Towards Facial Gestures Generation by Speech Signal Analysis Using HUGE Architecture

Goranka Zoric[1], Karlo Smid[2], and Igor S. Pandzic[1]

[1] Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10 000 Zagreb
{Igor.Pandzic,Goranka.Zoric}@fer.hr
[2] Ericsson Nikola Tesla, Krapinska 45, p.p. 93, HR-10 002 Zagreb
karlo.smid@ericsson.com

Abstract. In our current work we concentrate on finding correlation between speech signal and occurrence of facial gestures. Motivation behind this work is computer-generated human correspondent, ECA. In order to have a believable human representative it is important for an ECA to implement facial gestures in addition to verbal and emotional displays. Information needed for generation of facial gestures is extracted from speech prosody by analyzing natural speech in real-time. This work is based on the previously developed HUGE architecture for statistically-based facial gesturing and extends our previous work on automatic real-time lip sync.

Keywords: speech prosody, facial gestures, ECA, facial animation.

1 Introduction

Embodied conversational agent (ECA) assumes a human in a computational environment. In human-computer interaction it represents a multimodal interface where modalities are the natural modalities of human conversation: speech, facial displays, hand gestures, body stance [1]. When building believable ECAs, the rules of human behavior must be taken into account.

In face-to-face conversation among humans, both verbal and nonverbal communication takes part. As stated in [2] considerable research is being done on the topic of facial displays with focus on the relationship between facial displays and emotional states. But according to experiments in [3] only one third of nearly 6000 facial displays of psychiatric patients were classified as emotional. In our work we are interested in those facial displays that are not explicit emotional displays (i.e. an expression such as smile). Also, we are not interested in those facial displays that are explicit verbal displays (i.e. visemes) or have an explicit verbal message (i.e. an observation about the size of a massive object may be accompanied by widening of the eyes). In our work we concentrate on the facial gestures. A facial gesture is a form of non-verbal communication made with the face or head, used continuously instead of or in combination with verbal communication. There are a number of different facial gestures that humans use in everyday life, such as different head and eyebrow movements, blinking, eye gaze, frowning etc [4].

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 112–120, 2009.

In everyday communication humans use facial gestures consciously or unconsciously to regulate flow of speech, accentuate words or segments, or punctuate speech pauses [5]. Work in [6] differentiates several roles of facial gestures in this context:

- Conversational signals – correspond to facial gestures occurring on accented items clarifying and supporting what is being said; facial gestures in this category are eyebrow movements, rapid head movements, gaze directions and eye blinks.
- Punctuators – correspond to facial gestures that support pauses by grouping or separating sequences of words; the examples are specific head motions, blinks or eyebrow actions.
- Manipulators – correspond to the biological needs of a face, such as blinking to wet the eyes or random head nods and have nothing to do with the linguistic utterance
- Regulators - control the flow of conversation. A speaker breaks or looks for an eye contact with a listener. He turns his head towards or away from a listener during a conversation.

ECAs that we deal with in this work act only as presenters and are not involved in conversation so only the first three roles are applicable for them.

Further on, we want to animate our ECAs using only natural speech as input. Speech related facial gestures are connected with prosody and paralanguage. Prosody refers to characteristics of speech which cannot be extracted from the characteristics of phoneme segments, where pauses in speech are also included. Its acoustical correlates are pitch, intensity (amplitude), syllable length, spectral slope and the formant frequencies of speech sounds. Paralanguage refers to the non-verbal elements of communication used to modify meaning and convey emotion and includes pitch, volume and, in some cases, intonation of speech.

In this paper our statistically-based method for automatic facial gesturing in real-time is described. Current version of our system includes the following features:

- head and eyebrow movements and blinking during speech pauses
- eye blinking as manipulators
- amplitude of facial gestures dependent on speech intensity

The rest of the paper is organized as follows. After a description of related work in chapter 2, we describe our system and give discussion on obtained results in chapter 3. At the end, future work and further directions are given.

2 Related Work

In this chapter we summarize work that is related to ours. The first part is more connected to psychological and paralinguistic research relevant to synthesizing natural behavior of ECA with accent on facial gestures which are not directly connected with emotional and verbal signals. In the second part we give an overview of existing systems based on non-verbal speech-related full facial animation where we are specifically interested in systems that are driven by the natural speech. We will mention only

several relevant works that give information which helps us correlate characteristics of speech with facial gestures.

Ekman investigated systematically the relation of speech and eyebrow movement in [7]. According to his findings, eyebrow movements occur during word searching pauses, as punctuators or to emphasize certain words or parts of the sentence. Chovil in [2] concentrated on the role of facial displays in conversation. His research results showed that syntactic displays (emphasized words, punctuators) are the most frequent facial gestures accompanying speech and among those facial gestures, raising or lowering eyebrows are the most relevant. Cavé et al. in [8] investigated links between rapid eyebrow movement and fundamental frequency changes suggesting that these are not automatically linked but are consequence of linguistic and communicational choices. Honda in [9] connects pitch and head movement and Yehia et al. in [10] linearly map head motion and F0. Granström et al. investigated in [11] contribution of eyebrow movement to the perception of prominence and later added head movements emphasizing the importance of timing [12].

Existing ECA systems mainly use text or natural speech to drive full facial animation. There is considerable literature on the systems based on text input. To mention a few: [13][6][14][15][16][17]. These systems incorporate facial and head movements in addition to lip movements (voice). Many existing systems are capable of automatic lip synchronization from speech signal [18][19][20][21][22] producing rather correct lip movements but missing natural experience of the whole face because the rest of the face has a marble look. To add facial gestures to such systems, we need more information. As previously noted, knowledge needed for correlating facial gestures and features extracted from the speech signal is based on the results of psychological and paralinguistic research. However, state of the art literature lacks method which would be able to automatically generate a complete set of facial gestures, including head movements, by only analyzing speech signal. Existing systems in this field mainly concentrate on a particular gesture (i.e. head movement), or general dynamics of the face.

Works in [23][24][25][26][27][28][29] generate head movements based on recent evidence which shows that an audio feature, pitch contour (F0), is correlated with head motions [9][10]. In [23] preliminary evidence is given for the correlation between head motion and fundamental frequency. They measured and estimated face and head motion data to animate parametric talking heads. An extension of this work is given in [24], where authors used estimated head and face motion to animate a talking head. Later on, in [25], in addition to F0, they computed a root mean square (RMS) amplitude and concluded that both F0 and RMS amplitude were highly correlated with the kinematics of head motion during the speech. Fully automatic system for head motion synthesis is developed in [26], taking pitch, the lowest five formants, MFCC and LPC as audio features. Work in [27] generates expressive facial animation from speech and similarly as in previously mentioned systems adds head motion, while [28] additionally considers speech intensity as an audio feature. Recent work on automatic head motion prediction from speech data [29] is based on the thesis that temporal properties should be taken into account and therefore the data has to be segmented into longer parts.

Some systems use speech features to drive general facial animation. Work in [30] learns the dynamics of real human faces during speech using two-dimensional image

processing techniques. Lip movements and coarticulation is incorporated as well as additional speech-related facial animation. Similarly, the system in [31] learns speech-based orofacial dynamics from video, generating facial animation with realistic dynamics.

While in [32] a method to map audio features (F0, mean power) to video analyzing only eyebrow movements is proposed, Albrecht et al. in [33] introduce a method for automatic generation of several non-verbal 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 intensity of the utterance.

The systems described so far need a preprocessing step. Real-time speech driven facial animation is addressed in [34]. Speech energy is calculated and used as a variable parameter to control the facial modifications such as eyebrows frowning or forehead wrinkling. In our work we include a wider set of speech-driven facial gestures generated in real-time with on-the-fly rendering.

3 Speech-Driven Facial Gestures Based on HUGE Architecture

In speech-driven gestures the idea is to find a correlation between speech signal and occurrence of gestures in order to produce speech-driven facial animation. We are taking into consideration speech prosody features since prosody may reflect turn-taking in conversational interactions, types of utterance such as questions and statements, people's attitudes and feelings etc. – content that cannot be said with words, but shown with non-verbal signals. Prosody is derived from the acoustic characteristics of speech including pitch or frequency, length or duration, loudness or intensity, and pause. While relation between speech and lip movements is obvious, the relation between facial gestures (also gestures in general) and speech isn't so strong. Moreover, variations from person to person are bigger. Research on non-verbal communication of the face [35] [36] gave rules for generating facial gestures during thinking or word-search pauses (i.e. avoidance of gaze, eyebrows raise, frowning), also rules for the use of blinking as a manipulator [6] and rules considering gaze in the function of turn-taking [36]. Many of these rules may be applied when generating speech-driven facial gestures.

In the previous work an automatic lip-sync system was implemented [18]. It takes speech signal as input and performs audio to visual mapping in order to produce visemes. Our current work aims to develop an automatic system for full facial animation driven by speech in real-time. We currently distinguish between several facial animation components:

- head and eyebrow movements and blinking as punctuators
- head and eyebrow movements during thinking and word-search pauses
- blinking as a manipulator

Generation of speech-related facial gestures is based on our HUGE architecture.

3.1 HUGE Architecture

HUGE architecture [37] is a universal architecture for statistically-based human gesturing. It is capable of producing and using statistical models for facial gestures based on any kind of inducement (any signal that occurs in parallel to production of gestures in human behavior and may have a statistical correlation with the occurrence of gestures). The system works in two phases: the statistical model generation phase and the runtime phase.

In the statistical model generation phase, the raw training data is annotated and classified into a timed sequence of gestures and timed sequence of inducement states. An inducement state can be any state determined from the inducement that is expected to correlate well with production of gestures. Next, the statistical model is produced by correlating the gesture sequence with the inducement sequence. Facial gesture parameters that are incorporated in the statistical model are gesture type, duration and amplitude value.

The runtime phase runs in real time and is fully automatic. This phase takes a new sequence of inducement data and uses it to trigger the statistical model and to produce real time animation corresponding to the inducement.

A system based on the speech signal is a special case of HUGE architecture. Its components can be seen on Fig. 1.

Adaption of HUGE architecture to speech signal as inducement included the following crucial issues:

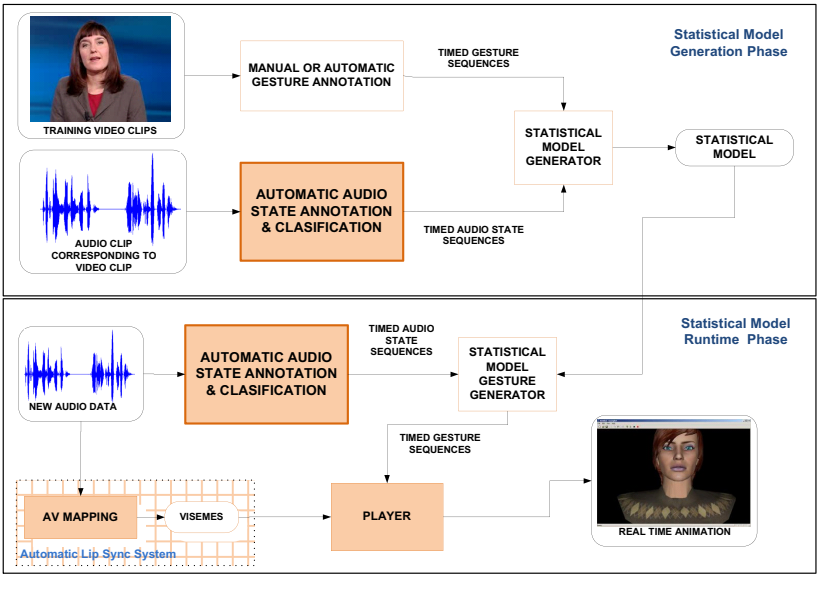

Fig. 1. Universal architecture of HUGE system adapted to audio data as inducement

- Definition of audio states correlated with specific speech signal features
- Implementation of the *Automatic Audio State Annotation & Classification* module, which assumes:
    - Speech signal analysis and feature set extraction
    - Speech signal classification into defined audio states
- Integration of the existing Lip Sync system

As a first step we have decided to use a pause in the speech as a base for the definition of audio states. Such decision is motivated by the existing connection between pauses in the speech and occurrence of certain facial gestures shown in the literature. The input speech signal is divided into silence and speech segments. Silence detection algorithm is based on the RMS amplitude value [38]. Speech segments will be used in further work for calculating speech prosody features such as pitch. Silences identified as pauses in speech are used for definition of two audio states – long and short pause. RMS amplitude is calculated for each frame of the speech signal, which is every 16ms just as in the existing Lip Sync system. Lip Sync system calculates visemes for every frame of audio. The results for four consecutive frames are summed and the viseme with the best score is modeled. We follow this idea when classifying speech into audio states. Using the obtained RMS amplitude value 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, speech signal is classified into the audio state called short pause. When pause in speech is longer than 32 frames, speech signal is classified into the audio state called 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.

3.2 Facial Gesture Generation

Based on the identified audio states, statistical model gives a sequence of facial gestures that might occur on the specific audio state. However, we additionally add certain rules which correspond to the rules that facial gestures have as a communications channel as stated in Introduction. It means that we might add or remove certain gestures depending e.g. on the indentified audio state.

When we consider pause in speech as a punctuator, eyebrow movements, head movements and eye blinks might happen, while during thinking and word-search pauses, eyebrow and head movements are supported. In addition to voluntary eye blinks, we have implemented periodic eye blinks (i.e. manipulators) based on the values given in [6]. If an eye blink does not happen as a punctuator within 5s it will be generated serving as a manipulator.

By using these rules we fine-tune the output from the statistical model. Once we know a gesture type, we know amplitude and duration of the specific gesture since they are also obtained from the statistical model. However, there is a possibility to additionally control amplitude of the facial gesture by calculating speech volume variations.

Having timed gesture sequences and also correct lip movements, we are able to create facial animation (visage|SDK is used). Figure. 2 shows snapshots from a facial animation generated from speech signal and incorporating non-verbal behavior.

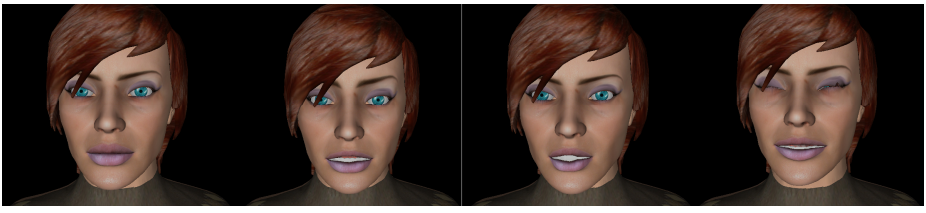

Fig. 2. From left to right: neutral pose, eyebrow movement, head movement, eye blink

4 Future Work

In this paper we have presented a system for creating facial gestures in real-time from a speech signal which incorporates non-verbal behavior. Even by incorporating a small number of prosodic and paralanguage features as well as communication rules, facial animation achieved a lot on naturalness.

However, the system is still in an early stage and there are many things left to do. Next we are planning to add head and eyebrow movements correlated with pitch changes. Adding gaze is an important issue since gaze contributes a lot to naturalness of the face. We will try to integrate as many of the rules found in literature on facial gestures as possible. Fine tuning and evaluation of our system remain an important step in building a believable virtual human.

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.

References

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.): Embodied Conversational Agents, p. 430. MIT press, Cambridge (2000)
2. Chovil, N.: Discourse-oriented facial displays in conversation, Research on Language and Social Interaction (1991)
3. Fridlund, A., Ekman, P., Oster, H.: Facial expressions of emotion. In: Siegman, A., Feldstein, S. (eds.) Nonverbal Behavior and Communication. Lawrence Erlbaum, Hillsdale (1987)
4. Zoric, G., Smid, K., Pandzic, I.: Facial Gestures: Taxonomy and Application of Nonverbal, Nonemotional Facial Displays for Emodied Conversational Agents. In: Nishida, T. (ed.) Conversational Informatics - An Engineering Approach, pp. 161–182. John Wiley & Sons, Chichester (2007)
5. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding, Semiotica (1969)
6. Pelachaud, C., Badler, N., Steedman, M.: Generating Facial Expressions for Speech. Cognitive Science 20(1), 1–46 (1996)
7. Ekman, P.: 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 (1979)

8. Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R.: About the relationship between eyebrow movements and F0 variations. In: Proceedings of Int'l Conf. Spoken Language Processing (1996)
9. Honda, K.: Interactions between vowel articulation and F0 control. In: Fujimura, B.D.J.O., Palek, B. (eds.) Proceedings of Linguistics and Phonetics: Item Order in Language and Speech (LP 1998) (2000)
10. Yehia, H., Kuratate, T., Vatikiotis-Bateson, E.: Facial animation and head motion driven by speech acoustics. In: Hoole, P. (ed.) 5th Seminar on Speech Production: Models and Data, Kloster Seeon (2000)
11. Granström, B., House, D., Lundeberg, M.: Eyebrow movements as a cue to prominence. In: The Third Swedish Symposium on Multimodal Communication (1999)
12. House, D., Beskow, J., Granström, B.: Timing and interaction of visual cues for prominence in audiovisual speech perception. In: Proceedings of Eurospeech 2001 (2001)
13. Graf, H.P., Cosatto, E., Strom, V., Huang, F.J.: Visual Prosody: Facial Movements Accompanying Speech. In: Proceedings of AFGR 2002, pp. 381–386 (2002)
14. Granström, B., House, D.: Audiovisual representation of prosody in expressive speech communication. Speech Communication 46, 473–484 (2005)
15. Cassell, J.: Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems. In: Luperfoy, S. (ed.) Spoken Dialogue Systems. MIT Press, Cambridge (1989)
16. Bui, T.D., Heylen, D., Nijholt, A.: Combination of facial movements on a 3D talking head. In: Proceedings of Computer Graphics International (2004)
17. Smid, K., Pandzic, I.S., Radman, V.: Autonomous Speaker Agent. In: Computer Animation and Social Agents Conference CASA 2004, Geneva, Switzerland (2004)
18. Zoric, G.: Automatic Lip Synchronization by Speech Signal Analysis, Master Thesis (03-Ac-17/2002-z) on Faculty of Electrical Engineering and Computing, University of Zagreb (2005)
19. Kshirsagar, S., Magnenat-Thalmann, N.: Lip synchronization using linear predictive analysis. In: Proceedings of IEE International Conference on Multimedia and Expo., New York (2000)
20. Lewis, J.: Automated Lip-Sync: Background and Techniques. Proceedings of J. Visualization and Computer Animation 2 (1991)
21. Huang, F.J., Chen, T.: Real-time lip-synch face animation driven by human voice. In: IEEE Workshop on Multimedia Signal Processing, Los Angeles, California (December 1998)
22. McAllister, D.F., Rodman, R.D., Bitzer, D.L., Freeman, A.S.: Lip synchronization of speech. In: Proceedings of AVSP 1997 (1997)
23. Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., Yehia, H.: Audio-visual synthesis of talking faces from speech production correlates. In: Proceedings of EuroSpeech 1999 (1999)
24. Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E.: Linking facial animation, head motion and speech acoustics. Journal of Phonetics (2002)
25. Munhall, K.G., Jones, J., Callan, D., Kuratate, T., Vatikiotis-Bateson, E.: Visual Prosody and Speech Intelligibility. Psychological Science 15(2), 133–137 (2003)
26. Deng, Z., Busso, C., Narayanan, S., Neumann, U.: Audio-based Head Motion Synthesis for Avatar-based Telepresence Systems. In: Proc. of ACM SIGMM Workshop on Effective Telepresence (ETP), NY, pp. 24–30 (October 2004)
27. Chuang, E., Bregler, C.: Mood swings: expressive speech animation. ACM Transactions on Graphics (TOG) 24(2), 331–347 (2005)
28. Sargin, M.E., Erzin, E., Yemez, Y., Tekalp, A.M., Erdem, A.T., Erdem, C., Ozkan, M.: Prosody-Driven Head-Gesture Animation. In: ICASSP 2007, Honolulu, USA (2007)

29. Hofer, G., Shimodaira, H.: Automatic Head Motion Prediction from Speech Data. In: Proceedings Interspeech 2007 (2007)
30. Brand, M.: Voice Puppetry. In: Proceedings of Siggraph 1999 (1999)
31. Gutierrez-Osuna, R., Kakumanu, P.K., Esposito, A., Garcia, O.N., Bojorquez, A., Castillo, J.L., Rudomin, I.: Speech-driven facial animation with realistic dynamics. IEEE Transactions on Multimedia (2005)
32. Costa, M., Lavagetto, F., Chen, T.: Visual Prosody Analysis for Realistic Motion Synthesis of 3D Head Models. In: Proceedings of International Conference on Augmented, Virtual Environments and 3D Imaging (2001)
33. Albrecht, I., Haber, J., Seidel, H.: Automatic Generation of Non-Verbal Facial Expressions from Speech. In: Proceedings of Computer Graphics International 2002 (CGI 2002), pp. 283–293 (2002)
34. Malcangi, M., de Tintis, R.: Audio Based Real-Time Speech Animation of Embodied Conversational Agents. LNCS (2004)
35. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated Conversation: Rule-based Generation of Facial Expressions, Jesture & Spoken Intonation for Multiple Conversational Agents. In: Proceedings of SIGGAPH 1994 (1994)
36. Lee, S.P., Badler, J.B., Badler, N.I.: Eyes Alive. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques 2002, San Antonio, Texas, USA, pp. 637–644. ACM Press, New York (2002)
37. Smid, K., Zoric, G., Pandzic, I.P.: [HUGE]: Universal Architecture for Statistically Based HUman GEsturing. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 256–269. Springer, Heidelberg (2006)
38. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech signals. Prentice-Hall Inc., Englewood Cliffs (1978)
39. http://www.visagetechnologies.com/