestures include various nods and head movements, blinks, eyebrow gestures and gaze, as well as frowning, nose wrinkling or lips moistening [26].

Speech related facial gestures are connected with prosody and paralinguistic information. Speech prosody refers to characteristics of speech signal, such as intonation, rhythm, and stress, which cannot be extracted from the characteristics of phoneme segments. Its acoustical correlates are fundamental frequency (pitch), intensity (loudness) and syllable length. Paralinguistic refers to the nonverbal elements of communication used to modify meaning and convey emotion and includes pitch, volume and, in some cases, intonation of speech.

While relation between speech and lip movements is obvious (e.g. phoneme-viseme), the relation between facial gestures (also gestures in general) and speech isn't so strong. Moreover, variations from person to person are large. Still, there is some psychological and paralinguistic research relevant to synthesizing natural behavior of virtual humans with accent on facial gestures.

2.1 Related psychological and paralinguistic research

Ekman investigated systematically the relation of speech and eyebrow movement in [8]. According to his findings, eyebrow movements occur during word searching pauses, as punctuators or to emphasize certain words or parts of the sentence. During thinking pauses (searching for word), raised eyebrows occur accompanied by an upward gaze direction. Eyebrows also may be lowered in this situation, especially in conjunction with filled pause ('errr'). Chovil in [3] concentrated on the role of facial displays in conversation. Her research results showed that syntactic displays (emphasized words, punctuators) are the most frequent facial gestures accompanying speech. Among those facial gestures, raising or lowering eyebrows are the most relevant. Cosnier in [6] showed that eyebrow raising in particular plays a role in question asking.

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

Cavé et al. in [2] investigated links between rapid (rising and falling) eyebrow movement and fundamental frequency (F0) 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 dialogs, assure the speaker of the listener's attention, and mirror the listener's degree of understanding, serving as back channel.

Kuratate et al. in [15] presented preliminary evidence for close relation between head motion and acoustic prosody (fundamental frequency). They concluded that production systems of the speech and head motion are internally linked. These results were further elaborated in [12, 17, 24] within the same group, adding head motion correlation with the amplitude (root mean square, RMS, amplitude) of the subjects' voice.

Granström et al. investigated in [10] contribution of eyebrow movement to the perception of prominence. Later, head movements were added, indicating that combined head and eyebrow movements are effective cues to prominence when synchronized with the stressed vowel [13] stating that perceptual sensitivity to timing is around 100 ms to 200 ms, which is about the average length of a syllable.

2.2 Related facial animation systems

State of the art literature lacks methods that would be able to automatically generate a complete set of facial gestures, including head movements, by only analyzing the speech signal. Existing systems in this field mainly concentrate on a particular aspect of facial gesturing, or on a general dynamics of the face. What follows, is the description of some relevant speech driven systems that include visual prosody.

Many works generate only head movements.

An automatic data-driven system for head motion synthesis is developed in [7], taking pitch, the lowest five formants, MFCC and LPC as audio features. A K-Nearest Neighbors

(KNN)-based dynamic programming algorithm is used to synthesize novel head motion given a new audio input.

Chuang and Bregler in [4] generate head motion in addition to expressive facial animation from speech. They use a database of examples that relate audio pitch to 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 [21] 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 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 [11] generates a sequence of head motion units using Hidden Markov Models given some speech based on the thesis that temporal properties should be taken into account and therefore the data has to be segmented into longer parts.

Albrecht et al. in [1] introduce a method for automatic generation of several nonverbal facial expressions from speech: head and eyebrow raising and lowering dependent on the 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. The intensity of facial expressions is additionally controlled by the power spectrum of the speech signal, which corresponds to the loudness and the intensity of the utterance.

Speech driven systems described so far need a preprocessing step.

Real time speech driven facial animation is addressed in [27]. 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 [16] 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. Hidden Markov Model drives the selection and speech features used are pitch, intensity and syllable duration.

SyncFace system [20] aims to enhance speech perception through visible articulation and nonarticulatory facial movements. They include speech-related facial movements linked to emphasis or prominence and those related to interaction control in a dialogue situation.

Our system works in real time and is capable of producing wider set of facial gestures as described in the next chapter.

3 The proposed hybrid approach to speech driven facial gesturing

3.1 Facial gestures included

Based on the information found in literature [8, 9, 19, 26], and as presented in previous chapter, in our work, we include the following gestures:

- **Nod.** That is an abrupt swing of the head with a similar abrupt motion back [9]. 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.

- **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 [9]. Possible directions are up, down, left, right and diagonal. It occurs at increased speech dynamics (when the 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 slow head motion coming to neutral position.
- **Eyebrow movements** are present generally in two variations. Eyebrow raise (eyebrows go up and down) is often used to accentuate a word or a sequence of words, while eyebrow frown (eyebrows go down and up) might appear at hesitation pauses or when thinking [8]. Besides, they appear frequently to mark a period (“,”).
- **Blinks.** They are described as rapid closing and opening of one or both eyes that 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 [5, 19].
- **Gaze.** The level of gaze falls at the hesitation pause, when thinking what to say (aversion of gaze—the speaker looking away from listener), while it rises at the end of an utterance in order to collect feedback from the listener (eye contact—the speaker is steadily looking toward to the listener for a period of time).

3.2 Training database

As a starting point, we had a database with annotated facial gestures. It consists of 56 news video clips (3 female and 2 male speakers) with 13 min total duration.

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, MNS0, units).

3.3 Proposed method

The key issue in speech driven facial gesturing is to find a mapping between speech features and facial gestures. In this work additional constraint was real time performance (no preprocessing step).

Our solution to this problem is a method that uses a hybrid data-driven and rule-based approach and runs in three subsequent steps. Such approach is chosen to overcome the lack of data that we were faced when tried to classify gestures into its types and subtypes. Instead, facial gesture is first classified into four basic types (blink, eyebrow movement, nod and swing) using Hidden Markov Models (HMMs). Than, it is additionally fine tuned using rules generated from the information found in the literature on nonverbal

communication. At the end, each facial gesture is further characterized with its subtype, amplitude and duration based on the statistics from the training database.

Since system requirement is real time performance, calculations in this work are based only on the preceding speech signal. Speech features are calculated for every frame. Frame length is 16 ms. The speech signal is classified into one of the existing groups every four frames giving maximal delay in gesture generation less than 80 ms.

The training of gesture HMMs is done in similar manner: four audio frames preceding the beginning of the corresponding gesture are used instead of using the frames that are actually covering the gesture (Fig. 1).

Such approach is chosen as a possible optimal solution and balance between severe real time requirements and complex timing relation between speech features and appearance of gestures. We were motivated by the human nature—before a gesture is shown, our brain already "knows" that the gesture will be done, so the assumption is that some info about coming gesture might be found in the speech signal which just precede that gesture. With such pretty simplified approach we can only roughly determine the possible gesture appearance, so we additionally apply rules and gesture statistics to fine tune results.

Our main idea in this work was to create statistically correct facial gesturing, which generally follows underlying speech signal within given real time requirements.

3.3.1 Hidden Markov model module

In the first step, incoming speech is classified every four frames as blink, eyebrow movement, nod, swing or neutral. For that purpose five continuous density left-right HMM models are trained (Fig. 2).

An acoustic feature vector calculated for every frame consists of fundamental frequency (F0), root mean square (RMS) amplitude and its delta and acceleration coefficients. An idea is to model speech signal that we observe and that is 4 frames of speech prosodic parameters. Having that in mind we model HMMs with 4 emitting states with left-right transition probabilities—every state represents one speech frame and it is only possible to move from one speech frame to the following one. We use three data streams (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 most connected to the facial gestures (visual prosody), the data stream consisting of delta coefficients is given the biggest weight.

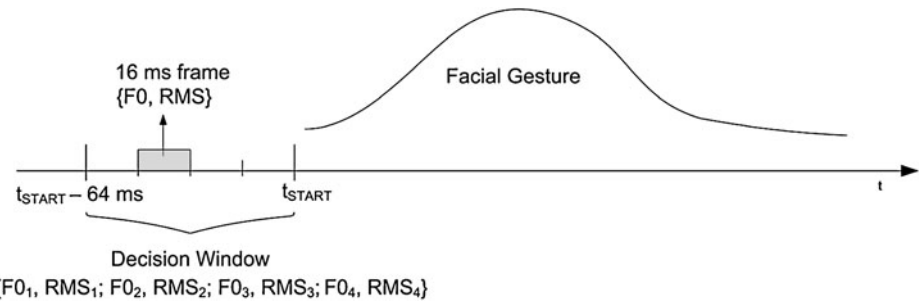

Fig. 1 Speech frames for analysis and the facial gesture shown on the time line

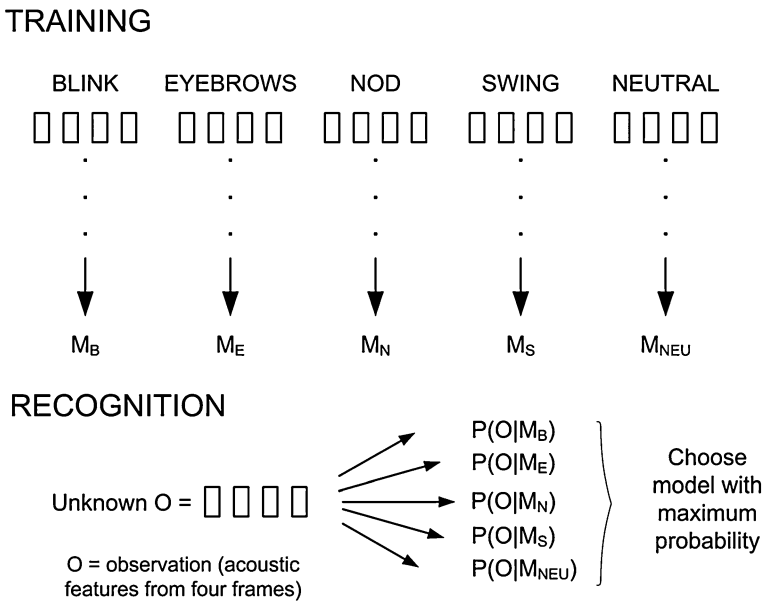

Fig. 2 Facial gesture HMM models, training and recognition

In order to observe each gesture in the context of the previous ones and in such way obtain statistically correct models, we use n-grams (sequence of n words/symbols) language models to predict each symbol in the sequence given its n-1 predecessors (i.e. 3-grams: -0.8405 BLINK EYEBROWS NEUTRAL).

3.3.2 Communication module incorporating different rules

In the second step, the set of facial gestures obtained by HMM models is shaped using different rules. Gestures are added or deleted having in mind the following: the gesture currently obtained by HMM module, gestures previously obtained by HMM module, but which still haven't finish, last occurrence for every gesture type and rules for each gesture type found in the literature (it's frequency of occurrence or gestures that often comes together).

If the gesture, which HMM models have chosen as the most probable one for the specific moment, is of the same type as the gesture that was chosen before, but still didn't finish, than that gesture is discarded.

Rules included are:

- **Gaze direction in accordance with head movements.** In gaze following animation model, the eyes of our virtual human are moving in opposite directions of a head movement, if head movement amplitude is smaller than the defined threshold.
- **Small head movements in periods where there is no nods or swings**. Since the head is never still, we add random head movements in accordance with the speech amplitude.
- **Blinking as manipulator.** In addition to voluntary eye blinks, we have implemented periodic eye blinks based on the values given in [19]. If an eye blink does not happen within 5 s, it will be added.
- **Gaze, eyebrow and head movements during thinking and word-search pauses.** Gaze up or down with correspondent head and eyebrow movement are added.

- **Head and eyebrow movements and blinking as punctuators**. Head nod, eyebrow raise and blinking are added.

The pause detection algorithm is based on the RMS amplitude value and we differentiate long and short pauses. Using the obtained amplitude value for every frame and thresholds set in the initial frames, each frame can be potentially marked as pause in the speech. If pause is identified in four consecutive frames, it is classified as a short one. When pause in speech is longer than 32 frames, it is classified as a long pause. Such distinction is motivated by different kinds of pauses that occur during speech. A short pause stands for a punctuator, meaning that its role is to separate or group sequences of words, while a long pause corresponds to a thinking and word-search pause.

3.3.3 Statistics module

After HMMs and communication module, we know a type of gesture and its start time. Further, statistic for gesture subtype, amplitude and duration is calculated based on the information collected from the training video clips.

Subtype frequency is calculated as frequency of occurrence based on the total number of certain gesture and the number of the given subtype. For amplitude and duration calculation the range of possible values is divided into 6 groups (intervals). For each group, gestures with suitable duration and amplitude are counted and frequency of occurrence for each group is calculated based on the total number of certain gesture and the number of the gestures in the group.

This process aims to produce the global statistics based on the existing database for the frequency of occurrence, amplitude (in Mouth-Nose Separation unit) and duration of various gestures (in ms). Figure 3 shows statistics for different properties of the gesture (subtype, duration and amplitude) for blinks, eyebrow movements, nods and swings.

4 System overview and validation

In this section a system for speech driven facial gesturing for virtual humans is described (an overview is shown on Fig. 4). We also include results of preliminary subjective evaluation.

The system works in two phases: training and runtime.

4.1 Training phase

In the system preparation process, we need to go through training phase only once for every data that we want to use.

Inputs of this phase are audio clips (PCM format, .wav file, 16 kHz sample rate, 16 bit sample size) corresponding to training video clips and manually annotated facial gestures corresponding to training video clips in XML format, as described in Section 3.2. One of the outputs are HMM models with corresponding resources needed for later use. Also, database gesture statistics is generated and it includes number of gestures identified in database and corresponding subtype, duration and amplitude statistic.

As noted before, the acoustic feature vector used for training HMMs consists of two prosodic features, pitch and intensity. For pitch calculation SPTK (Speech Processing ToolKit) is used [22]. Cepstrum method is used to calculate the pitch period values corresponding to frames of input data. For unvoiced frames the output value is zero, and for

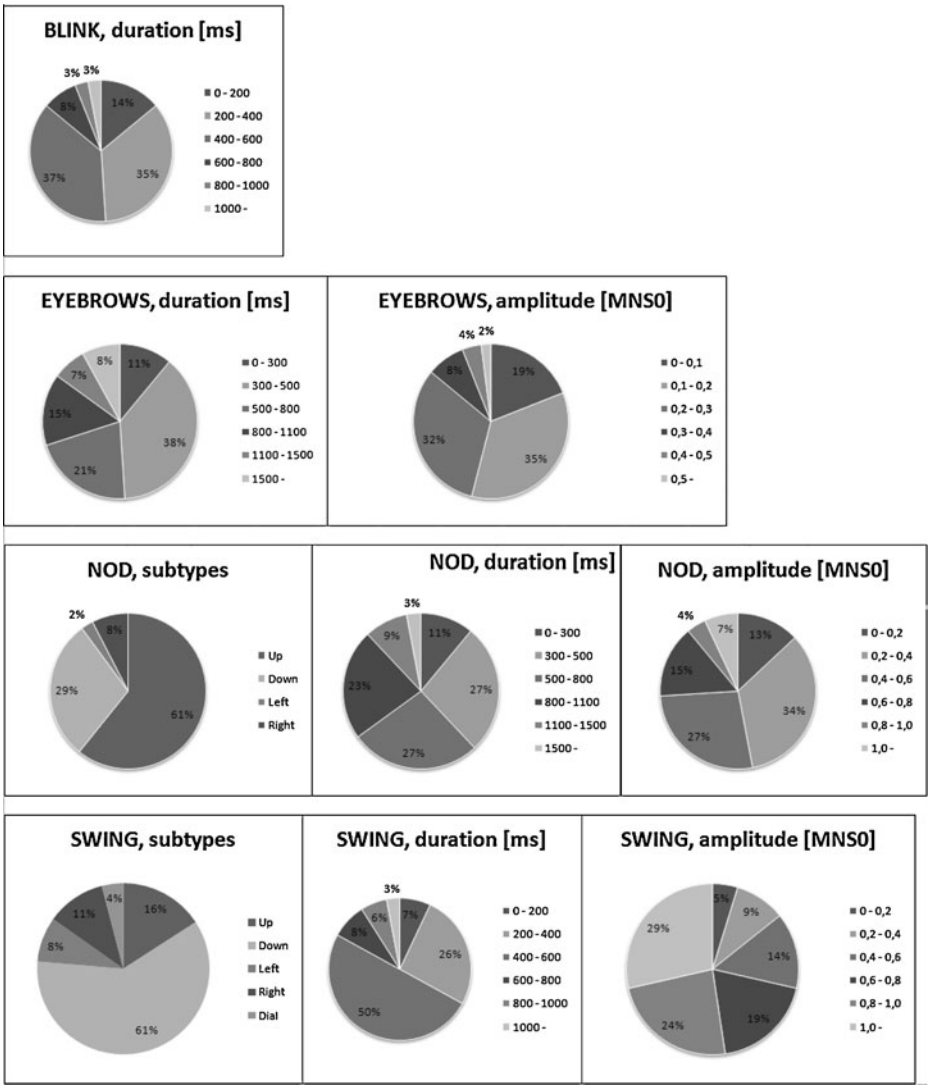

Fig. 3 Facial gesture statistics—subtype, duration and amplitude of the gesture

voiced frames, the output value is proportional to the pitch period. HMM training is performed using HMM ToolKit, HTK [14]. HMMs are modeled using label files, which were extracted from XML files with annotated gestures, and corresponding feature vectors. As an output, besides HMMs, dictionary and language model are created and all together are called HMM Resource Group.

4.2 The runtime phase

The runtime phase runs in real time and is fully automatic. This phase takes a new speech signal, triggers HMM, communication and statistical modules and produces set of facial gestures with all details (subtype, duration and amplitude) every four frames

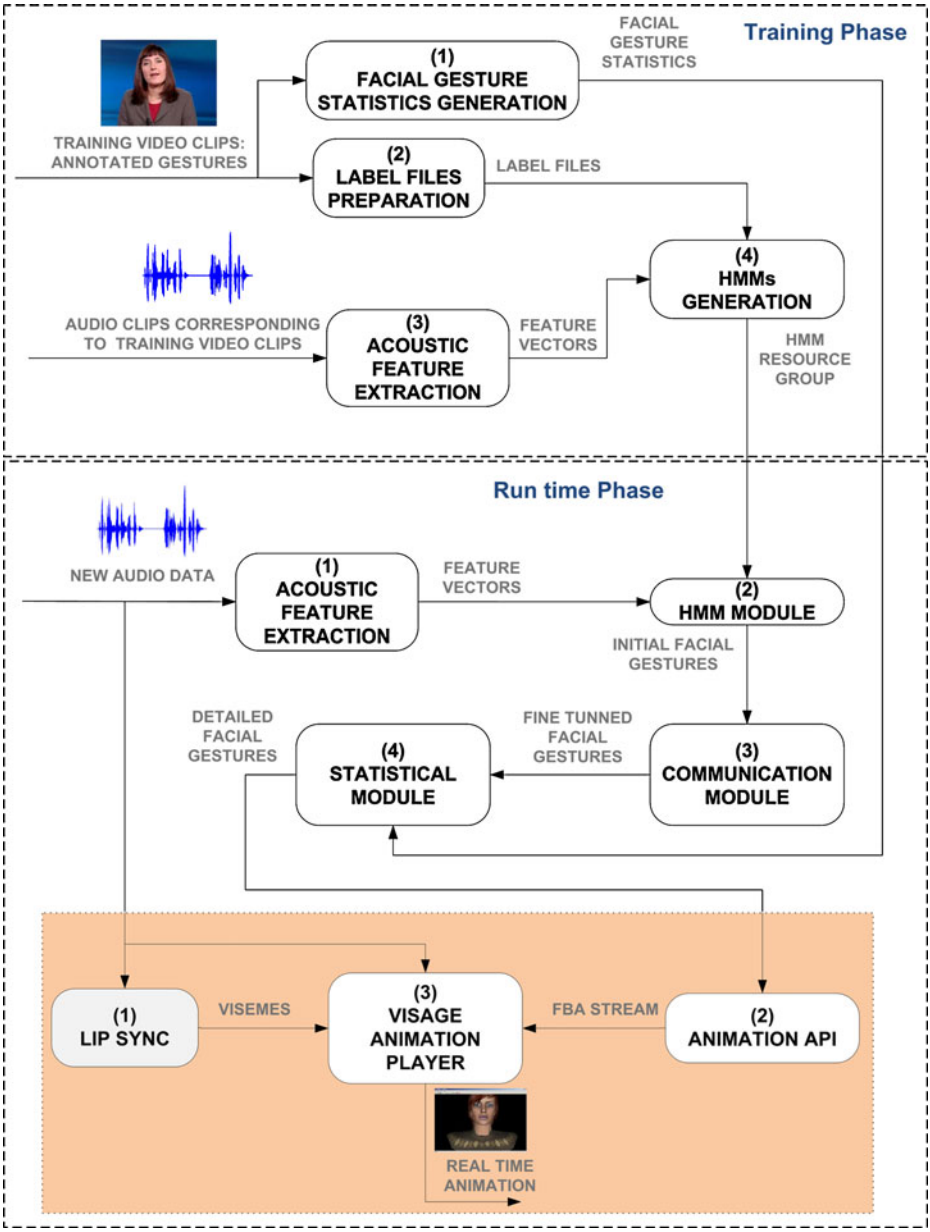

Fig. 4 System overview

(as explained in the previous Section). Those gestures are turned into MPEG-4 FBA (Facial and Body Animation) bitstream using Animation API and together with visemes (Lip Sync) are used in an animation player to animate the face of virtual human (shaded part).

More details about lip synchronization can be found in [25]. As animation player the Visage|SDK is used [23]. Based on MPEG-4 standard, it merges two tracks—calculated

visemes and FBA stream containing facial gestures and renders it on the screen. Once the required information is extracted from the speech any parameterized face model can be animated. More details on the MPEG-4 standard (i.e. MPEG-4 compatible 3D faces, MPEG-4 FA player) can be found in [18].

Figure 5 shows snapshots of facial gestures generated from speech signal.

4.3 System validation

To validate our system we use subjective evaluation. In it we aim to compare facial animation produced using our statistical method with two others—the first one containing facial gestures directly copied from the training video clips and the second one with the facial gestures produced randomly.

Using three previously described methods and a single audio file we are able to synthesize a facial animation. With the screen capture, we than create videos and present them to testing subjects. For every video, subjects are asked to rate facial movements using a scale, which specifies level of agreement to a statement. Statements used in our questionnaire are:

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

Preliminary results confirmed that the facial movements synthesized using our statistical method are generally consistent and time aligned with the underlying speech and therefore a virtual human is perceived natural and pleasant.

Fig. 5 Facial gesture snapshots—neutral, head movement, eyebrow raise and blink

5 Conclusion

In this paper we have presented a method for creating facial gestures for virtual humans in real time from a speech signal. It uses a hybrid approach—HMMs, rules and global statistics to generate head and eyebrow movements, blinks and gaze from the speech prosody. Based on such audio to visual mapping, a prototype system for speech driven facial gesturing is developed. Initial results obtained from subjective evaluation show that incorporating nonverbal behavior in addition to lip synchronization contributes in more positive perception of virtual human.

However, thoroughly evaluation (using more subjects and different audio files) of the method as well as of the whole system is needed in order to get opinion of potential users. It will help us to come closer in building believable virtual humans. Some improvements that we plan to work on are adding more gestures such as head posture or lip moistening, improving facial animation (e.g. for gaze since it contributes a lot to naturalness of the face) and expanding facial animation (e.g. for wrinkles). Further, since our database consists of videos with only news presentations, virtual humans built using our system for speech driven facial gesturing are suitable for use as virtual news presenters. Extension on the video clips covering broader topics as well as its enlargement in necessary for our system to be applicable for wider spectra of applications.

Acknowledgments The work was partly carried out within the research project "Embodied Conversational Agents as interface for networked and mobile services" supported by the Ministry of Science, Education and Sports of the Republic of Croatia. This work was partly supported by grants from The National Foundation for Science, Higher Education and Technological Development of the Republic of Croatia and The Swedish Institute, Sweden.

References

1. Albrecht I, Haber J, Seidel H (2002) Automatic generation of non-verbal facial expressions from speech. In Proceedings of Computer Graphics International 2002 (CGI 2002), pages 283–293
2. Cavé C, Guaïtella I, Bertrand R, Santi S, Harlay F, Espesser R (1996) About the relationship between eyebrow movements and F0 variations, In Proceedings of Int'l Conf. Spoken Language Processing
3. Chovil N (1991) Discourse-oriented facial displays in conversation. Research on Language and Social Interaction
4. Chuang E, Bregler C (2005) Mood swings: expressive speech animation. ACM Trans Graph 2:331–347
5. Condon WS, Ogston WD (1971) Speech and body motion synchrony of speaker-hearer. In Horton DL, Jenkins JJ (eds) Perception of language, 150–184
6. Cosnier J (1991) Les gestes de la question. In Kerbrat-Orecchioni, editor, La Question, 163–171, Presses Universitaires de Lyon
7. Deng Z, Busso C, Narayanan S, Neumann U (2004) Audio-based head motion synthesis for avatar-based telepresence systems. Proc ACM SIGMM Workshop on Effective Telepresence (ETP), 24–30
8. Ekman P (1979) About brows: Emotional and conversational signals. In von Cranach M, Foppa K, Lepenies W, Ploog D (eds) Human ethology: Claims and limits of a new discipline.
9. Graf HP, Cosatto E, Strom V, Huang FJ (2002) Visual prosody: facial movements accompanying speech. In Proceedings of AFGR 2002, 381–386
10. Granström B, House D, Lundeberg M (1999) Eyebrow movements as a cue to prominence. In The Third Swedish Symposium on Multimodal Communication
11. Hofer G, Shimodaira H (2007) Automatic head motion prediction from speech data, in In Proceedings of Interspeech
12. Honda K (2000) Interactions between vowel articulation and F0 control. In Proceedings of Linguistics and Phonetics: Item Order in Language and Speech (LP'98). Fujimura BDJO, Palek B (eds)
13. House D, Beskow J, Granström B (2001) Timing and interaction of visual cues for prominence in audiovisual speech perception. In Proceedings of Eurospeech

14. HTK, The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/
15. Kuratate T, Munhall KG, Rubin PE, Vatikiotis-Bateson E, Yehia HC (1999) Audio-visual synthesis of talking faces from speech production correlates. EuroSpeech99 3:1279–1282
16. Levine S, Theobalt C, Koltun V (2009) Real-time prosody-driven synthesis of body language. In proceedings of ACM SIGGRAPH Asia
17. Munhall KG, Jones JA, Callan DE, Kuratate T, Bateson EV (2004) Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol Sci 15(2):133–137
18. Pandzic IS, Forchheimer R Editors (2002) MPEG-4 Facial Animation—The Standard, Implementation and Applications, John Wiley & Sons Ltd, ISBN 0-470-84465-5
19. Pelachaud C, Badler N, Steedman M (1996) Generating facial expressions for speech. Cogn Sci 20 (1):1–46
20. Salvi G, Beskow J, Al Moubayed S, Granström B (2009) SynFace—speech-driven facial animation for virtual speech-reading support. EURASIP Journal on Audio, Speech, and Music Processing, vol. 2009
21. Sargin ME, Erzin E, Yemez Y, Tekalp AM, Erdem AT, Erdem C, Ozkan M (2007) Prosody-Driven Head-Gesture Animation, ICASSP'07
22. SPTK, The Speech Signal Processing Toolkit, http://sp-tk.sourceforge.net/
23. Visage Technologies, http://www.visagetechnologies.com/
24. Yehia H, Kuratate T, Vatikiotis-Bateson E (2000) Facial animation and head motion driven by speech acoustics, 5th Seminar on Speech Production: Models and Data. Hoole P, (ed) Kloster Seeon
25. Zoric G (2005) Automatic Lip Synchronization by Speech Signal Analysis, Master Thesis (03-Ac-17/ 2002-z) on Faculty of Electrical Engineering and Computing, University of Zagreb
26. Zoric G, Smid K, Pandzic I (2007) Facial gestures: Taxonomy and application of nonverbal, nonemotional facial displays for emodied conversational agents. In Toyoaki Nishida (ed) Conversational Informatics—An Engineering Approach. John Wiley & Sons, pp. 161–182, ISBN 978-0-470-02699-1
27. Zoric G, Smid K, Pandzic IS (2009) Towards facial gestures generation by speech signal analysis using HUGE architecture. Multimodal Signals Cogn Algorithmic Issues Lect Notes Comput Sci LNCS 5398:112–120

Goranka Zoric is a Ph.D. student and research and teaching assistant at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia since 2002. She is involved in several undergraduate courses in the fields of virtual environments and multimedia communications. Her current research interests are in the field of face animation of virtual characters, with particular focus on the speech driven animation, and their applications in networked and mobile environments. Further interests include multimedia communication and services using interdisciplinary human centric approach.She has published 20 papers, out of which one book chapter and 8 journal papers. Goranka received her B.Sc. (Dipl.-Ing.) and M.Sc. degrees in electrical engineering with a major in telecommunications and information science from the University of Zagreb in 2002 and 2005, respectively.

In 2005 she worked as visiting scientist at the department of Signal Theory and Communications (TSC) at the Technical University of Catalonia, Barcelona, Spain. Two months research fellowship provided by

European project SIMILAR. During the summer 2006 she had participated in eNTERFACE'06 workshop on Multimodal Interfaces on the project "An Agent Based Multicultural Customer Service Application". This project is a part of NG-ECA project, a collaborative research work of University of Zagreb, University of Tokyo and University of Kyoto. During 2008/2009 she stayed for 15 months as 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.

Robert Forchheimer received the M.S. degree in electrical engineering from the Royal Institute of Technology, Stockholm (RIT) in 1972 and the Ph.D degree from Linköping University in 1979. During the academic year 1979 to 1980 he was a visiting research scientist at University of Southern California. He has further worked at The Royal Institute of Technology (RIT) in Stockholm and at the University of Hannover, Germany.

Dr Forchheimer's research areas have involved data security, packet radio transmission, optical computing, integrated circuits for image processing, image coding, organic electronics and bioinformatics. He has authored or coauthered more than 160 papers in these areas and also holds several patents. He is the cofounder of five companies. Dr Forchheimer is currently in charge of the Information Coding Group at Linköping University. There, his main work concerns algorithms and systems for networked real-time communication, organic electronics and bioinformatics.

Igor S. Pandzic is an Associate Professor at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia and director of Human-Oriented Technologies

Laboratory (HOTLab). He teaches undergraduate and postgraduate courses in the fields of virtual environments and communications. His main research interests are in the field of computer graphics and virtual environments, with particular focus on character animation, embodied conversational agents, and their applications in networked and mobile environments. Igor also worked on networked collaborative virtual environments, computer generated film production and parallel computing. He published four books and around 90 papers on these topics.

Formerly he worked as a Senior Assistant at MIRALab, University of Geneva, Switzerland, where he obtained his PhD in 1998. The same year he worked as visiting scientist at AT&T Labs, USA. In 2001–2002 he was a visiting scientist in the Image Coding Group at the University of Linköping, Sweden, and in 2005 at the Department of Intelligence Science and Technology, Kyoto University, on a Fellowship awarded by Japan Society for Promotion of Science. He received his BSc degree in Electrical Engineering from the University of Zagreb in 1993, and MSc degrees from the Swiss Federal Institute of Technology (EPFL) and the University of Geneva in 1994 and 1995, respectively.

Igor was one of the key contributors to the Facial Animation specification in the MPEG-4 International Standard for which he received an ISO Certificate of Appreciation in 2000. He held various functions in organizing numerous international conferences and served as guest editor for a special topic in IEEE Communications Magazine. He has been active in international research collaborations within the EU research programmes since the 3rd Framework Programme, as well as in national research projects and collaboration with industry.