Published in the Proceedings of International Conference on Auditory-Visual Speech Processing AVSP 2008

Towards Real-time Speech-based Facial Animation Applications built on HUGE architecture

Goranka Zoric, Igor S. Pandzic

Department of Telecommunications, Faculty of Electrical Engineering and Computing
{Goranka.Zoric, Igor.Pandzic}@fer.hr

Abstract

This paper examines how our existing HUGE architecture aids fulfilling requirements emerging from the everyday expansion of new multimedia applications. Trends in human-computer interaction stream towards creating natural user interfaces which apply rules from human-human interaction. Virtual humans who look natural, and behave believably fit perfectly in the concept of creating natural user interfaces.

Universal architecture for statistically based HUman GEsturing (HUGE) is a system for producing and using statistical models for facial gestures based on any kind of inducement. We present the general architecture and its adoption to speech-based facial gesturing, and further justify it through the analysis of requirements for building multimedia application. We believe that this universal architecture is useful for experimenting with various kinds of potential applications since it enables automatic generation of full facial animation in the real time.

Index Terms: facial gesturing, lip syncing, MPEG-4 FBA, virtual characters, speech prosody, multimedia applications

1. Introduction

The human face is an extremely important communication channel. The face can express lots of information, such as emotions, intention or general condition of the person. In noisy environments, lip movements can compensate for a possible loss in speech signal. Moreover, the visual component of speech plays a key role for hearing impaired people.

Besides the communication functions, the human face is primary element in human recognition. Composed of a complex structure of bones and muscles, it is extremely flexible and capable for various movements and face expressions. Such anatomical complexity accompanied by human sensitivity to discontinuities in simulated face movements, makes face animation one of the most difficult and challenging research areas in the computer animation.

Virtual humans are graphical simulations of real or imaginary persons capable of human-like behaviour, most importantly talking and gesturing [1]. When integrated into an application, a virtual human representing a real human, brings life and personality, improves realism and in general provides a more natural interface. The rules of human behaviour, among others, imply speech and facial displays - in face-to-face conversation among humans, both verbal and nonverbal communication takes part. For a realistic result, lip movements must be perfectly synchronized with the audio. Other than lip sync, realistic face animation includes facial displays. In our work we are interested in those facial displays that are not explicit emotional displays (i.e. expression such as smile) and also those which are not explicit verbal displays. We call them facial gestures. Some examples are different head and eyebrow movements, blinking, eye gaze, frowning etc [2].

Systems based on virtual humans generally give rich visual output and therefore are suitable for the various multimedia applications such as teleconferencing, messaging, news delivery, advertising, games and in advanced user interfaces (i.e. for education and commerce). Although such systems already exist, most often they do not work in real time, what might be essential for some applications.

The rest of the paper is organized as follows. In chapter 2, we first give related work and then describe our system. Chapter 3 gives discussion on multimedia applications requirements and how HUGE architecture aids fulfilling them. At the end, future work and further directions are given.

2. System overview

Virtual characters that we deal with in this work act only as presenters and are not involved in conversation. Further, we animate our virtual characters using only natural speech as input in real time. Examples of several existing systems for speech driven facial gesturing are given in the next paragraph.

Works in [3][4][5][6] generate head movements from pitch (F0). In [3] 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 head. Fully automatic system for head motion synthesis is developed in [4] taking the pitch, the lowest five formants, MFCC and LPC as audio features. Work in [5] generates expressive facial animation from speech and similarly as in previously mentioned systems adds head motion, while [6] additionally consider speech intensity as audio feature. Some systems use speech features to drive general facial animation. Work in [7] learns dynamics of real human faces during speech using two-dimensional image processing techniques. Similarly, the system in [8] learns speech-based orofacial dynamics from video generating facial animation with realistic dynamics. While in [9] a method to map audio features (F0, mean power) to video analyzing only eyebrow movements is proposed, Albrecht et al. in [10] introduce a method for automatic generation of several non-verbal facial expressions from speech: head and eyebrow raising and lowering, gaze direction, movement of eyelids, random eye movement during normal speech. The intensity of facial expressions is additionally controlled by the intensity of the utterance. The systems described so far need a preprocessing step. Real time speech driven facial animation is addressed in [11]. 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 wider set of speech-driven facial gestures generated in the real time with on-the-fly rendering.

2.1. HUGE architecture

Generation of speech-related facial gestures is based on our HUGE architecture [12]. It 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 kind of signal that occurs in parallel to the production of gestures in human behaviour and that 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 (Figure 1) the raw training data is annotated and classified into the 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.

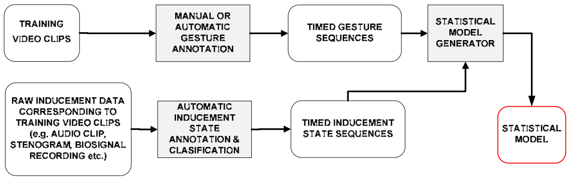
TRAINING VIDEO CLIPS
MANUAL OR AUTOMATIC GESTURE ANNOTATION
TIMED GESTURE SEQUENCES
STATISTICAL MODEL GENERATOR
RAW INDUCEMENT DATA CORRESPONDING TO TRAINING VIDEO CLIPS (e.g. AUDIO CLIP, STENOGRAM, BIOSIGNAL RECORDING etc.)
AUTOMATIC INDUCEMENT STATE ANNOTATION & CLASIFICATION
TIMED INDUCEMENT STATE SEQUENCES
STATISTICAL MODEL

Figure 1: Statistical model generation phase

The runtime phase runs in real time and is fully automatic (Figure 2). 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.

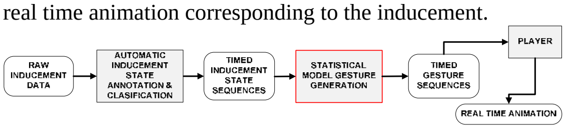
RAW INDUCEMENT DATA
AUTOMATIC INDUCEMENT STATE ANNOTATION & CLASIFICATION
TIMED INDUCEMENT STATE SEQUENCES
STATISTICAL MODEL GESTURE GENERATION
TIMED GESTURE SEQUENCES
PLAYER
REAL TIME ANIMATION

Figure 2: Statistical model run-time phase

2.2. Speech-based HUGE architecture

As we want to use audio as inducement we adopted the HUGE architecture to include acoustic speech features. What we needed to do first is an integration of the existing Lip Sync system [13] since we need mouth movements in addition to facial gestures. It takes speech signal as input and performs audio to visual mapping in order to produce the viseme.

Next we have defined audio states correlated with specific speech signal features. In this part the key component was *Automatic Audio (Inducement) State Annotation & Classification* module whose implementation assumed speech signal analysis and feature set extraction as well as its classification into defined audio states. Knowledge needed for correlating facial gestures and features extracted from the speech signal, is based on the results of psychological and paralinguistic research. Acoustical correlates of prosody and paralanguage features are pitch, intensity (amplitude), syllable length, spectral slope and the formant frequencies of speech sounds.

Current version of our system divides input audio into voiced and unvoiced segments. Voiced segments will be used in further work for calculating speech prosody features such as pitch, and unvoiced segments we use as an audio state determining the speech pause in speech signal. We base a pause detection algorithm on the amplitude and zero crossing rate calculated for each frame of the speech signal [14]. Audio state is determined based on the obtained values and experimentally set thresholds.

Once we have generated statistical model based on the existing audio states, we use it in the runtime to determine facial gestures from the new audio data (Figure 3). At the moment our system supports eyebrow movements, head movements and eye blinks. Once we know a gesture type, we know amplitude and duration of the specific gesture since they are also obtained from the statistical model. Having timed gesture sequences and also correct lip movements, we are able to create facial animation (visage|SDK is used [15]).

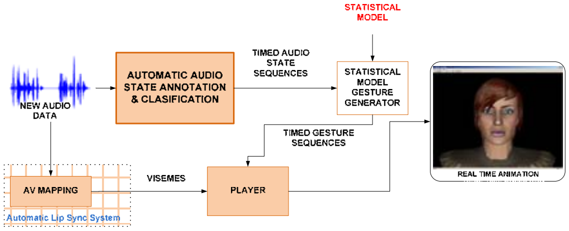
STATISTICAL MODEL
NEW AUDIO DATA
AUTOMATIC AUDIO STATE ANNOTATION & CLASIFICATION
TIMED AUDIO STATE SEQUENCES
STATISTICAL MODEL GESTURE GENERATOR
TIMED GESTURE SEQUENCES
AV MAPPING
Automatic Lip Sync System
VISEMES
PLAYER
REAL TIME ANIMATION

Figure 3: Statistical model runtime phase with audio data as inducement

Existing Lip System takes frames of audio every 16ms and calculates the corresponding viseme. The results for four consecutive frames are summed and the viseme with the best score is modelled. We do similarly with the facial gesture generation. Doing so, we manage to achieve real-time goal. The total time delay consists of the time needed to perform calculations and the time length of the frames taken into account. As we have 16 kHz sampled frames of 256 samples, the time needed for playing one frame is 16 ms. Consequently, a calculation time must be less than 16 ms. Since we analyse 4 frames before deciding about correct viseme and audio state additional time delay is 64 ms. That makes a total time delay less than 80 ms, what is short enough not to lose the impression of real time.

Figure 4 shows snapshots from a facial animation generated from speech signal and incorporating non-verbal behavior. Next figures show snapshots of facial gestures generated from speech signal.

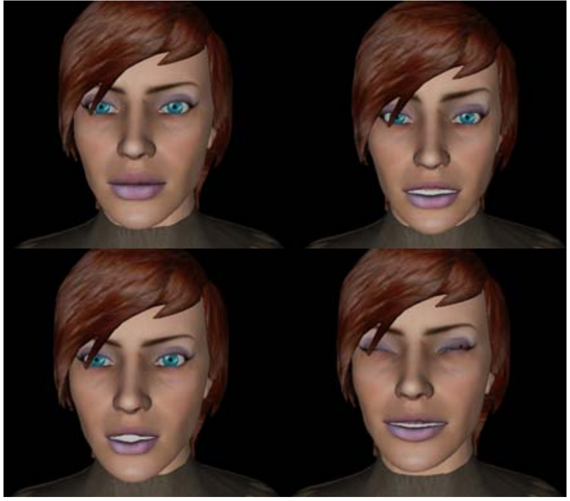

Figure 4: From left to right, up to down: neutral pose, eyebrow movement, head movement, eye blink

3. HUGE Architecture and Multimedia Applications

Typically, a face animation system consist of three main parts [16] which are face model production, face animation production, and specific platform delivery.

Making a face model and preparing it for animation is typically time consuming, but it need to be done only once for specific usage and is usually done by a professional designer. Face animation is in our case generated automatically using the HUGE architecture. Having a face model and automatic facial gesturing it becomes simple to experiment with the use of such a system with various applications and platforms. Our facial animation produces standard MPEG-4 FBA bitstreams, with bit rates that can be as low as 0.5 kbit/sec if only viseme-based speech animation is used. The delivery is based on the very small and portable Face Animation Player core which can easily be ported on top of any software environment supporting 3D graphics. The player is essentially an MPEG-4 FBA decoder. When the MPEG-4 Face Animation Parameters (FAPs) are decoded, the player applies them to a face model. Due to its simplicity and low requirements, the face animation player is easy to implement on a variety of platforms (e.g. 3D animation tools, PCs, games consoles, Web or mobile phones).

What follows is description of several facial animation application scenarios. Although variety of similar approaches exists, here are given only the most representative ones, considering the way of use and creation.

3.1. Virtual character as a web guide

A talking virtual character can be integrated into a Web site to provide services as a virtual guide. Appearing on the site, virtual character presents Web site’s information in an attractive way by walking visitors through the content. The virtual guide presents a service to the visitors giving brief explanations. A user can navigate the site and be talked to or can determine required information from the virtual guide using interactive maps or get help while filling in a form.

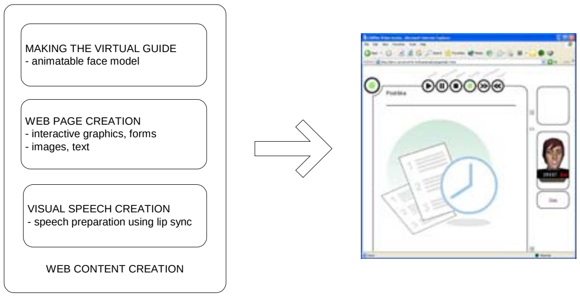

Figure 5: A virtual human as a web guide

There are three steps in creating Web content needed for virtual guide, as shown on the Figure 5.

- Making the virtual guide
- Web page creation
- Visual speech synthesis

The virtual guide’s face is an animatable face model which can be driven by Facial Animation Player implemented as Java applet [17]. Typically, the guide needs to be prepared only once for specific service. The second step is preparing the actual content. After the information to be presented is determined, the presentation form has to be chosen – plain descriptions, interactive graphics or forms. Depending on the chosen method, interactive graphics, forms or images must be prepared, as well as suitable speech for the virtual guide. No matter if the virtual guide is only welcoming visitors or acting in some other way, a speech for each item must be prepared. Suitable speaker is recorded reading the desired text and these recordings are used as input in the facial animation system in order to produce animation.

Since the virtual guide uses natural voice and interacts with users, Web sites become even more pleasant and interesting, improving realism in human-machine communication.

For the end-user delivery, only standard Web browser is needed. Bandwidth and CPU requirements are modest which means that Web site containing the virtual guide is accessible to practically anyone who can access the Web.

3.2. Virtual presentation

Many people gather on the big fairs or shows. Presentations advertising some products or speeches on various themes are being held. In all those situations virtual characters might replace humans making them more attractive to the broad audience. Important is to create adequate talking virtual character for the assigned role which audience will like. In doing so, the things that are already said many times, could become again interesting. Additionally, real speaker do not have to stand on the stage in front of the audience, but can sit in comfortable chair somewhere in the background, reading the prepared text and not worrying about appearance (Figure 6).

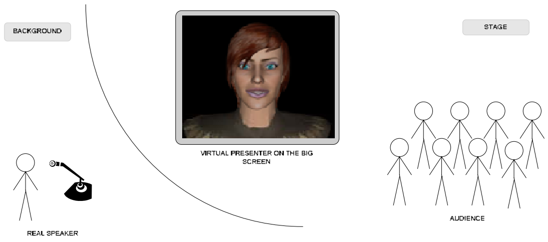

Figure 6: A symbolic view of the virtual presenter use

A virtual character is driven by speech in the real time. The speaker talks in the microphone. At the same time lip sync process is performed and the animation is projected on the big screen. Very few requirements are set for such application. Only interesting virtual character should be created, but this is on the artists and the marketing team. Considering that all is done in the real time, interaction with the audience is also possible.

3.3. Personalized virtual human for mobile devices

With the expansion of various mobile devices such as internet-enabled cellular phones or wireless handheld devices (PDAs), new applications for virtual humans' technologies are opening. On the other hand, for the mobile device users, applications with interfaces using speech are especially important. Consequently, creation of virtual characters’ facial animation based on the speech signal analysis could become notable method in the face animation production for such applications.

Portable computing devices have limited computational and memory resources and strict power consumption constraints. So, various demands are being set on the applications for the mobile devices. Important issues to be considered when designing such applications are: processing power, bandwidth, handoffs, battery power, network instability caused with the nature of wireless medium or mobility of the device, and QoS (Quality of Service). Further challenges arise on the terminal application using 3D faces driven by speech processing algorithms for the automatic facial gesturing, primarily involving speech processing, graphics and animation.

Talking virtual characters, especially if driven by a voice, enable rich multimedia services on mobile platforms and at the same time bring personality and human touch into everyday use of mobile devices, i.e. personalised virtual humans. An idea is to create personalised virtual humans that can be used as part of multimedia services on mobile phones. The personalised 3D face model is created from the 2D facial image and it looks like the person from the picture. It is animated using speech analysis (lip sync). Personalised virtual human is used on the mobile phones as part of the everyday communication via phone calls or MMS messages. In the phone call, personalised virtual human representing caller is animated using caller's voice (speech signal) in the real time. It means that user during conversation is able to see on the mobile phone screen 3D face model looking like the caller and possible also behavioring (gesturing) like him. In the second case, facial animation is generated on the server from the sender’s voiced message and sent as MMS message. By using own voice for creating messages, widespread messaging on the mobile phones is becoming personalized, and the perception more intense. Also, personalised virtual humans might be used in the teleconference or similar multimedia service instead of real human.

4. Conclusions

In this paper we have presented HUGE architecture for creating facial gestures in the real time from the speech signal which incorporates nonverbal behaviour. As well we have described how such architecture helps experimenting with new multimedia applications.

However, the system is still in the early stage and there are still many things which might be improved. Next we are planning to add head and eyebrow movements correlated with pitch. Adding gaze is important issue since gaze contribute a lot to naturalness of the face. Also use of our system in various multimedia applications and evaluation remain as an important step in fine tuning our system for automatic facial gesturing.

5. Acknowledgements

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.

6. References

[1] "Facial Animation Framework for the Web and Mobile Platforms", Igor S. Pandzic, Proceedings of Web3D’02, 2002.

[2] "Facial Gestures: Taxonomy and Application of Nonverbal, Nonemotional Facial Displays for Emodied Conversational Agents", in Conversational Informatics - An Engineering Approach (Toyoaki Nishida, eds.), Goranka Zoric, Karlo Smid and Igor S. Pandzic, 2007.

[3] "Audio-visual synthesis of talking faces from speech production correlates", T. Kuratate, K.G. Munhall, P.E. Rubin, E. Vatikiotis-Bateson and H. Yehia, In Proceedings of EuroSpeech’99, 1999.

[4] "Audio-based Head Motion Synthesis for Avatar-based Telepresence Systems", Z. Deng, C. Busso, S. Narayanan, and U. Neumann, In Proceedings of ACM SIGMM Workshop on Effective Telepresence (ETP), 2004.

[5] "Mood swings: expressive speech animation", Chuang, E., Bregler, C., In ACM Transactions on Graphics (TOG), 2005.

[6] "Prosody-Driven Head-Gesture Animation", M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, M. Ozkan, ICASSP'07 Honolulu, USA., 2007.

[7] "Voice Puppetry"M. Brand, In Proceedings of Siggraph 99, 1999.

[8] "Speech-driven facial animation with realistic dynamics", Gutierrez-Osuna, R., Kakumanu, P.K., Esposito, A., Garcia, O.N. Bojorquez, A., Castillo, J.L., Rudomin, I., IEEE Transactions on Multimedia, 2005.

[9] "Visual Prosody Analysis for Realistic Motion Synthesis of 3D Head Models", Costa, M., Lavagetto, F., Chen, T., In Proceedings of International Conference on Augmented, Virtual Environments and 3D Imaging, 2001.

[10] "Automatic Generation of Non-Verbal Facial Expressions from Speech", Albrecht, I., Haber, J. and Seidel, H., In Proceedings of Computer Graphics International 2002 (CGI 2002), 2002.

[11] "Audio Based Real-Time Speech Animation of Embodied Conversational Agents", Malcangi, M., de Tintis, R., Lecture Notes in Computer Science, 2004.

[12] "[HUGE]: Universal Architecture for Statistically Based HUman GEsturing", Karlo Smid, Goranka Zoric and Igor S. Pandzic, Lecture Notes on Artificial Intelligence LNAI 4133, pp. 256-269 (Proceedings of the 6th International Conference on Intelligent Virtual Agents IVA 2006), 2006.

[13] "Real-Time Language Independent Lip Synchronization Method Using a Genetic Algorithm", Goranka Zoric and Igor S. Pandzic, special issue of Signal Processing Journal on Multimodal Human-Computer Interfaces, 2006.

[14] "Digital Processing of Speech signals", L.R. Rabiner and R.W. Schafer, Prentice-Hall Inc. 1978.

[15] Visage Technologies, www.visagetechnologies.com

[16] "Faces Everywhere: Towards Ubiquitous Production and Delivery of Face Animation", I. S. Pandzic, J. Ahlberg, M. Wzorek, P. Rudol, M. Mosmondor, In Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia, Norrkoping, Sweden, 2003;

[17] "Using a Virtual Human as Web Guide" , G. Zoric, I.S. Pandzic, In Proceedings of the SoftCOM, 2003.