VISUAL PROSODY IN SPEECH-DRIVEN FACIAL ANIMATION: ELICITATION, PREDICTION, AND PERCEPTUAL EVALUATION

A Thesis

by

MARCO ENRIQUE ZAVALA CHMELICKA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2005

Major Subject: Computer Science

VISUAL PROSODY IN SPEECH-DRIVEN FACIAL ANIMATION: ELICITATION, PREDICTION, AND PERCEPTUAL EVALUATION

A Thesis

by

MARCO ENRIQUE ZAVALA CHMELICKA

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

Ricardo Gutierrez-Osuna (Chair of Committee)	Nancy Amato (Member)
Heather Bortfeld (Member)	Valerie Taylor (Head of Department)

May 2005

Major Subject: Computer Science

ABSTRACT

Visual Prosody in Speech-Driven Facial Animation: Elicitation, Prediction, and Perceptual Evaluation. (May 2005)

Marco Enrique Zavala Chmelicka, B.S., Army Polytechnic School

Chair of Advisory Committee: Dr. Ricardo Gutierrez-Osuna

Facial animations capable of articulating accurate movements in synchrony with a speech track have become a subject of much research during the past decade. Most of these efforts have focused on articulation of lip and tongue movements, since these are the primary sources of information in speech reading. However, a wealth of paralinguistic information is implicitly conveyed through visual prosody (e.g., head and eyebrow movements). In contrast with lip/tongue movements, however, for which the articulation rules are fairly well known (i.e., viseme-phoneme mappings, coarticulation), little is known about the generation of visual prosody.

The objective of this thesis is to explore the perceptual contributions of visual prosody in speech-driven facial avatars. Our main hypothesis is that visual prosody driven by acoustics of the speech signal, as opposed to random or no visual prosody, results in more realistic, coherent and convincing facial animations. To test this hypothesis, we have developed an audio-visual system capable of capturing synchronized speech and facial motion from a speaker using infrared illumination and retro-reflective markers. In

order to elicit natural visual prosody, a story-telling experiment was designed in which the actors were shown a short cartoon video, and subsequently asked to narrate the episode. From this audio-visual data, four different facial animations were generated, articulating no visual prosody, Perlin-noise, speech-driven movements, and ground truth movements. Speech-driven movements were driven by acoustic features of the speech signal (e.g., fundamental frequency and energy) using rule-based heuristics and autoregressive models. A pair-wise perceptual evaluation shows that subjects can clearly discriminate among the four visual prosody animations. It also shows that speech-driven movements and Perlin-noise, in that order, approach the performance of veridical motion. The results are quite promising and suggest that speech-driven motion could outperform Perlin-noise if more powerful motion prediction models are used. In addition, our results also show that exaggeration can bias the viewer to perceive a computer generated character to be more realistic motion-wise.

DEDICATION

To God Almighty who has allowed me to live this breath-taking learning experience. To my loving family: my mom Valeria, my dad Enrique, my sister Gabriela, and my brother in law Alex, who have always provided me care, support, and above all the courage to stand still at all times.

ACKNOWLEDGMENTS

The author would like to thank all the persons that have made possible this journey, First of all to my family and friends from the Graduate and Professional Group (GAP) at First Baptist Bryan, in particular to the Joiner's, the Miers', the Hoover's, and the Mitchell's for making me feel at home at their homes.

A special mention to the Fulbright program, Ecuador directed by Mrs. Susana Cabeza de Vaca, the Institute of International Education (IIE) through Paetra Hauck, the Sponsored Student Programs office at Texas A&M, directed by Mrs. Violetta B. Cook, and the Computer Science Department at Texas A&M for providing the economical funds that allowed me to pursue and finally complete this personal goal I had since I obtained my B.S. degree.

This work would not have been possible without the asserted guidance and direction of Dr. Ricardo Gutierrez-Osuna who neglected many lunch hours just to put up with me. In addition, my sincere gratitude to Nancy Amato, who not only served as a Committee Member for my Thesis but also guided me during my freshman year. The author wishes also to acknowledge the advice and support of Dr. Heather Bortfeld and Andruid Kerne; and the prompt help of Dr. Donald Friesen, Elena Rodriguez, and Patricia Rudkin in all the formalities required for graduation.

My special recognition to the contributions of Karl Jablonski in the previous working version of the tracking system used for this project which provided me a fruitful learning experience in C++ and the libraries used to build Graphical-User-Interface applications. In addition, the author would also like to recognize the remarkable work done by the students of the Computer System Design class: Todd Belote, Bryan Harris, Aaron Brown, and Brad Busse, to whom this project owes many things such as the implementation of the real-time queuing and servicing of audio/video events and the new class hierarchy structure of the application just to mention a few.

My gratitude to the members of the Pattern Recognition and Intelligent Sensor Machines (PRISM) Laboratory: Dr. Alex Perera-Lluna, Dr. Takao Yamanaka, Baranidharan Raman, Agustín Gutierrez-Galvez, and Steven Ortiz for their countless hours of advice and suggestions towards the solution of many tasks related to my project despite their packed agenda.

Finally, thanks to my friends Lewis, Miso, Jonna, Steven, James, Tim, Casey, Keri, Lindsay, Dinara, Nishant, Norman, Humberto, Kreshna, Jesse, Jenny, Jason, and many others that shared with me the good times and the best times during my residence in College Station.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Page

Page

Page

1 INTRODUCTION

Facial animations capable of articulating accurate lip motion in synchrony with a speech track have become increasingly available during the past decade [1]-[2]. Visual speech (i.e., lip and tongue motion) is accompanied by a variety of motion, such as eyebrow raises, head shakes and nods, and eye gaze. These movements are the visual counterpart to the prosody of the spoken language (i.e., intonation, rhythm); hence they are commonly referred to as "visual prosody." Visual prosody carries information that is complementary to that provided by the lexical content of the message. In contrast with visual speech, however, for which the articulation rules are fairly well known (i.e., viseme-phoneme mappings, coarticulation), little is known about the generation of visual prosody. It is for this reason that most speech-driven facial animations do not display visual prosody or resort to randomly generated movements.

1.1 Research hypothesis

The work presented in this thesis is preliminary study of the perceptual contributions of visual prosody in animated characters. *Our main hypothesis is that visual prosody driven by acoustics of the speech signal, as opposed to random or no visual prosody, results in more realistic, coherent and convincing facial animations*.

This thesis follows the style and format of *IEEE Sensors Journal*.

1.2 Organization of the manuscript

The remaining sections of this thesis are organized as follows. Section 2 provides an introduction to computer generated facial animation and the different techniques used to capture motion for animation. Section 3 describes the motion capture system build at the Texas A&M University Pattern Recognition and Intelligent Sensor Machines (PRISM) Lab, and the manner in which facial motion is determined to generate computer animation. Section 4 includes a brief description of the protocol designed to elicit visual prosody, as well as the two different computational models that were built to synthesize it. Section 5 describes the perceptual evaluation of the facial animation with different visual prosody conditions, as well as the statistical analysis of results. Finally, Section 6 presents the conclusions of this research, and promising directions for future work.

2 BACKGROUND REVIEW

This section provides a broad introduction to computer generated character animation, with special emphasis on facial parameterization and facial models. We also present an overview of the most commonly used techniques to capture facial motion for animation purposes. The issue of visual prosody is also reviewed in the final subsection.

2.1 Computer facial animation

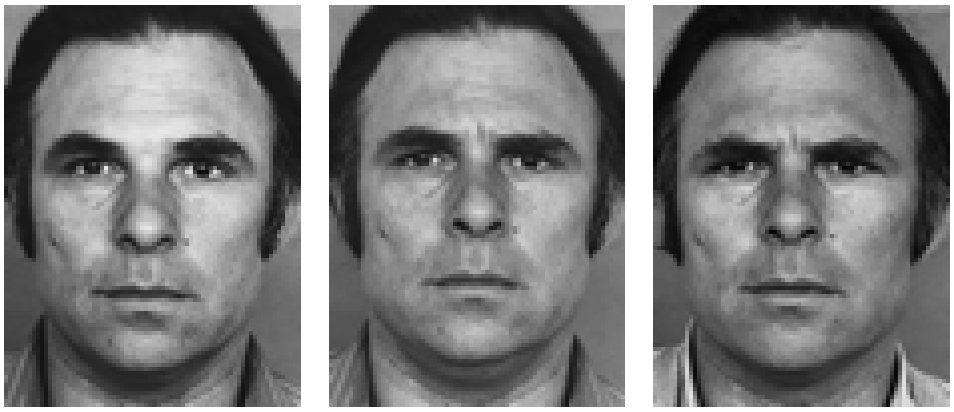
One of the most interesting and challenging areas in computer animation is the synthesis of human faces. Computer facial animation has been an intense subject of study in a variety of scientific disciplines ranging from psychology to computer science, as well as in art. Interest from psychology stems from the acuity with which humans can recognize faces and extract meaning from facial expressions [3]-[6]. Interest within computer science tends to focus on the synthesis of facial avatars for the purpose of multimodal human-computer and computer-mediated interaction. Artists, on the other hand, are interested in aesthetic facets that can be used to convey emotion [7].

As processors and graphic accelerators have increased throughput, it has become easier and more affordable to create computer-animated human characters. Along with these advances, it has also become important to produce realistic images. Perceptual experiments have shown that the more photo-realistic the character appears, the less

forgiving the audience is to details in lip synchronization, saccadic movements, eyebrow motion, head motion and, in general, overall audio-video coherence [8]. Therefore, it has become pressing to study not only the synthesis of faces, but also the rules governing facial feature movement and their relationship with prosody content.

2.1.1 Facial parameterization

The earliest attempt to parameterize facial movements was the Facial Action Coding System (FACS), developed by Ekman and Friesen in 1978 to allow psychologists to study human emotions using facial movements/postures [9]. FACS is based on a detailed study of facial muscle physiology, and the necessary interactions needed in order to produce a visible (noticeable) displacement. In total, they isolated 66 Action Units (AU) that describe a single muscle movement or a group of muscles involved in the movement of a facial feature. For instance, a lowering eyebrow movement (see Fig. 1), is encoded as AU4 (Brow Lowerer). This Action Unit is composed of the union of three muscle strands that affect the forehead, the glabella region (root of nose) and the eyelids. The relevance of this system is that it allowed the description of facial movements in terms of parameters, which in turn nourished the development of computer-based facial animation models [10].



(a) (b) (c)

Fig. 1. Sample pictures borrowed from [9] of a person portraying Action Units (AU): (a) neutral face; (b) AU4 (eyebrow lowerer) sample I; and (c) AU4 sample II.

Based on FACS parameters, Parke [11] developed a computer-generated character in 1974, which has come to be most influential in the computer-animation community. Pearce et al. [12] extended the facial motion parameters of Parke's talking head to support phoneme-based speech animation. The extension provided phoneme-based control by a direct mapping of phonemes into a set of parameters, as well as the timing for each one of them. Additionally, DiPaola [13] extended the parameterization descriptors to include a broader range of facial types and facial expression libraries, support for asymmetric faces, improved eye and ear modeling, and added facial hair and neck parameters. Essa [14] proposed an extension of facial coding called FACS+, which employed computer vision techniques to normalize head photo-shots and extract features based on optical flow. Following a similar approach to the AU coding in FACS, Magnenat-Talman et al. [15] proposed the Abstract Muscle Action (AMA) procedure,

which represents facial expressions by simulating specific face muscle actions that roughly correspond to a muscle or a bone structure. It is important to note, however, that AMA actions are not independent, and the order in which they are executed affects the final result. More recently, in 1999, the Moving Picture Expert Group (MPEG) established a standard for face animation with the MPEG-4 FAP specification [16]. Though controversial, this standard has stimulated the development of commercial applications such as Instant Messaging avatars, MPEG-4 compliant computer animated characters, and MPEG-4 player devices, to mention a few. [17]-[21].

2.1.2 Computer facial models

Broadly speaking, there are three types of facial models: parameterized, muscle-based, and image-based. The first computer-generated faces were modeled by Parke in 1972 [11], and belong to the class of parameterized models. Parke's model consisted of a three-dimensional (3D) mesh of polygons whose movements were limited by physical constraints of the human face. For instance, polygons in the upper lip were adjacent to those in the lower lip, but they were not connected, thus a deformation in the lower area did not affect the upper part. This model aimed to quickly generate a convincing reproduction of a talking head without understanding the physiological events that produced the voice. Several descendants of Parke's talking head have evolved: Sven from the Royal Institute Technology (KTH, Stockholm) [22], Baldi from the Perceptual Science Lab at the University of California Santa Cruz [23], and the Talking Head from

the Laboratory of Computational Engineering at the University of Helsinky [24], to mention a few.

Waters and Terzopoulus [25] designed a facial model using an abstraction of physiological muscle behavior embedded in a non-uniform polygonal 3D structure. They argued that a lineal model could not accurately describe facial movement because muscle contraction and relaxation are inherently non linear. Their 3D physics-based face model with texture-mapped skin used estimates of primary facial muscle contractions as control parameters. In order to generate novel animation the system acquired muscle tension from 2D images in which the subject's facial features (e.g., eyebrows, nasolabial furrow contours, chin) were highlighted with make-up to facilitate tracking, which was performed with snakes [26]. In addition, their model could be customized to a specific subject by texture mapping 3D scanned data from a Cyberware Color 3D digitizer.

An interesting twist was introduced to the visage synthesis field by Bregler et al. [27]. Why not use real human face photographs instead of trying to emulate it using computer graphics (CG) objects and skin textures? Using an audiovisual database of a subject, they were able to generate new footage of the subject producing new utterances not included in the training set. Their technique, called Video-rewrite, automatically mapped training video frames into audio phonemes and produced new sequences by combining these frames according to the desired new audio. Head position and orientation was tracked with computer vision techniques, and mouth images corresponding to new

utterances were stitched in the video using morphing techniques. Their model was able to achieve photorealistic results. In a similar fashion, Ezzat et al. [28] developed a video-realistic speech animation system by means of a Multidimensional Morphable Model (MMM). The only requirement for generation of novel video in this system is to provide annotated and aligned text. Also along these lines, Cosatto [29] developed a photorealistic talking head with texture mapping over a 3D polygonal structure representing the face.

2.2 Computer animation production

Facial animation is the result of concatenating video frames featuring the synthetic actors/objects in different positions. Editing each frame manually can be a very time-consuming task. Therefore, a number of tracking tools, ranging from partially to completely automatic, have been developed to facilitate the generation of motion for synthetic characters. These techniques include key framing, performance-driven animation, and speech-driven animation. Key framing is a process in which several video frames (known as key frames), and the interval between them, are provided as input; the resulting animation is produced by interpolating between key frames. Of more interest to our research are performance-driven and speech-driven animations, which are examined in more detail in the following subsection.

2.2.1 Performance-driven animation

In performance-driven animation, the facial model is driven by data from a motion-capture (mocap) system. These systems can be classified according to the number of cameras used for tracking motion, which range from multiple-camera systems to monocular video. Mocap systems can also be classified based on whether they employ visual markers, or rely on computer vision techniques to extract feature information (marker-less). The former, more intrusive method facilitates the tracking of facial motion by placing visual markers at strategic locations in the subject's face (e.g., eyebrows, lips, chin), whereas marker-less techniques rely on complex computer vision algorithms to extract distinctive facial features and track their positions.

Multiple camera tracking systems generally employ infrared strobe lights to illuminate passive retro-reflective markers, and cameras specially suited to record images in the infrared band. Several companies provide such systems and the required software for image processing and data editing [30]-[31]. These systems have become very popular in the entertainment industry since they allow computer animations to be generated in a very short time. Depending on the number and placement of the cameras, these systems allow a wide range of full-body performances, such as dance and fight scenes, or may be restricted to a confined region. Markers can be obviated at the expense of computationally-intensive processing algorithms such as optical flow [32] and disparity maps [33] to extract the 3D position of facial landmarks, in some cases requiring dedicated supercomputers [34].

The motion-capture system used in this thesis is a semi-custom monocular tracking system with infra-red illumination and retro-reflective markers, an economical alternative to the prohibitive equipment employed in the entertainment industry. The use of a single camera comes at the expense of losing depth information, which prevents us from recovering the 3D position of facial landmarks. This limitation is partly overcome by using a head-mounted frame, which will be discussed in Section 3.2.1.

2.2.2 Speech driven facial animation

In speech-driven animation, facial motion is synthesized from an audio track containing speech utterances. There are two general approaches for audio-driven facial animation: phonetic and subphonetic.

In phonetics-based approaches, a direct mapping from phonemes to visemes (the visual counterpart of a phoneme) is used. This technique requires that a phoneme transcription of the utterance be available by either manual annotation or automated speech recognition. Alternatively, a Text To Speech (TTS) system may be used to synthesize a speech utterance, which implicitly provides the phonetic transcription for the phoneme-viseme mapping. The target configuration of a given viseme in natural speech depends not only on the corresponding phoneme but also on the context (i.e., forward and backward coarticulation). For this purpose, Cohen and Massaro [35] have proposed a

coarticulation model to improve the naturalness of facial animations produced by phoneme-viseme mappings. Their model uses the temporal dependence of visemes by means of so-called “dominance functions” to smooth the transition between viseme targets.

In sub-phonetic approaches, synthesis of facial motion is performed by mapping acoustic parameters (e.g., Linear Predictive Coefficients) directly onto facial motion. Sub-phonetic methods are advantageous because they preserve prosodic information (i.e., intonation, rhythm), which is otherwise lost when an utterance is transcribed into its phonetic sequence. On the other hand, sub-phonemic approaches are computationally intensive since they do not make use of the underlying language structure [36]-[37].

2.3 Visual prosody

Human spoken communication not only uses voice, but also complements it with visual information in parallel. Many of the accompanying gestures filling the visual channel, such as head nodding, eyebrow raises, or pupil dilation, are innate movements that contribute to validate the message content. Cavé et al. [38] has argued that communication is trimodal, requiring the integration of verbal, vocal and gestural channels. The verbal component contains the choice of wording employed in a communicative context, whereas the vocal component embodies the pitch or tone in

which the speech is articulated, and the gestural constituent comprises the use of facial features in a semiotic fashion.

The relationship between message content and visual prosody is complex and not well understood. Dohen et al. [39] have observed that in French there exists a correlation between the word of focus and visually perceptible signals such as jaw opening and lip closure. Granström et al. [40] have used facial gestures to convey affirmative and negative settings in Swedish. They have reported that smile, speech intonation, eyebrow rise, head nodding, and eye closure (in this order) contribute to discriminating the proper setting. In a cross-language study, Krahmer et al. [41] have shown that eyebrow movement accompanies pitch accents. In fact, for Dutch this signal aided in the localization of the word of focus, while in Italian it did not, probably due to prosodic language differences among both languages. Pelachaud et al. [42] have proposed a model for facial expression (eye and head motions) based on discourse semantics that takes into account several dimensions: phonemic, intonational, informational, and affectual.

Recent work in facial dynamics and speech perception ([43] and references therein) has shown that humans are able to correctly identify the source of an utterance. When exposed to a sequence of audio followed by mute video of two speakers (one at a time) performing different utterances, subjects were able to correctly match faces and voices significantly above chance level. These results suggest that information in the speech

channel is coupled across the visual channel, aiding in the proper identification of the speaker.

The contribution of visual prosody to message content is, without a doubt, an area that deserves further study. Breakthroughs in this area will not only increase the naturalness of virtual characters, but will also help understand human communication in a broader sense. In the next subsections the reader will find a detailed review of the different visual prosody channels not involved in speech production, such as head, eyes, and eyebrows.

2.3.1 Head prosody

No facial animation would be complete without the integration of head motion. Head motion is essential in the production of facial animation because not only gives it a sense of vitality, but also contributes to emphasize the message content and characterize the avatar personality [44]. Cosatto [45] noted that low frequency head movements extend to the length of words and phrases and are most probably related to a change of posture, whereas higher frequencies (in the order of 2 to 15Hz) are closely related to prosody content. Deng et al. [46] have developed a model for head motion driven by speech to facilitate animators in the production of novel head motion given new utterances and desired key frames. This model, as acknowledged by the authors, has limitations in its ability to generate head motion if the key frames are not among the information included in the training information, which is stored in the Audio-Head-motion Database (AHD).

Another factor that was not considered in this model is the effect of linguistic context on head motion. Albrecht et al. [47] used a mixed technique to drive head motion with pitch level combined with random tilts introduced from time to time to avoid monotony.

In summary, although head motion is not intrinsically involved in the production of utterances in human spoken language, it provides discernible information that is related not only to voice pitch but also to the semantics of the speech and to the speaker itself.

2.3.2 Pupil prosody (gaze)

Lee et al. [48] has shown that gaze in a communication context serves mainly for: "*1) sending social signals; 2) open a channel to receive information; and 3) regulate the flow of conversation*". The authors developed a statistical model for saccadic eye movement that synthesizes realistic gaze in two modalities: talking mode and listening mode. In a similar approach, Deng et al. [49] generated novel pupil animation using non-parametric sampling techniques from a pool of stored pupil images. Although these models drive gaze autonomously without any input feedback from the environment, the resulting pupil motion looks very realistic.

2.3.3 Eyebrow prosody

Aside from the seminal contribution of Ekman [50], the work of Grammer et al. [51] is one of the earliest studies on eyebrow motion. Through a cross cultural analysis, the

authors showed that there is an innate eyebrow motion, referred to as an eyebrow flash, at the beginning of a human-human interaction that signals invitation or recognition. In addition, Krahmer et al. [41] have noted that eyebrow motion can serve not only as an asynchronous event prior to speech intercourse, but also as a gestural channel conveying complementary information (i.e., the word of focus) to the verbal channel. Eyebrow position, in conjunction with other facial feature postures was used by Ekman and Rosenberg to describe facial expressions that can be interpreted as emotional states [5]. Cosnier [52] also ascribed to eyebrows a role in inquisitive locution. Yet, further investigation is needed to decipher the intricacies of eyebrow function in gestural communication.

3 AUDIO-VISUAL PROCESSING SYSTEM

This section presents the audio-visual capture system that has been developed by us over the past two years for the purpose of tracking facial motion, specifically head, lips, and eyebrows. The section starts with an overview of the imaging hardware based on IR retro-reflective markers, as well as software tools that facilitate the synchronized acquisition of audio and facial-motion. It also describes the detectors that have been developed to extract lip, head and eyebrow motion from raw motion-capture data.

3.1 Audio and motion capture system

An audio-visual system has been developed at the Texas A&M University Pattern Recognition and Intelligent Sensor Machines (PRISM) Lab over the past two years. The system was conceived as a low-cost (under $1,000) alternative to professional motion-capture equipment (i.e., Vicon®), which have a price tag close to $50,000 circa 2004. The PRISM mocap system consists of the following components:

- IBM Blue Eyes pupil camera.
- Acoustic Magic microphone array.
- Winnov Videum 1000 Plus audio/video acquisition card.
- A Personal Computer (PC) (Pentium IV 2GHz, 512MB RAM was used for the experiments described here).

- A Graphical User Interface (GUI) developed with the help of senior-design students at Texas A&M University.
- Retro-reflective adhesive markers, which are placed at key locations on the subject's face, e.g., eyebrows, nose and lips.

The camera and microphone are connected to a custom-off-the-shelf data-acquisition card (Winnov Videum 1000 Plus) capable of capturing hardware-synchronized audio-visual streams at 30 video frames per second (fps) and audio at 16KHz. A GUI developed specifically to manage the capture process allows the user to select the desired procedure (e.g. video capture, video tracking, video playback, etc.). The system is able to record synchronized audio and video, and track facial points in real-time (at ½ video resolution) or off-line (at full resolution) to produce Facial Animation Parameter (FAP) streams that can be read by MPEG-4 compliant readers such as The Facial Animation Engine (FAE) from the University of Geneva [21].

In the next sub-sections the reader will find detailed information regarding necessary modifications that were performed to the camera in order to enhance the tracking of reflective markers, as well as the functional block diagram description of the managing software.

3.1.1 Camera enhancements

The Blue Eyes PupilCam was designed at IBM Almaden Research Center to detect the pupil of a subject using the same principle by which one occasionally gets the annoying "red eyes" with flash pictures [53]. The PupilCam consists of two arrays of infrared Light Emitting Diodes (IR-LEDs): the first array is aligned on-axis (around the camera lens), whereas the second one is aligned off-axis (top hand side and bottom hand side), as shown in Fig. 2. The former is hardware-synchronized to illuminate the pupil area for even video frames ("red eyes") while the latter illuminates during odd frames to ensure that the scene has equal illumination intensity. In this way, the pupils can be detected by a simple image subtraction [54].

The camera was slightly modified to allow tracking of small retro-reflective markers (less than 2x2mm) at full 30fps rates (as opposed to the 15fps rate for pupil tracking) by maintaining the on-axis LED array illuminated at all times and disconnecting the off-axis LED array. In addition, an optical filter (Wratten no.87) was placed on the camera lens to filter out visible light while allowing infrared light to reach the CCD array. Finally, a fine coating of polytetrafluoroethylene (PTFE) was applied to the LEDs to diffuse their light emission and avoid saturation of the CCD array. All of these steps significantly contributed to enhance the contrast between infrared light reflected from the facial markers and the background, producing a clean image for subsequent segmentation.

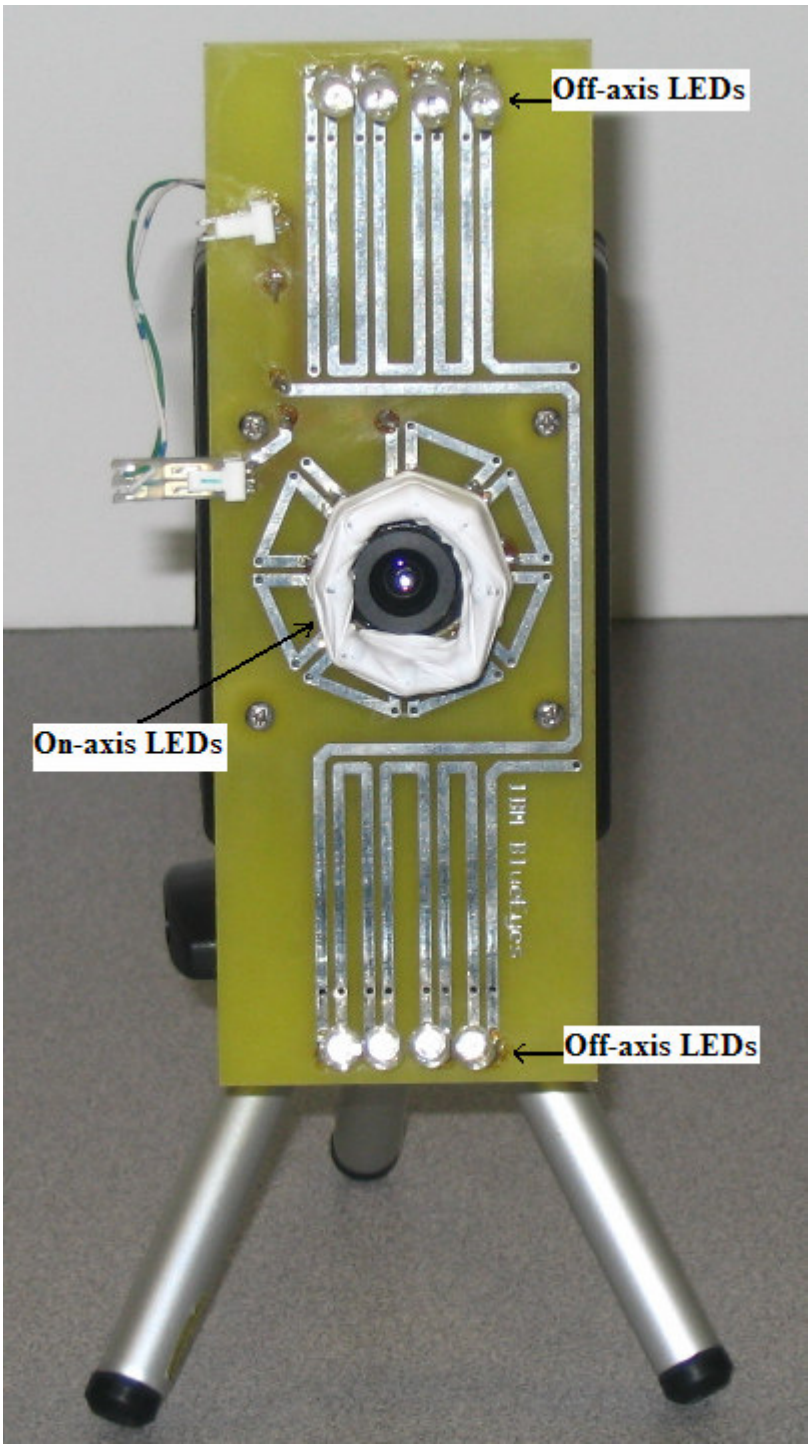


Fig. 2. IBM's Blue Eyes PupilCam system used to acquire motion capture.

3.1.2 Facial motion tracking manager

A custom application has been developed at the Texas A&M University PRISM Lab to manage the acquisition, storage, and post-processing of audio and video. The original system was implemented by Karl Jablonski as part of his Undergraduate Honor's Thesis in 2002-2003 [55]. The system was refined by Todd Belote, Bryan Harris, Aaron Brown, and Brad Busse, as part of their Computer Engineering Capstone Design project in Spring 2004. In addition to co-directing the Capstone Design project, the author's contributions to the development of this software have been:

- Improved memory management for extended video recording and processing (up to 4 minutes of video at 640x480 pixels and 30 fps). Due to the complexity of the application, several orphan memory allocations were created during a typical run-time execution; therefore, limiting the amount of resources available for subsequent processes and consequently the maximum video recording time.
- Addition of video processing capabilities for editing, playback, and analysis of video subsections. This feature, not available in the previous implementation, enabled the user to select a desired subsection from a video or join different videos into a single file. Thus, allowing more flexibility for animation generation.

- Enabling of full-resolution off-line image processing (up to 640x480 pixels with the current camera). The original application was designed for real-time tracking, and could only operate at ½ resolution (320x240).
- Detection of dropped frames and corresponding corrective procedures. Even tough the original application ensured that all the events generated by the capturing card were serviced -by means of a priority queue implementation-, lip synchronization was lost after prolonged recordings (typically three minutes or more). Later, it was determined that the true output from the video card at the maximum resolution (640x480 pixels) was 29.97 fps. Therefore, a copy of a previous frame was inserted in the video file to ensure 30 fps throughput.
- Determination of MPEG-4 compliant facial expression parameters taking into account head motion and appearance using heuristic methods. The previous implementation used a plane transformation matrix based on [56] to project the feature points into the plane position at frame zero prior to compute the displacements. This approach was discarded in the current implementation due to a lack of naturalness in lip motion using an informal perceptual evaluation (MZC and RGO).

The application allows live audio to be saved in a standard wave file format (WAV), whereas video can either be processed on-line (to yield MPEG-4 FAP) or saved in a proprietary format for off-line processing. In addition, the application allows the user to

play back a video sequence in slow motion as well as produce FAP streams with a variety of options. Fig. 3 shows the main screen of the application's GUI.

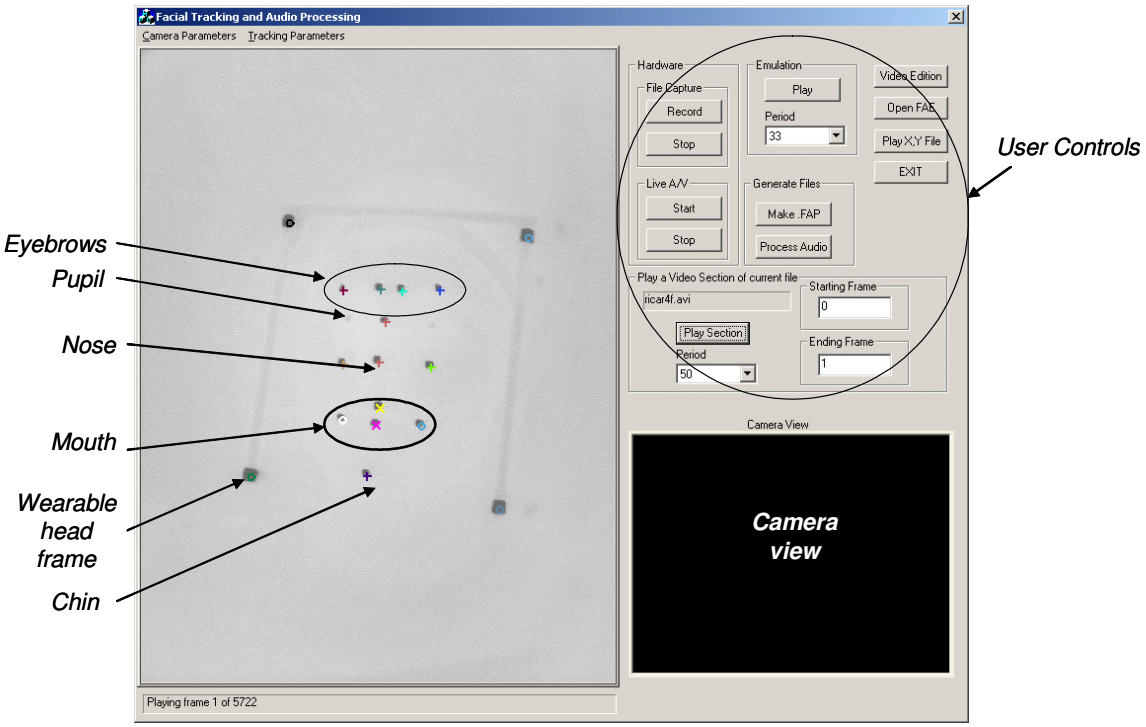


Fig. 3. Graphical user interface for managing motion capture, process audio/video files, and generate new FAP stream files. The left-hand side of the screen shows a processed video frame with color markers overlaid at the location of the recognized landmarks. The right-hand side shows the user controls.

As mentioned earlier, data acquisition is performed with a Winnov Videum 1000 Plus audio-video capture card. During initialization, the facial tracking GUI configures

various parameters of the capture card, including video frame rate, video codec, video resolution, audio sample rate and audio sample size. Once data-acquisition is started, video and audio events are fired periodically to capture new data. The audio/video events are handled directly by the GUI. Audio events are registered every second and are processed by saving the incoming information directly to disk each time they occur.

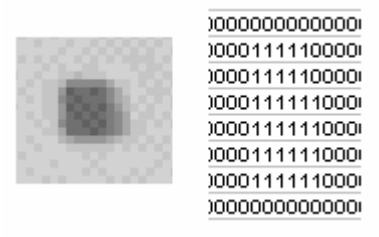
Video events on the other hand, take place every time a frame is available which, at the current frame rate of 30 frames per second, is every 33.3 ms. These events can be handled either by (i) saving the data to memory or (ii) tracking the markers on-line. Each video event is spawned in its own thread and the data is stored in a circular queue until it can be saved to disk. This feature is particularly useful in cases where the thread cannot be serviced in a timely fashion due to operating system tasks in process. Such case is not unlikely considering that a typical hard disk drive access alone takes approximately 15ms to 20ms, which represents almost 60% of the time allotted to process the video data before another frame becomes available.

3.1.3 Video processing

Video processing consists basically of extracting the location of the markers from the raw image and labeling each marker based on its position relative to the other markers in the image. The complete process can be performed on the fly at 30 fps for a maximum image size of 320x240 (one half of the camera resolution) with a Dell Pentium IV 2.0

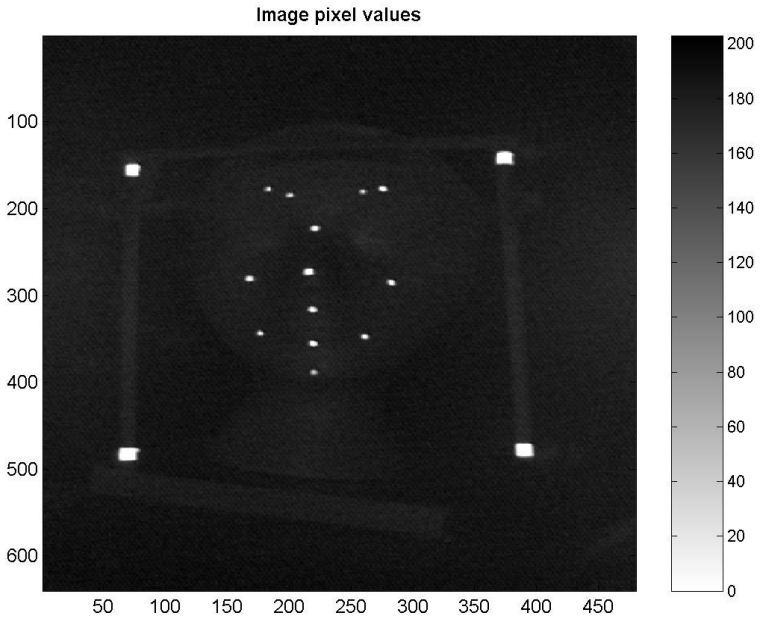
GHZ, 512 MB RAM. Alternatively, the video can be saved at full resolution (640x480 pixels @ 30 fps) and processed off-line, as mentioned earlier.

Marker segmentation. The raw IR image is initially segmented with a pre-specified threshold to produce a binary image. Fig. 4 shows the detail of a reflective marker as seen by the camera before thresholding, and the binary result after applying the threshold criteria (1's denote potential markers, 0's denote background areas). A histogram analysis of a typical image is presented in Fig. 5. It can be seen that reflective marker pixel values have a high contrast when compared to the background.

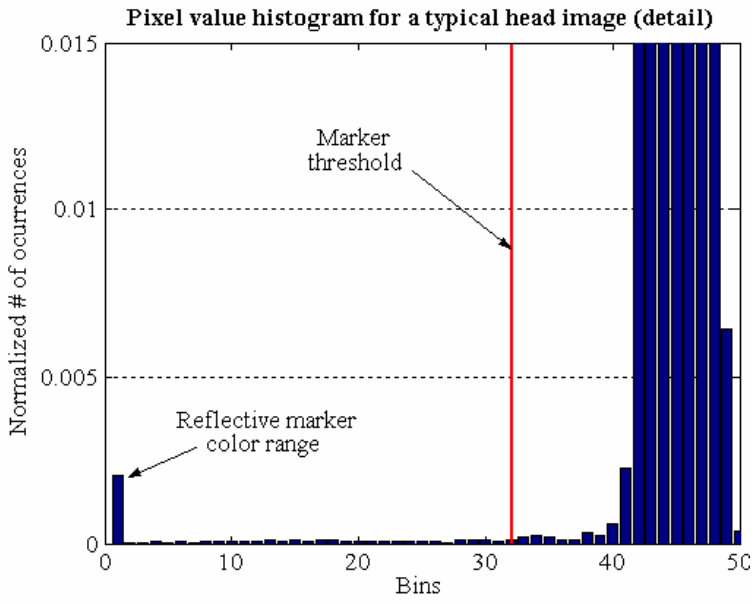


(a) (b)

Fig. 4. Detail of a retro-reflective marker as seen by the camera. (a) Amplified gray scale image of the marker; and (b) corresponding binary map after filter application.



(a)



(b)

Fig. 5. Pixel values from a typical IR head image. (a) Gray scale values; and (b) detail of the normalized histogram for image in (a).

The entire image is scanned to find areas of 3x3 pixels with a density greater than or equal to 0.36, where density is defined as the ratio between the number of pixels with a value of 1 to the total pixel area. When a candidate area is found, its center is located using the following procedure:

a) the searching area is increased from 3x3 pixels to 4x4 and so forth until the measured density falls below 0.36, up to a maximum size of 10x10
b) the center of gravity of the area is determined using equations (1) and (2).

$$x = \frac{\sum_{i=1}^{i=N} \#\, pixelsAboveThresholdInColumn_i \times i}{N} \quad (1)$$

$$y = \frac{\sum_{i=1}^{i=N} \#\, pixelsAboveThresholdInRow_i \times i}{N} \quad (2)$$

This simple procedure proved to be extremely reliable for tracking flat reflective markers as well as semi-spherical shaped markers in different orientations and positions relative to the camera. It is however possible that, due to partial occlusions or odd orientations relative to the image plane, a marker may go undetected. Error recovery from this situation is possible, and is carried by interpolating the last position at which the marker was still present and the first occurrence at which it is reacquired.

Marker labeling. The next step in video processing is labeling each marker based on its extracted position in the image. Such task can be very difficult if the target's initial position is unknown. Therefore, the videotaped subject is first asked to wear a head-mounted light-weight frame and pose with the tip of his/her nose pointing in the direction of the camera, looking at the camera lens, and maintain his/her head straight during the first seconds of recording (in as much as possible). In addition, the subject is also asked to keep his/her facial muscles in a relaxed state, with the lips together, and a closed jaw during this initial period. This neutral head position guarantees a relative location for all target markers. For instance, the frame markers are easily identified by selecting the two uppermost and the two bottommost markers in the image. In the same manner, the eyebrow markers are located immediately below the frame's top markers. Subsequently, cheeks, nose, and lips are identified in a similar fashion. Fig. 6 shows a dummy face and the marker identification result (e.g., lips with blue and cyan diamonds and yellow and magenta x's).

(a) (b)

Fig. 6. Images of a dummy head wearing reflective markers. (a) Taken with a normal digital camera; and (b) as seen by the IBM Pupilcam with a color overlay of the position and identification of the markers (e.g., lips in blue and cyan diamonds and yellow and magenta x's).

The labeling of markers in subsequent frames is facilitated by the spatial locality of the problem. For instance, a mark identified as Upper-Right eyebrow in the first frame will most likely be found in or near the same area in the following frame, thus allowing an

easier marker labeling procedure while relaxing constraints for head orientation in subsequent frames.

After the markers are labeled, their 2D position is stored in a file, and used subsequently to generate the appropriate parameters for the MPEG-4 facial animation engine, as described in the next section. At this point, the video file, which grows at a rate of 1GB every 2 minutes of 640x480 at 30 fps, is no longer needed and can be discarded.

3.2 Facial motion determination

The MPEG-4 standard [57] includes a series of Facial Animation Parameters (FAP) that allow facial expressions to be parameterized. There are a total of 68 FAPs categorized in ten groups shown in TABLE 1. These parameters are defined as relative displacements from a reference face in which the muscles are relaxed, the lips are closed, the upper teeth are in contact with the lower ones, the head is oriented frontally towards the camera, the eyelids are open, and the pupil diameter is 1/3 of the iris diameter (see Fig. 7). This reference face is normally referred to as a neutral face.

TABLE 1

FAP groups adapted from [16], pp 20

Group	Number of FAPs
Visemes ands expressions	2
Jaw, chin, inner lower lip, corner lips, mid lip	16
Eyeballs, pupils, eyelids	12
Eyebrow	8
Cheeks	4
Tongue	5
Head rotation	3
Outer-lip positions	10
Nose	4
Ears	4

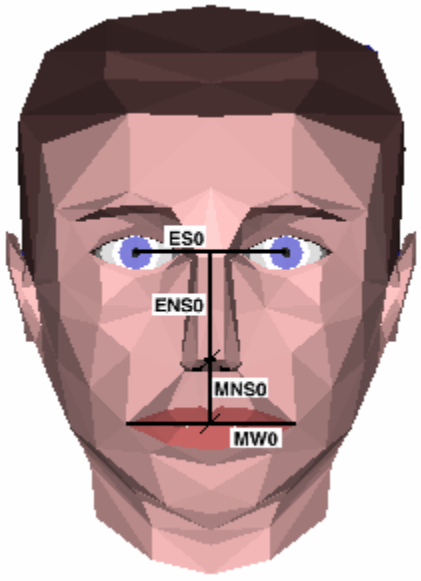


Fig. 7. Neutral face and referential distances used to compute facial animation parameters (adapted from [21]): eye separation (ES0), eye to nose separation (ENS0), mouth to nose separation (MNS0), and mouth width (MW0).

3.2.1 Placement of retro-reflective markers

The placement of markers was optimized to facilitate tracking of facial motion in regions most relevant to our perceptual studies, which are described in Section 4.3. These regions include eyebrows, cheeks, nose, lips, and chin. Eyebrow motion was initially tracked by placing markers at each of the six feature points defined in the MPEG-4 standard (see figure Fig. 8).

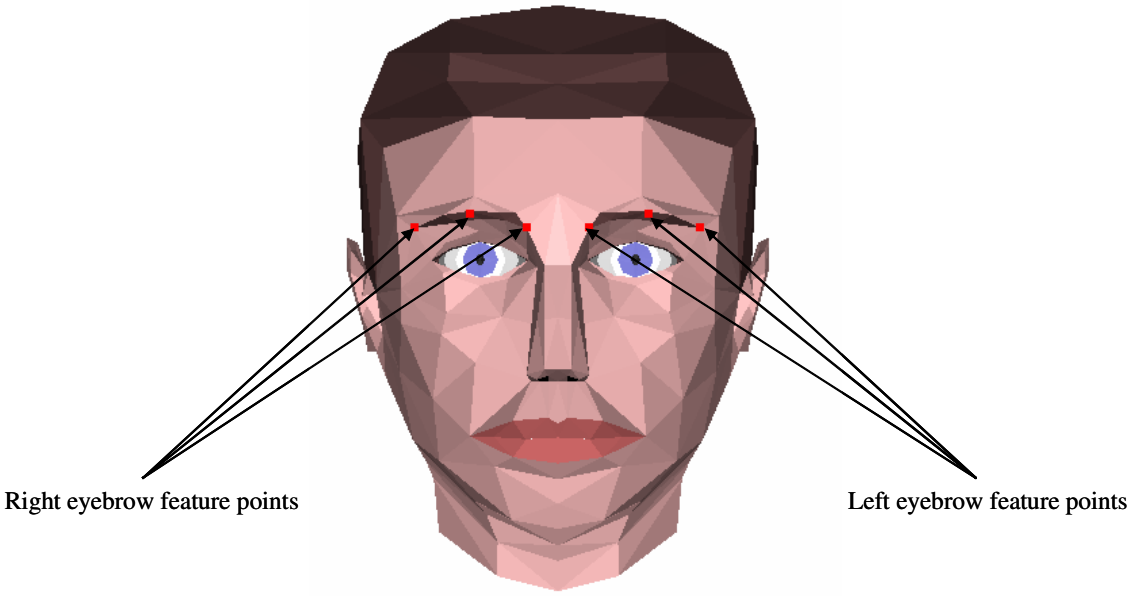
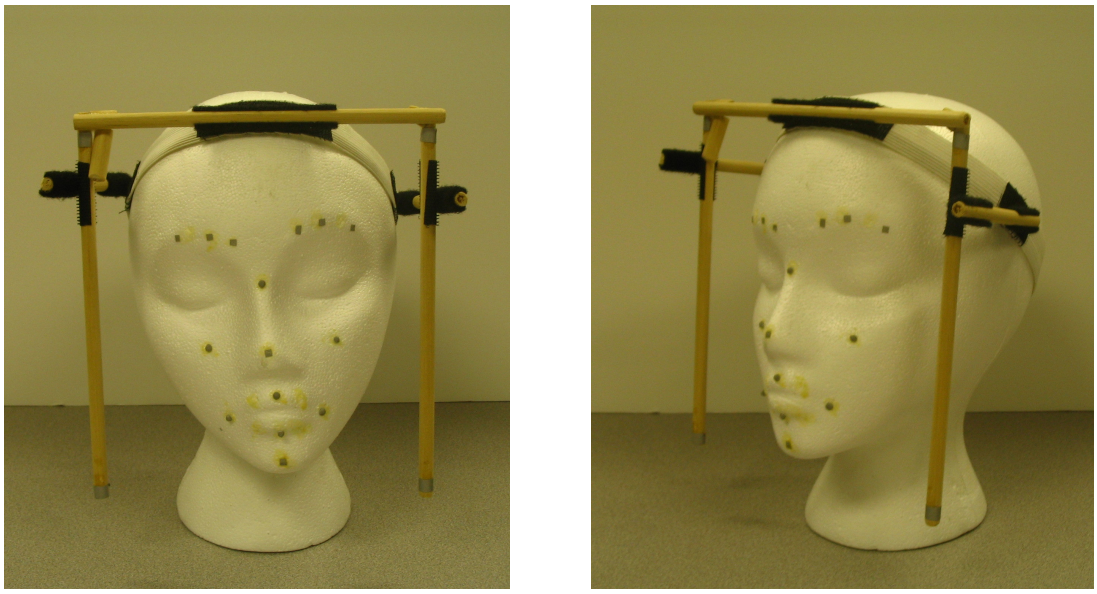


Fig. 8. The six feature points -marked with red squares- defined in the MPEG-4 standard for eyebrow motion.

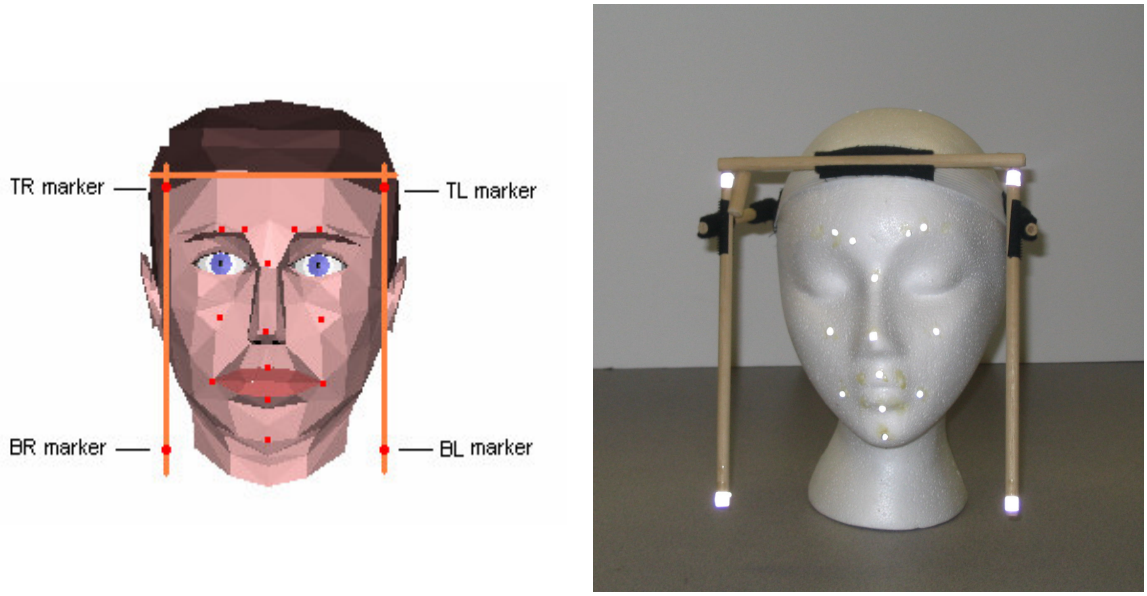
However, it was later found that the outer markers were not easily tracked because, due to their orientation with respect to the camera plane, they would oftentimes disappear during yaw rotations (refer to Fig. 9). For this reason, it was finally decided that only mid and inner eyebrow markers would be used. Similarly, lip motion was initially tracked with eight markers, but this often caused the marker-labeling algorithm to swap labels due to the proximity of the markers. Therefore, it was later decided that only four markers at the extremes of the oral cavity would be used to track lip motion: top, bottom, right, and left.



(a) (b)

Fig. 9. Illustration of marker occlusion caused by head rotation: (a) Styrofoam head in frontal orientation; and (b) with a yaw rotation (note that the right outer eyebrow is almost unnoticeable).

To determine the displacement of a feature relative to the neutral face, it is first necessary to decouple non-rigid motion (e.g., lip motion, facial expressions) from rigid motion (e.g., head rotations). This can be resolved by placing reference markers at locations that are unaffected by non-rigid motion. Initially, the nose tip was used to determine the new head position. However, it was found that the estimates of head pose were not accurate enough to recover subtle facial movements such as eyebrow raises. For this reason, it was decided to assist the process with a head-mounted light-weight frame as shown in Fig. 6(a).



(a) (b)

Fig. 10. Placement of retro-reflective markers and wearable frame on subject's head: (a) on the FAE [21] default model marked in red; and (b) on a styrofoam head (note the flash reflection on the markers).

3.2.2 Estimation of head pose

Fig. 11 shows the three canonical head rotations that can be estimated by the system. Head roll can be estimated directly from the markers on the left and right posts of the wearable frame as depicted in Fig. 12. First, the orientation of the left post in the head-mounted frame is determined from markers TL and BL, and the direction of right post from markers TR and BR. An average of these two orientations (at frame *t*) is computed using equation (3). Later, the angle between the head's vertical orientation $\overrightarrow{VPostAvg_t}$ and the camera vertical direction $\vec{y}$ is determined with a dot product operation. Additionally, a conversion factor is applied since the FAP head rotation units are given in 10^{-5}rad (refer to equation (4)).

$$\overrightarrow{VPostAvg_t} = \frac{\overrightarrow{RightPost_t} + \overrightarrow{LeftPost_t}}{2} \tag{3}$$

$$RollFAPunit_t = 100000 \times a\cos\left(\frac{\overrightarrow{VPostAvg_t} \bullet (\vec{y})}{\left|\overrightarrow{VPostAvg_t}\right|}\right) \tag{4}$$

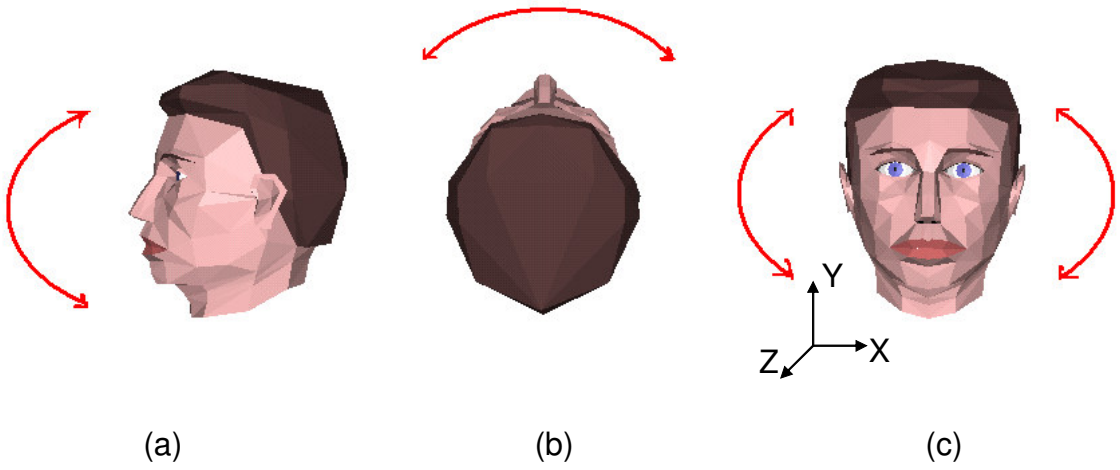


Fig. 11. Canonical head rotations (a) pitch, (b) yaw, and (c) roll.

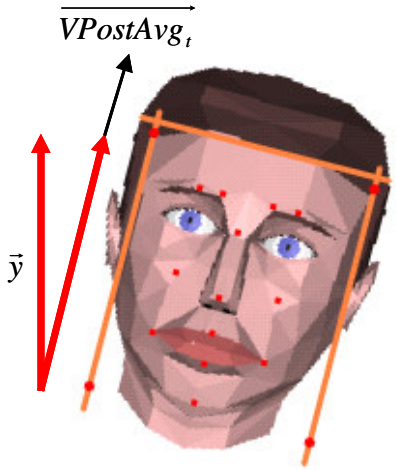


Fig. 12. Head roll is determined by the angle between the vertical ($\vec{y}$) and the average direction between left and right posts ($\overrightarrow{VPostAvg_t}$).

Determination of pitch and yaw angles is not straightforward since a 2D image does not provide depth information. However, an approximate measure of these rotations can be computed by exploiting perspective projection. The approach is illustrated in Fig. 13. Two line segments with the same length, one at a distance d and the other at a distance $d+\Delta d$ from the plane of projection, will have a different apparent length on the image plane. The farther the line segment is from the plane of projection, the smaller it appears. A similar effect is produced in the projection of the reference frame when the head orientation changes due to pitch or yaw rotations. Using this rationale in the reference frame's appearance problem, we find that when the head is leaned forward the distance between TL and TR markers is greater than the distance between BL and BR markers, as shown in Fig. 14(b). Analogously, the distance between TR and BR markers appears larger than the distance between TL and BL markers when the head is turned left (cf. Fig. 14(c)).

Fig. 15 illustrates the effect of pitch rotations on the relative distance between top markers (TL-TR), and bottom markers (BL-BR) in the frame. The left image shows a head in the neutral pose, whereas the right image portraits the same head leaned forward, both images as captured by the camera.

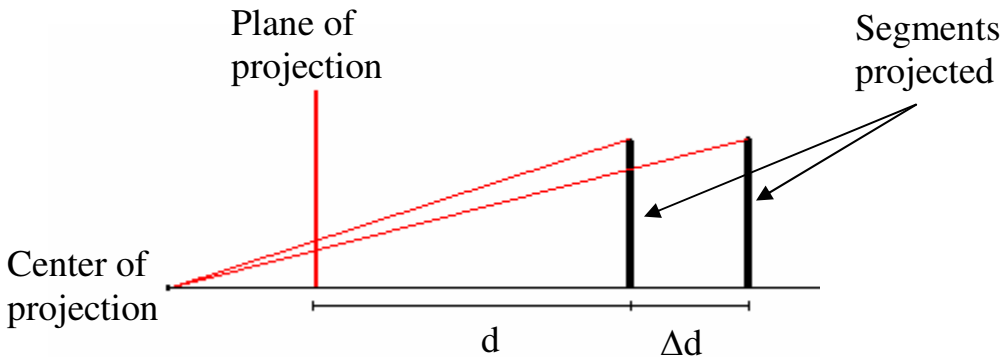
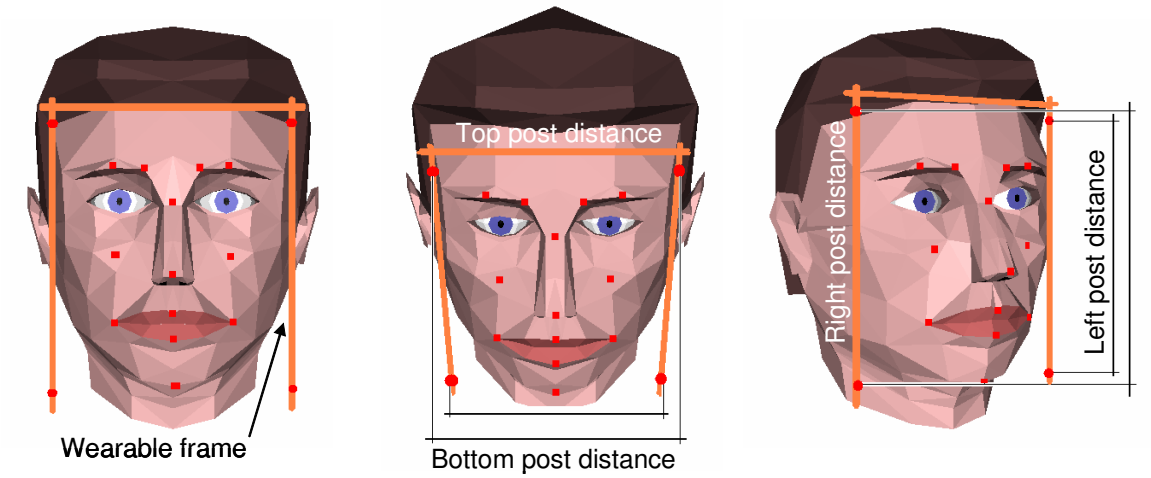
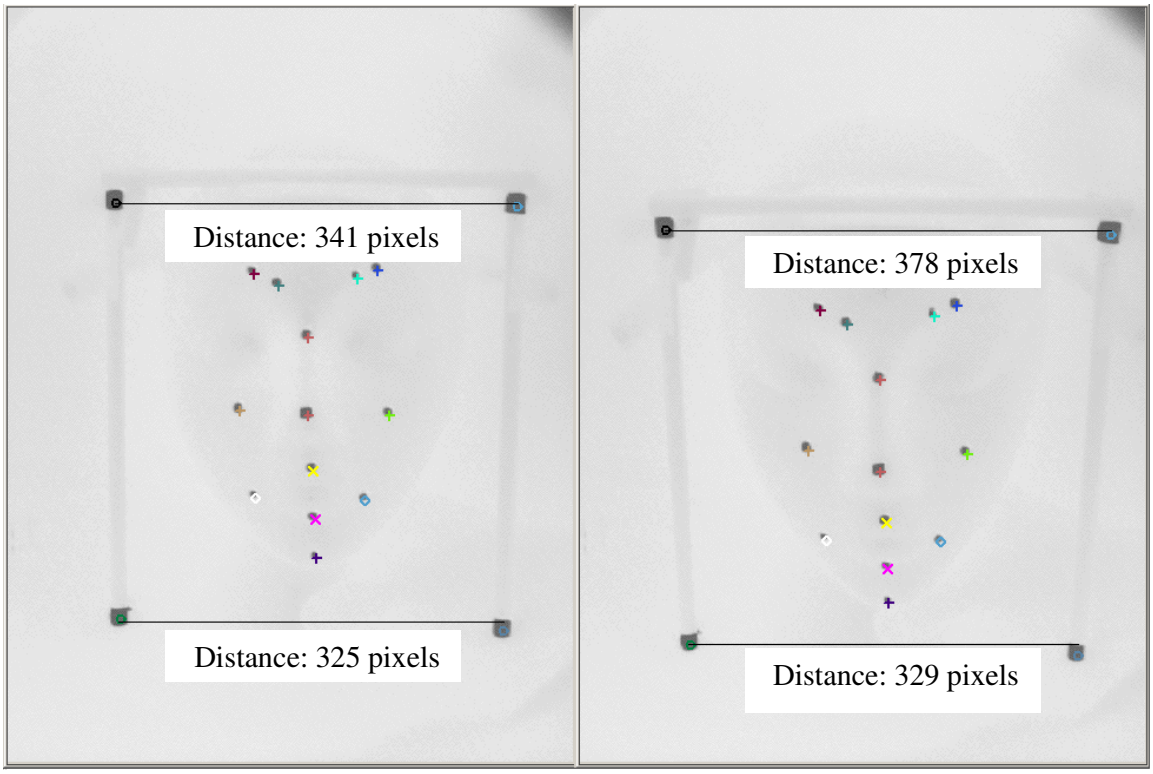


Fig. 13. Perspective projection. A segment at a distance d from a plane of projection appears larger than the same segment placed at a distance d+Δd.



(a) (b) (c)

Fig. 14. Frame appearance in different head orientations. (a) Head in neutral posture; (b) head leaning forward makes the top post distance appear larger than the bottom post distance; and (c) head turned left makes the right post distance appear larger than the left post distance.



(a) (b)

Fig. 15. Actual frames from a motion capture showing the effect of perspective projection on head appearance and *HRatio* (refer to equation (5)). (a) Head in neutral posture, *HRatio*=1.049; and (b) head with a pronounced pitch inclination, *HRatio*=1.148.

It can be seen that the horizontal ratio at a given frame t, $HRatio_t$, of the top distance to the bottom distance changes as a result of pitch angle, in this case 1.049 for the neutral position and 1.148 for the leaned forward head. The same heuristic can be applied to approximate yaw rotation, in this case by comparing the magnitude between the left and right posts. Equations (5) and (6) define the horizontal and vertical ratios respectively:

$$HRatio_t = \frac{\left|\overrightarrow{TopPost_t}\right|}{\left|\overrightarrow{BottomPost_t}\right|} \tag{5}$$

$$VRatio_t = \frac{\left|\overrightarrow{RightPost_t}\right|}{\left|\overrightarrow{LeftPost_t}\right|} \tag{6}$$

Conversion of these ratios into actual head rotation FAP units is performed using equations (8) and (10). The constant values for the conversion were obtained through calibration. For instance, the pitch constant k_p in equation (7) was experimentally determined to be 0.15 using the styrofoam head model in Fig. 10(b) and leaning it forward and backward 17°, ($Pitch\theta_{\max} = 0.296 rad$). Angles greater than 17° for the case of head pitch were considered out of range during a normal interview process. The same applies for head yaw. The only difference was that the maximum angle allowed was 8° ($Yaw\theta_{\max} = 0.139 rad$) and the constant k_y was determined to be 0.05. It is important to note that that pitch and yaw motion are tightly coupled; hence, the corresponding FAP values computed using the proposed method are just mere approximations.

$$k_p = abs(HRatio_t - 1.0) \quad measured\ when\ head\ leaned\ forward\ 17° \tag{7}$$

$$PitchFAPunit_t = \mathrm{sgn}(HRatio_t - 1.0) \times \frac{Pitch\theta_{\max}(rad) \times 100000}{k_p \Big/ \min(abs(HRatio_t - 1.0), k_p)} \tag{8}$$

$$k_y = abs(VRatio_t - 1.0) \quad measured\ when\ head\ rotated\ right\ 8° \tag{9}$$

$$YawFAPunit_t = \mathrm{sgn}(VRatio_t - 1.0) \times \frac{Yaw\theta_{\max}(rad) \times 100000}{k_y \Big/ \min(abs(VRatio_t - 1.0), k_y)} \tag{10}$$

3.2.3 Estimation of lip motion

Lip motion is estimated from four markers placed at the top and bottom lips, and the right and left corners of the mouth. Mouth opening at any given frame *t*, $Vlip_t$, could in principle be determined from the vertical distance between top and bottom markers on the lips. Unfortunately, this distance varies not only with mouth aperture but also with head orientation (particularly pitch movements) as a result of the projection onto the image plane (cf. Fig. 13). For this reason, a correction factor is applied to account for the coupling with pitch angles. The correction factor has the same magnitude whether the pitch movement is forward or backward. Therefore, the parameter $HRatio_t$ in equation (5) was slightly modified to account for this fact as shown in equation (11), and used to determine the correction factor in equation (12).

$$HRatio'_t = \max\left(\frac{\left|\overrightarrow{TopPost_t}\right|}{\left|\overrightarrow{BottomPost_t}\right|}, \frac{\left|\overrightarrow{BottomPost_t}\right|}{\left|\overrightarrow{TopPost_t}\right|} \right) \tag{11}$$

$$PitchCorrectionFactor_t = \frac{HRatio'_t}{HRatio'_0} \tag{12}$$

In addition, a non-lineal transformation was applied afterwards to the lip motion depending on the measured aperture of the mouth. Several transformations were considered, as shown in Fig. 16. Using an informal perceptual evaluation (MZC and RGO), it was concluded that the best performance was achieved with the Gaussian shaped function defined in equation (13), which filters out small lip openings (jitter) and amplifies larger ones:

$$GL\left(thr\big|_{thr>0}, Input\right) = \begin{cases} Input, & if\ \ Input/thr \le 0 \\ thr \times \exp\left(\dfrac{-\left(Input/thr - 1.0\right)^2}{2 \times 0.25^2}\right), & if\ \ 0 < Input/thr < 1 \\ Input, & if\ \ Input/thr \ge 1 \end{cases} \quad (13)$$

In summary, the vertical opening of the mouth is determined by:

(1) subtracting from the raw lip aperture ($VLip_t$) the aperture recorded at the initial frame $VLip_0$, since MPEG-4 FAPs are a measure relative to the neutral face.

(2) applying the multiplicative term $PitchCorrectionFactor$ to account for coupling with pitch rotations of the head:

$$CorrectedLipOpening_t = PitchCorrectionFactor \times \frac{VLip_t - VLip_0}{MNS_0 / 1024} \quad (14)$$

(3) applying the Gaussian transformation in equation (13) to emphasize larger openings of the oral cavity:

$$LipOpening_t = GL\left(200, CorrectedLipOpening_t\right) \quad (15)$$

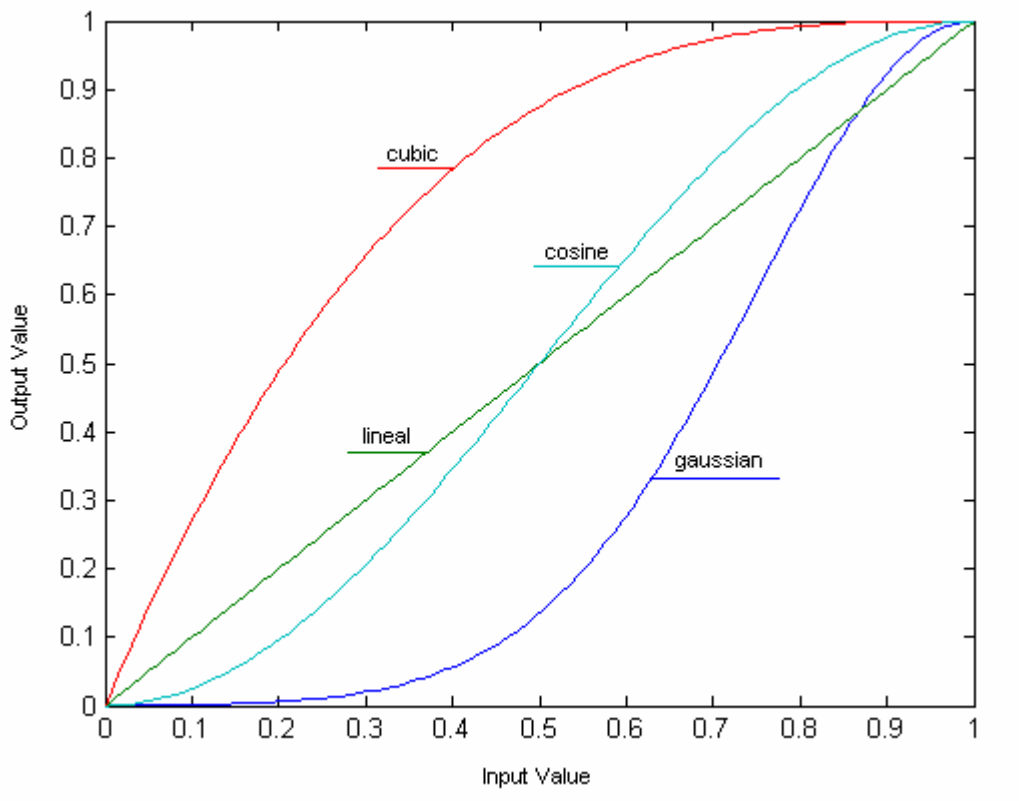


Fig. 16. Various transformations considered to smooth the effect of aliasing and head rotation coupling in lip aperture determination (shown normalized). The green curve shows the original values without modification. The red and cyan curves do not achieve the desired smoothing effect while the blue curve, Gaussian shaped, attenuates values below 0.3 and reaches the apex zone rapidly for values above it.

Finally, this lip vertical opening was proportionally converted into displacements for the bottom lip, top lip, right, and left lip FAPs using empirically-determined weights, as given in equations (16) to (19).

$$BottomLipFAP_t = -0.9 \times LipOpening_t \tag{16}$$

$$TopLipFAP_t = -0.1 \times LipOpening_t \tag{17}$$

$$RightLipFAP_t = -0.4 \times LipOpening_t \tag{18}$$

$$LeftLipFAP_t = -0.4 \times LipOpening_t \tag{19}$$

The horizontal aperture of the mouth is determined by subtracting the distance between the right and left markers on the lips in the current frame ($HLip_t$), from that in the neutral frame $HLip_0$. Although head rotations and perspective projection have an effect on this magnitude, a decision was made to keep this FAP computation as simple as possible since its contribution to lip synchronization and lip animation was at this point quite acceptable. Therefore, the horizontal displacement for these parameters is given by equations (20) and (21).

$$HRightLipFAP_t = \frac{1}{2} \times \frac{HLip_t - HLip_0}{MW_0 / 1024} \tag{20}$$

$$HLeftLipFAP_t = \frac{1}{2} \times \frac{HLip_t - HLip_0}{MW_0 / 1024} \tag{21}$$

It must be noted that the manner in which the lip FAPs are computed reproduces real lip motion quite accurately, but it is unable to capture asymmetric lip movements, nor idiosyncratic grins and smiles accompanying speech.

3.2.4 Estimation of eyebrow motion

The MPEG-4 standard employs six FAP parameters to describe eyebrow motion. As noted earlier, a decision was made to remove the outer eyebrow markers since these were frequently missed by our trackers. (cf. Section 3.2.1). Two additional parameters are introduced at this point, the distance between the marker at the nose bridge and the midpoint between inner eyebrows, $VIEyebr_t$ (defined in equation (22)), and the distance between the marker at the nose bridge and the middle eyebrows denoted as $VMEyebr_t$, defined in equation (23). These parameters, which are illustrated in Fig. 17, will be later used to compute the eyebrow displacements.

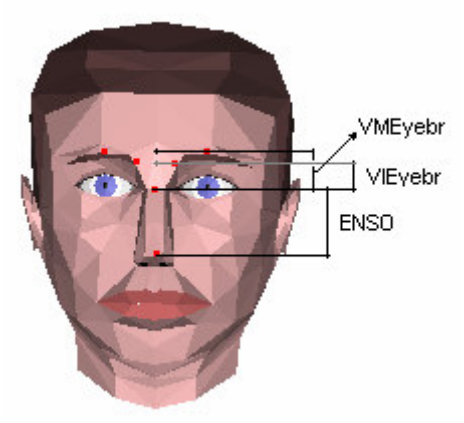


Fig. 17. Parameters associated with eyebrow motion . VMEyebr is defined as the vertical distance between the nose bridge and the mid-point of the middle eyebrows. VIEyebr is defined as the distance between the bridge and the inner eyebrows mid-point. Finally, ENS0 is the eye to nose separation in the neutral posture.

$$VIEyebr_t = \text{distance}(InnerEyebrowsMidPoint_t, NoseBridge) \quad (22)$$

$$VMEyebr_t = \text{distance}(MiddleEyebrowsMidPoint_t, NoseBridge) \quad (23)$$

As it occurs with vertical lip motion, eyebrow displacements are subject to coupling with pitch rotation. Unfortunately, the correction factor applied for lips was found not to work well for eyebrows. This phenomenon is attributed to the fact that the fiduciary point (nose bridge) does not lie in the same vertical line as the eyebrow points but is oblique and consequently the compensation factor applied for head pitch does not correct the distortion caused by head yaw. After some experimentation, a custom correction factor based on $VPostAvg_t$ (equation (24)) was applied to the measured distances $VIEyebr_t$ and $VMEyebr_t$.

$$VPostAvg_t = \frac{\left|\overrightarrow{RightPost_t}\right| + \left|\overrightarrow{LeftPost_t}\right|}{2} \quad (24)$$

This correction factor simply attempts to neutralize the difference in appearance caused by head motion on the parameters $VIEyebr_t$ and $VMEyebr_t$ at frame *t*, prior to the subtraction to the reference parameters at frame zero, ($VIEyebr_0$ and $VMEyebr_0$). This is achieved by multiplying the terms at the current frame by $VPostAvg_0 / VPostAvg_t$. Subsequently, a Gaussian shaped transformation (equation (25)) was applied in the same fashion as for lip FAPs.

$$GE\left(thr\big|_{thr>0}, Input\right) = \begin{cases} Input, & if\ \ abs\left(Input/thr\right) \geq 1 \\ thr \times \exp\left(\dfrac{-\left(abs(Input/thr)-1.0\right)^2}{2\times 0.25^2}\right), & if\ \ 0 < Input/thr < 1 \\ -thr \times \exp\left(\dfrac{-\left(abs(Input/thr)-1.0\right)^2}{2\times 0.25^2}\right), & if\ \ 0 > Input/thr > -1 \end{cases} \quad (25)$$

Corrected displacements were computed using equations (26) and (27), and later converted (equations (28) and (29)) to obtain the final FAP displacements for eyebrow motion.

$$CorrectedInnerEyebrow_t = \frac{\left(\dfrac{VPostAvg_0}{VPostAvg_t} \times VIEyebr_t\right) - VIEyebr_0}{ENS_0 / 512} \quad (26)$$

$$CorrectedMiddleLeftEyebrow_t = \frac{\left(\dfrac{VPostAvg_0}{VPostFrmAvg_t} \times VMEyebr_t\right) - VMEyebr_0}{ENS_0 / 512} \quad (27)$$

$$InnerLeftEyebrowFAP_t = GE\left(50, CorrectedInnerEyebrow_t\right) \quad (28)$$

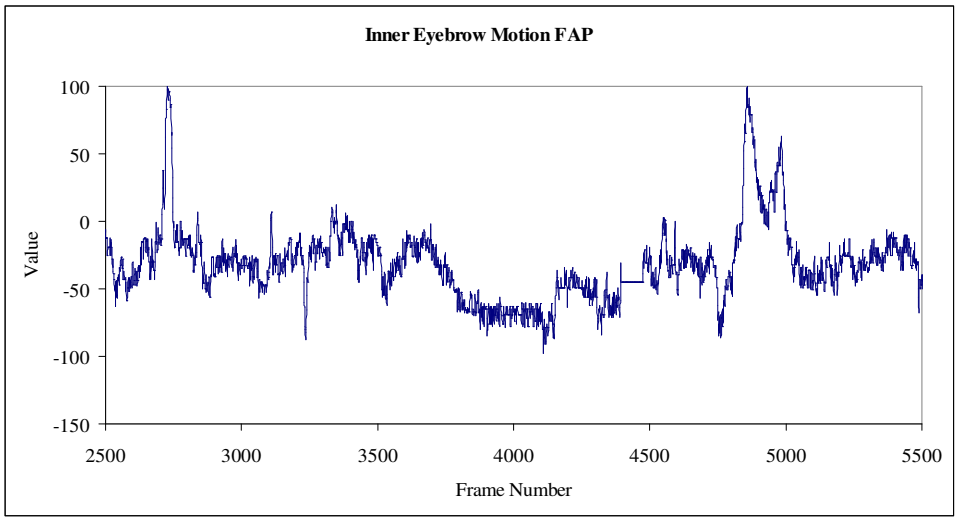
$$MiddleLeftEyebrowFAP_t = GE\left(100, CorrectedMiddleEyebrow_t\right) \quad (29)$$

$$InnerRightEyebrowFAP_t = InnerLeftEyebrowFAP_t \quad (30)$$

$$MiddleRightEyebrowFAP_t = MiddleLeftEyebrowFAP_t \quad (31)$$

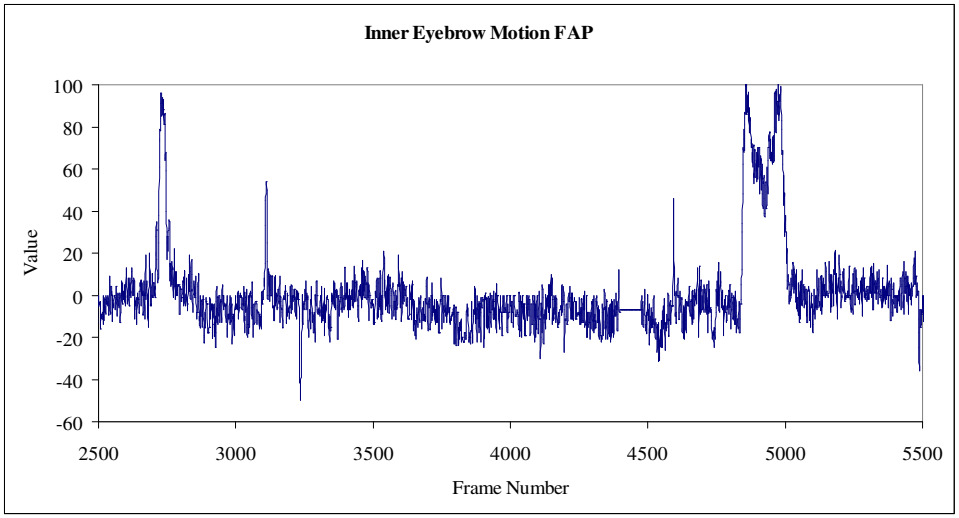
Finally, to eliminate high-frequency jitter, a non-causal average filter was applied to the estimated eyebrow FAPs. The window size for the filter was set to 5 frames (or 167ms)

since typical “eye-greeting” lasts approximately 160ms, whereas “eyebrow flash” persists approximately 300ms [51]. Fig. 18 shows the eyebrow motion signals at different stages during the process.

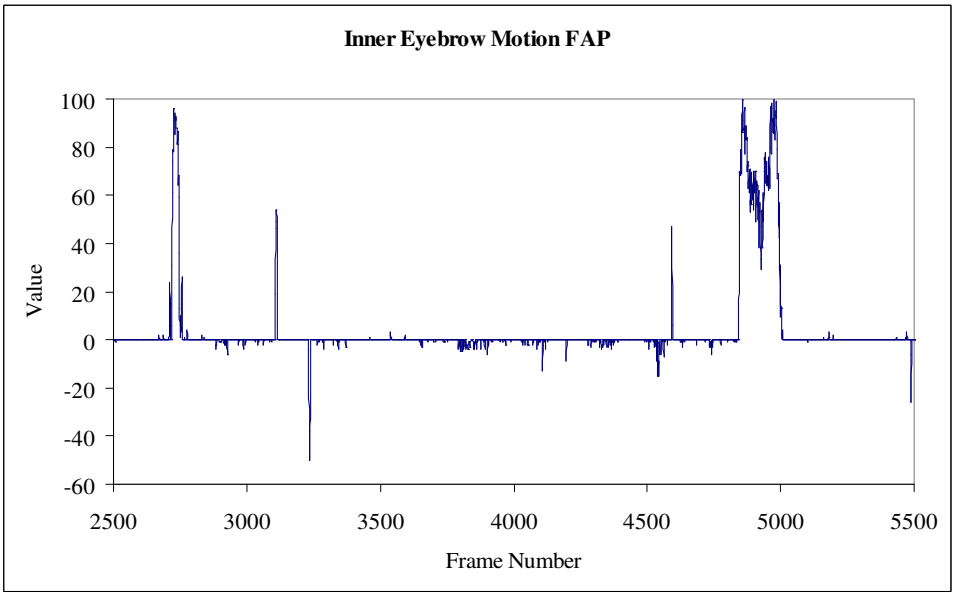


(a)

Fig. 18. Waveforms from a real motion capture. (a) Inner eyebrow motion without correction factor; (b) inner eyebrow motion after the application of the correction factor; (c) following Gaussian squashing; and (d) following average filtering.

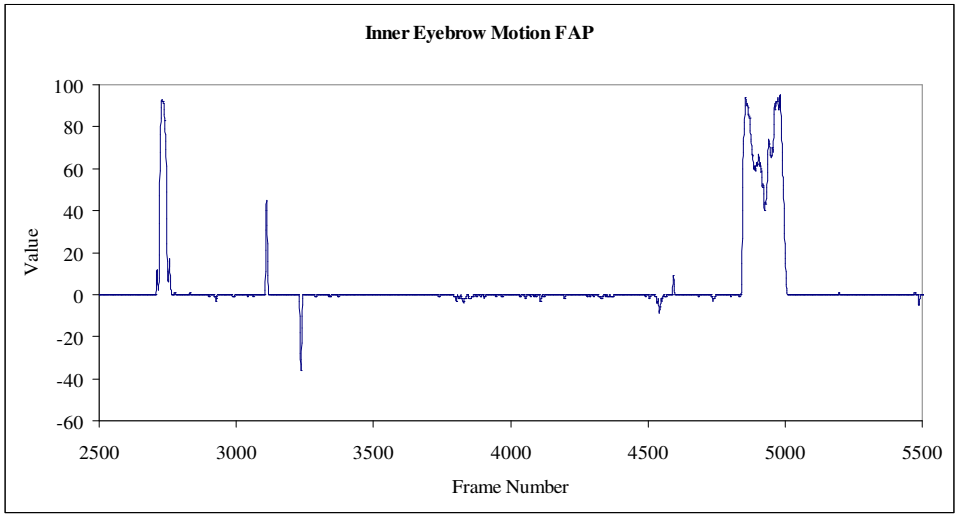


(b)



(c)

Fig. 18. Continued.



(d)

Fig. 18. Continued.

4 ELICITATION AND PREDICTION OF VISUAL PROSODY

This section describes the protocol that was devised to elicit and capture natural facial motion from English speakers, as well as the two computational models that were built to generate visual prosody. The two different forms of visual prosody: randomly generated movements, which served as the baseline stimulus, and speech-driven prosody by means of heuristics and autoregressive models.

4.1 Prosody elicitation protocol

Using the motion capture system described in the previous section, nine interviews were carried out to acquire facial motion data. Due to the limitations of the tracking system, some constraints were introduced at the time of motion capture: a) the use of eyeglasses was not allowed during the whole interview, and b) male subjects had no facial hair, such as beard or moustache. Small adhesive reflective markers (less than 2x2mm) were placed on the individual's face at thirteen points previously defined, (refer to Section 3.2.1, Fig. 10(b) for more details).

An interview protocol was designed to establish a baseline for facial motion across subjects. The process consisted of three distinct sections:

1. Description of a childhood game or a life threatening experience. This section was intended to familiarize the individual with the overall motion-capture

process, the reflective markers and the head frame. Since the topic was selected by the subject, this greatly helped him/her get relaxed in front of the camera and microphone.

2. View of a selected video sequence, in this case a Looney Tunes cartoon entitled "Putty Tat Trouble" [58]. This section served to establish a common story-telling scenario.
3. Scene description. The person was asked to describe the Tweety/Silvester cartoon in front of the camera. No specific format was followed. The aim of this stage was to elicit and capture idiosyncratic visual prosody from the subject.

Each recording session lasted approximately two hours. After each session, the data was processed in order to store the relative position of each marker during the recording, and later generate FAP displacements. About ten minutes of animation were produced for each individual, containing several narrations and the common story-telling of the cartoon. In most cases, little or no intervention was needed or recorded from the interviewer's side.

4.2 Video selection

Two videos were manually selected from among the nine recordings for further study due to their quality in terms of head and eyebrow motion. The remaining videos had to be discarded since they only displayed subtle facial motion and, hence, less opportunity

to exploit and acquire meaningful relationships between utterances and visual prosody. The two selected segments (those of subjects S2 and S7) have high variance on the head and eyebrow articulators, as shown in TABLE 2. However, higher variance does not necessarily equate to 'pleasant' animation. Such is the case of segment S9, in which the motion appears rhythmic and more the result of nervousness than natural story-telling.

TABLE 2

Statistics from selected FAP features in video samples

Video Name	**Gender**	**Duration mm:ss**	**Statistic**	**Inner eyebrow FAP 31,32**	**Middle eyebrow FAP 33,34**	**Head pitch FAP 48**	**Head yaw FAP 49**	**Head roll FAP 50**
S1	Male	1:20	std. dev.	0.14	6.07	1281.31	2159.74	1842.55
			mean	-0.01	-1.02	-204.04	6141.41	-2320.05
S2	Male	1:48	std. dev.	**11.81**	**11.81**	**3890.67**	**4677.11**	**3267.26**
			mean	5.54	5.54	-1102.69	-2575.43	-1489.92
S3	Male	2:05	std. dev.	5.55	4.77	2653.34	3901.66	3063.32
			mean	0.60	0.37	-3164.00	-852.93	987.31
S4	Male	1:35	std. dev.	2.57	0.13	1281.49	2234.09	2601.38
			mean	0.17	0.01	-903.16	1850.82	4977.86
S5	Male	1:24	std. dev.	0.40	0.00	752.93	854.61	408.18
			mean	0.18	0.00	-3545.07	-1312.69	110.82
S6	Male	3:40	std. dev.	10.86	13.83	1911.56	995.31	1176.29
			mean	-15.93	-26.35	-6495.01	4412.84	-2906.73
S7	Male	1:26	std. dev.	**12.75**	**12.75**	**3615.77**	**4291.83**	**3385.36**
			mean	6.59	6.59	-640.30	-1789.50	-1205.07
S8	Male	1:34	std. dev.	1.34	2.31	1833.98	3822.31	3706.54
			mean	0.28	0.51	1411.45	-1744.87	7132.25
S9	Female	1:49	std. dev.	13.26	11.79	4756.03	4333.57	5254.98
			mean	7.00	3.14	131.30	6475.49	-4354.76

4.3 Visual prosody models

In order to investigate the perceptual role of head and eyebrow motion in the context of facial avatars, four animations were produced for each of the two video snippets that were selected in the previous section:

- No visual prosody (NO_PROSODY)
- Random visual prosody (RANDOM)
- Speech driven visual prosody (SPEECH_DRIVEN)
- Ground truth visual prosody (GROUND_TRUTH)

In the four cases, lip movement was produced using the ground truth from video, since lip articulation was not a variable of interest in the present study.

The first and fourth models are straightforward, and are described in the next two paragraphs. Random and speech-driven models are more involved, and deserve separate treatment in independent subsections.

4.3.1 No visual prosody

The production of the first animation model is trivial, since it only involves setting to zero the corresponding FAP values for inner eyebrows (FAP 31, FAP 32), middle eyebrows (FAP 33, FAP 34), head pitch (FAP 48), head yaw (FAP 49), and head roll (FAP 50).

4.3.2 Ground truth visual prosody

The production of the last animation model is also trivial, as it employs the head and eyebrow movements that were extracted from the video. This animation model is important, as it provides a best-case scenario for visual prosody.

4.3.3 Random visual prosody

Randomly generated head and eyebrow motion employed a special noise function widely used in computer animation, known as Perlin Noise [59]- [62]. This special type of noise function has also been used by Perlin and Goldberg [63] in a scripting system to generate real-time animated characters capable of displaying behavioral motion. Perlin Noise is based on a fractal summation of pseudo-random functions:

$$PerlinNoise = \sum_{i=0}^{i=\#octaves} persistence^{i} \times Noise\left(frequency^{i} \times input\right) \tag{32}$$

The behavior of this noise function is controlled by means of the persistence and frequency parameters, as well as with the number of octaves. To better understand the function of persistence, frequency, and number of octaves an example is adapted from reference [61]. Fig. 19 shows the gradual summation of noise functions to produce a given output. As it can be seen, the persistence parameter diminishes the power of subsequent octaves, so called because the frequency of an octave is a multiple of the previous one. The persistence regulates the influence of subsequent octaves in the total

summation, while the sampling frequency changes its frequency content. The number of octaves can be interpreted as the level of granularity desired.

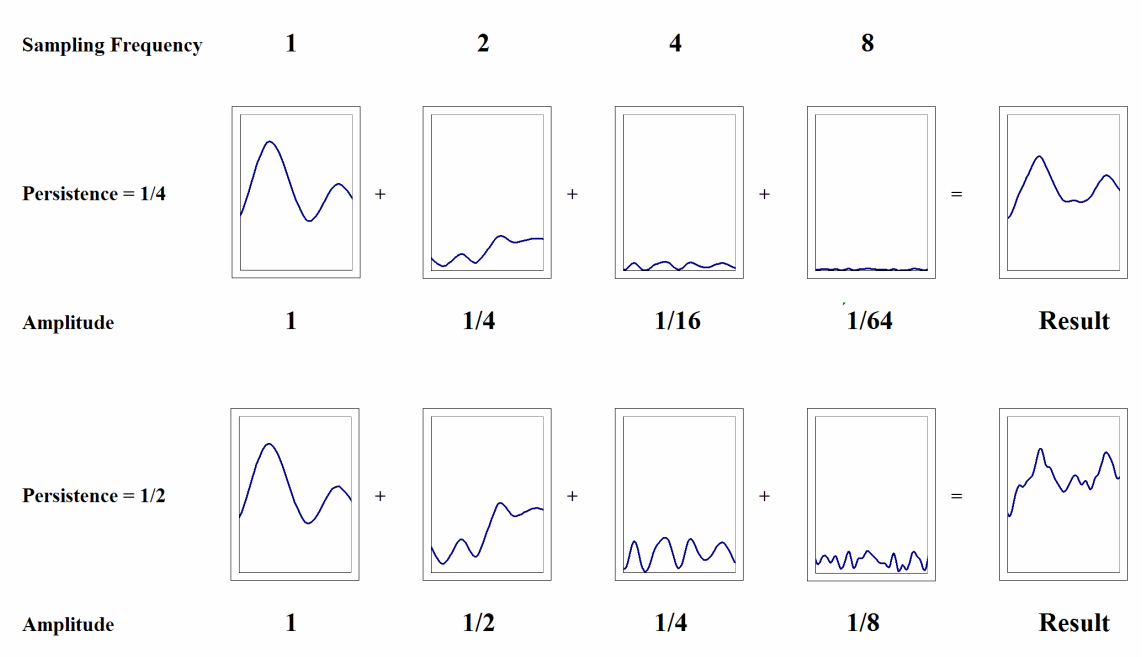


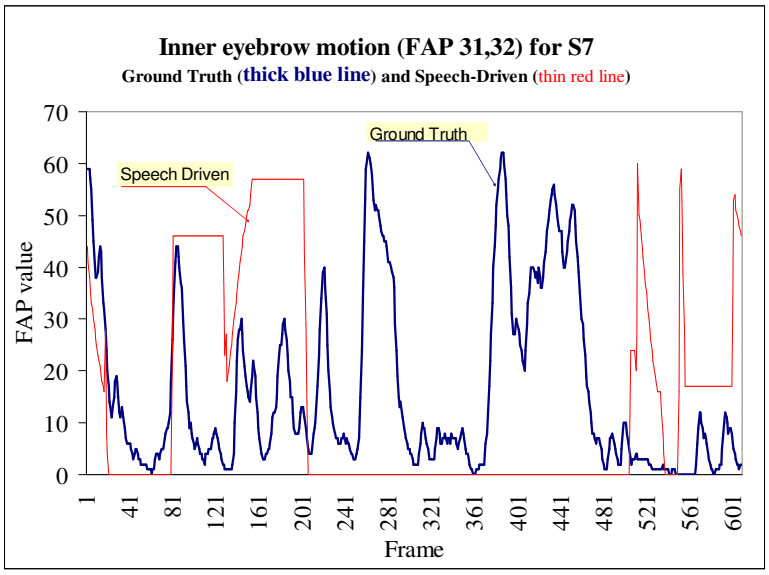
Fig. 19. Two examples of Perlin noise generation with a different persistence parameter. The final output is the summation of several octaves with a decreasing magnitude (due to a persistence parameter) and increasing seed frequency (due to a frequency modifier). Adapted from [61].

In order to generate eyebrow motion, the number of octaves was set to 6, while persistence was set to 0.8 and the seed was a function of the frame number as shown in equations (33) and (34). These values were set empirically to approximate typical values of eyebrow motion from the real motion captured data.

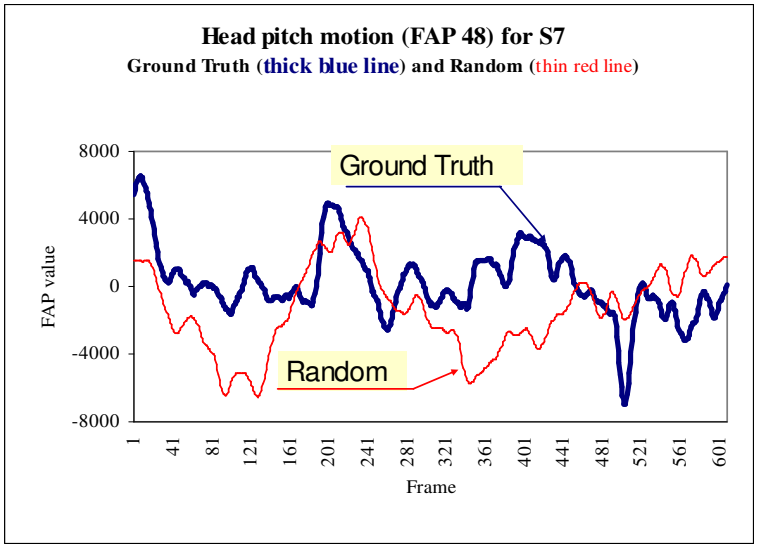
$$PerlinNoise = \sum_{0}^{6} 0.8^{i} \times Noise\left(2^{i} \times \left(\frac{frame\#}{100} + frameOffset \right) \right) \tag{33}$$

$$EyebrowFAP = \begin{cases} EyebrowScaleFactor \times PerlinNoise, & if\ \ PerlinNoise > 0 \\ 0, & if\ \ PerlinNoise \le 0 \end{cases} \tag{34}$$

The same noise function was used for head motion, with scale factors and frequency values adjusted to account for differences in units since eyebrow FAPs use the eye to nose separation (ENS0/1024) as a displacement unit whereas head rotations use 10^{-5}rad as the angular unit. In addition, negative values were allowed, as opposed to eyebrow generated motion in which negative values were neglected. Finally, each head inclination (head, yaw, and roll) contained a different *frameOffset*. Fig. 20 shows the final trajectories of the randomly generated visual prosody for one of the subjects; obviously, these trajectories are uncorrelated with the ground truth motion, but nonetheless have similar frequency content.

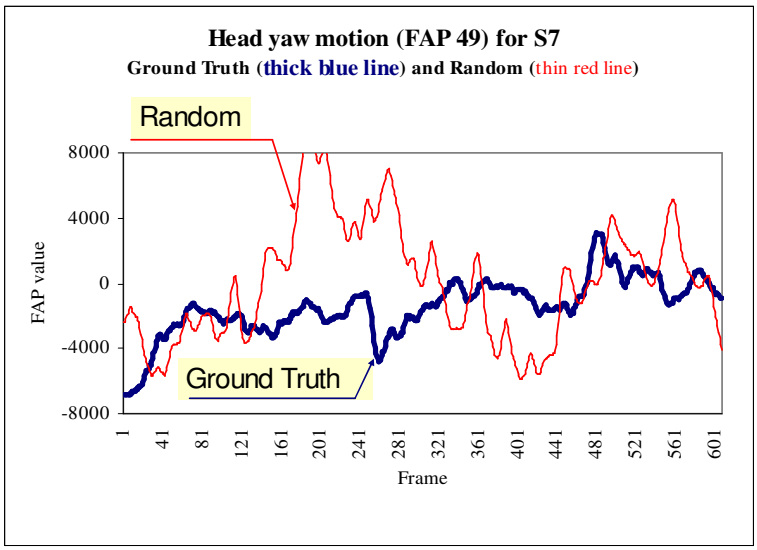


(a)

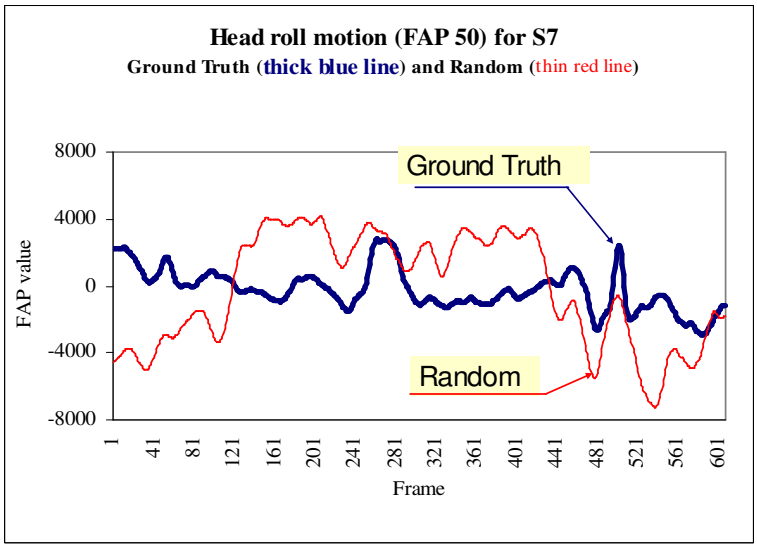


(b)

Fig. 20. Ground truth data (solid blue line) compared to random motion (dashed red line) for a video segment of 600 frames (40 seconds) containing S7 idiosyncratic for: (a) inner eyebrow motion; (b) pitch motion; (c) yaw motion; and (d) roll motion.



(c)



(d)

Fig. 20. Continued.

4.3.4 Speech driven visual prosody

Cavé et al. [38] have shown that there is a high correlation between rising Fundamental Frequency (*F0*) events and eyebrow rising-falling movement. Nonetheless, that study also showed that eyebrow motion can occur in silent segments as well as in flat *F0* regions, which were attributed to linguistic communicational choices. This result suggests that the relationship between *F0* and eyebrow motion is non-trivial. Even though full recovery of the visual prosody from speech acoustics may not be possible, we hypothesize that visual prosody driven by simple acoustic features (e.g., pitch and energy contours) may still be perceptually more realistic than randomly-generated or no visual prosody at all. To test this hypothesis, two simple computational models were used to generate eyebrow and head movements. Eyebrows were animated using a rule-based heuristic, whereas head movements were predicted using a linear autoregressive model.

4.3.4.1 Generation of eyebrow movements

Eyebrow motion was driven by the fundamental frequency component and the energy of the speech signal. Both variables were computed using the PRAAT tool from Boersma and Weenink [64]. Fig. 21 shows the analysis of the sound channel for video segment S7. The top signal in solid black line represents the corresponding Pulse Code Modulation values, while the bottom drawn in cyan line represents the fundamental frequency candidate computed for voiced segments.

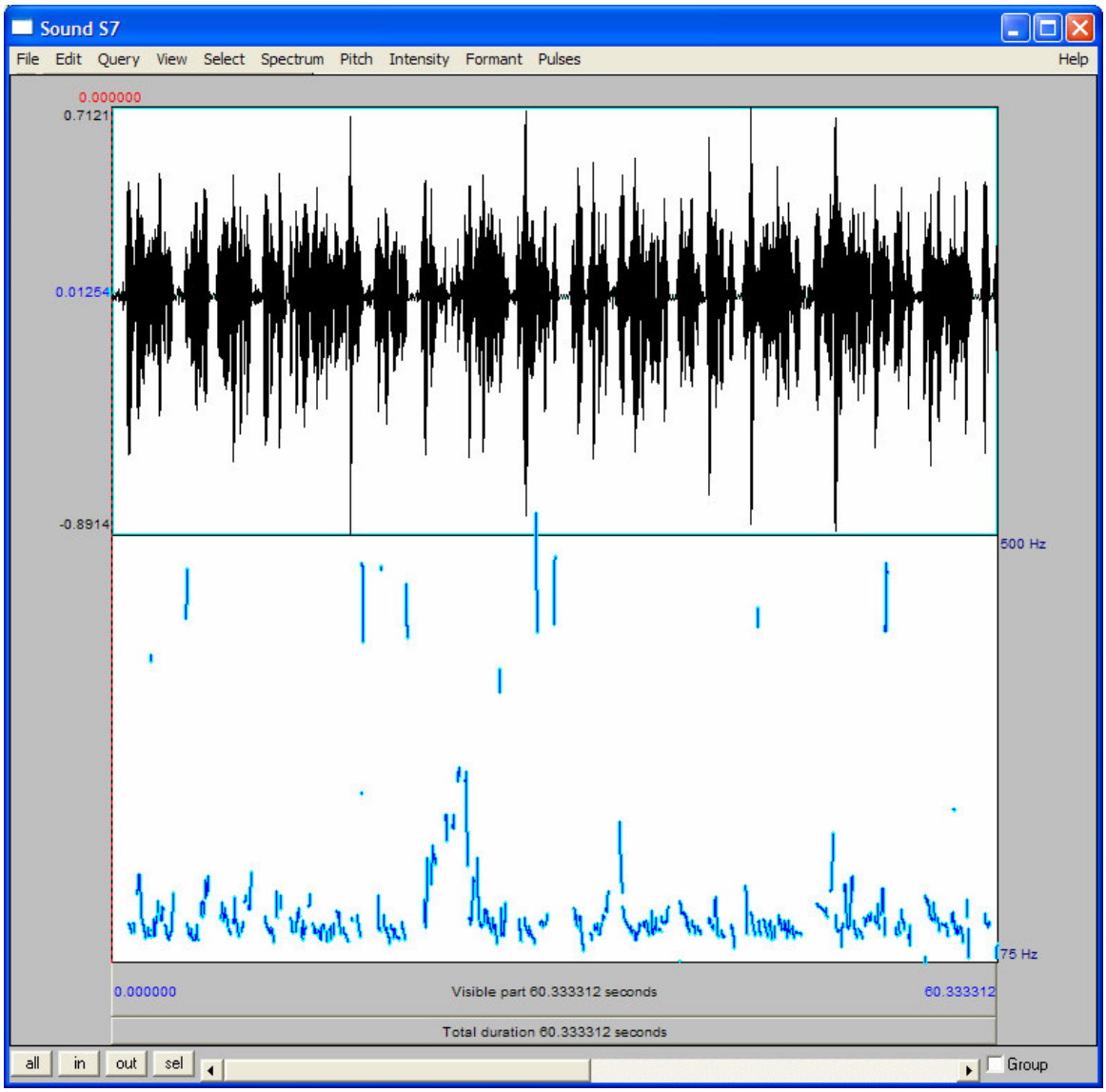


Fig. 21. Screen sample from the PRAAT tool from Boersma and Weenink [64] showing the fundamental frequency (*F0*) analysis for the test segment S7.

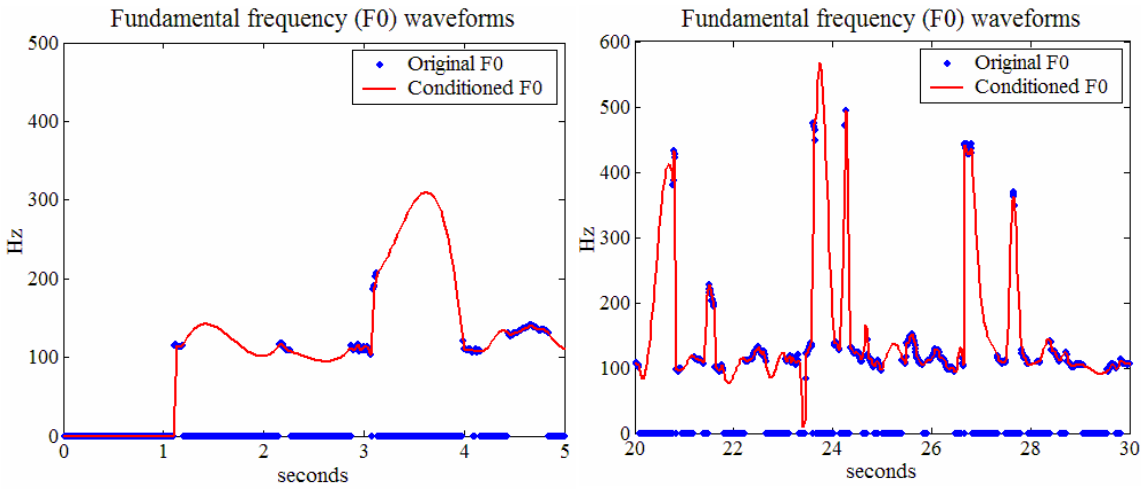
Since *F0* values for unvoiced segments such as the ones produced by certain consonants (e.g., /p/ or /t/) are not defined, the undefined segments were treated as missing points,

and interpolated values were obtained using a cubic spline. Consequently, transitions between *F0* regions were smoothed. In addition, the resulting function was cropped to limit its values between zero and the original maximum frequency to avoid outliers:

$$F0Conditioned = \begin{cases} \max(F0), & if\ \ spline(F0) > \max(F0) \\ spline(F0), & if\ \ 0 < spline(F0) < \max(F0) \\ 0, & if\ \ spline(F0) < 0 \end{cases} \tag{35}$$

where *spline()* refers to the output of the interpolating function. Fig. 22 shows an example of the raw and conditioned fundamental frequency for an audio signal at two different time scales. The spline follows the original signal during voiced segments, and provides a gross reconstruction during unvoiced utterances.

Once *F0Conditioned* was obtained, the rising edges were analyzed to determine if a pre-set limit of 207Hz was crossed. If such condition occurred, an eyebrow-rising event was automatically triggered. The eyebrow FAP displacement magnitude was determined as a scaled version of the fundamental frequency signal, with an appropriate offset value. In addition, the eyebrow displacement was maintained for a minimum of 300ms and terminated once the energy level dropped below 45dB with a gradual motion to neutral state that lasted three additional video frames.



(a) (b)

Fig. 22. Fundamental frequency waveforms from the audio channel of a motion capture. In blue dotted line the original *F0* signal, while the conditioned signal is displayed in solid red line. (a) First 5 seconds of a motion capture; and (b) another time segment belonging to the same motion capture section.

Fig. 23 to Fig. 25 show the audio parameters (*F0* and energy) for a motion capture segment and the corresponding speech driven eyebrow motion. It must be noted that the eyebrow displacement during the lapse between 10 and 20 seconds is an instance of eyebrow displacement hold due to an energy level above the threshold.

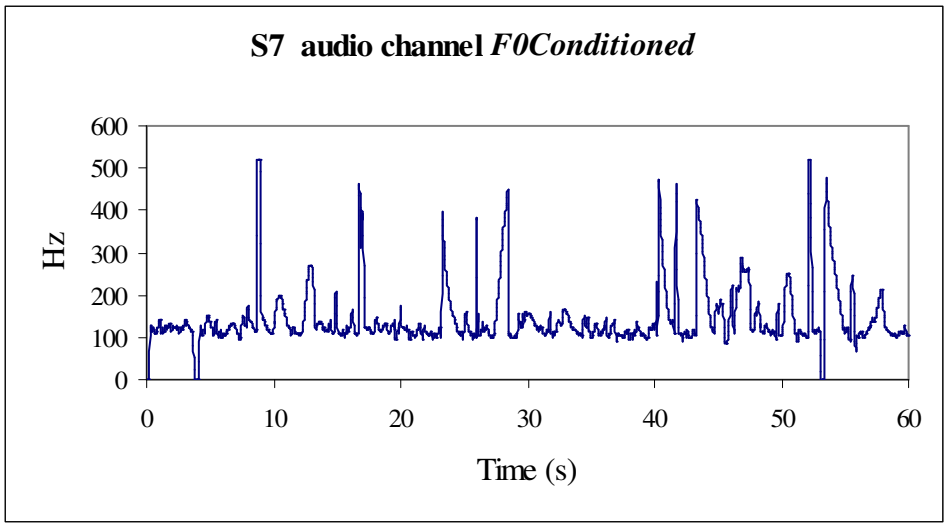


Fig. 23. S7 segment's *F0Conditioned* signal (refer to Eq. 35) for use in speech-driven facial animation.

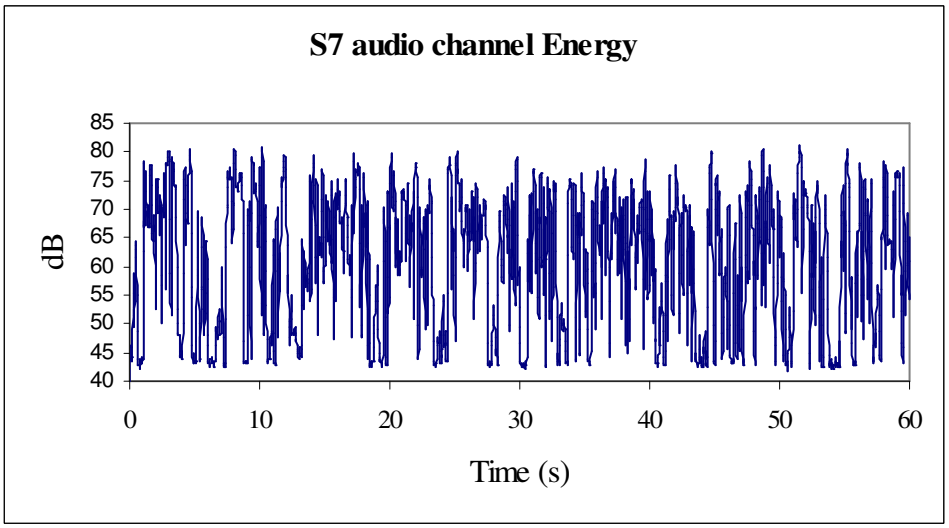


Fig. 24. S7 segment's energy signal for use in speech-driven facial animation.

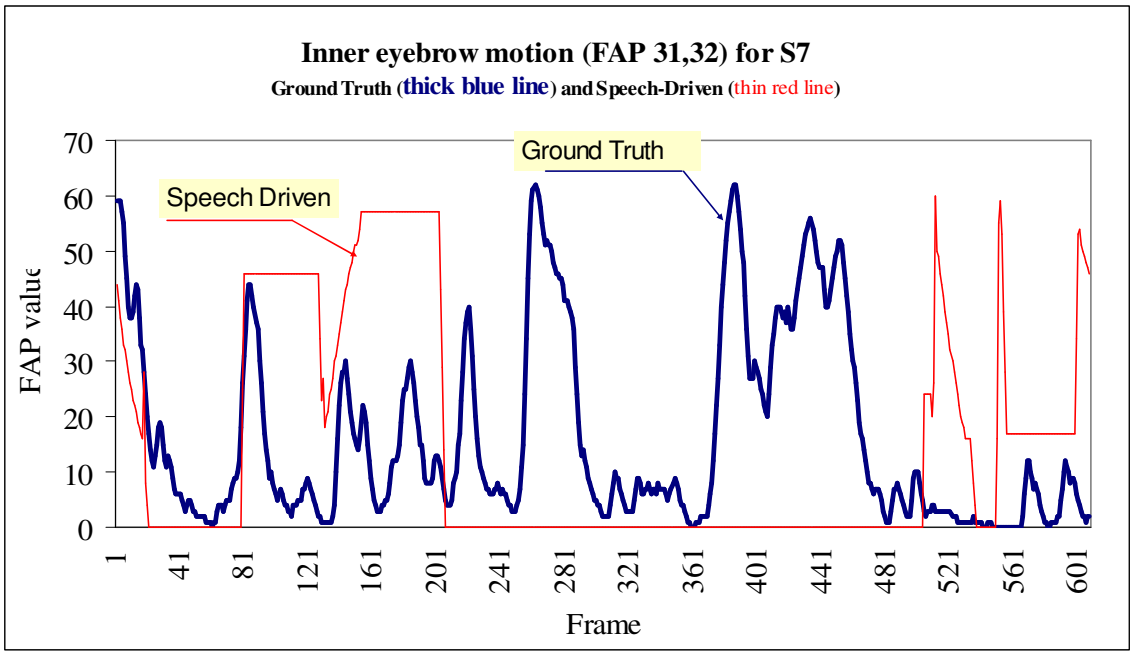


Fig. 25. Portion of the inner eyebrow motion (FAP 31 and 32) generated using the conditioned fundamental frequency and the energy parameters extracted from audio.

4.3.4.2 Generation of head movements

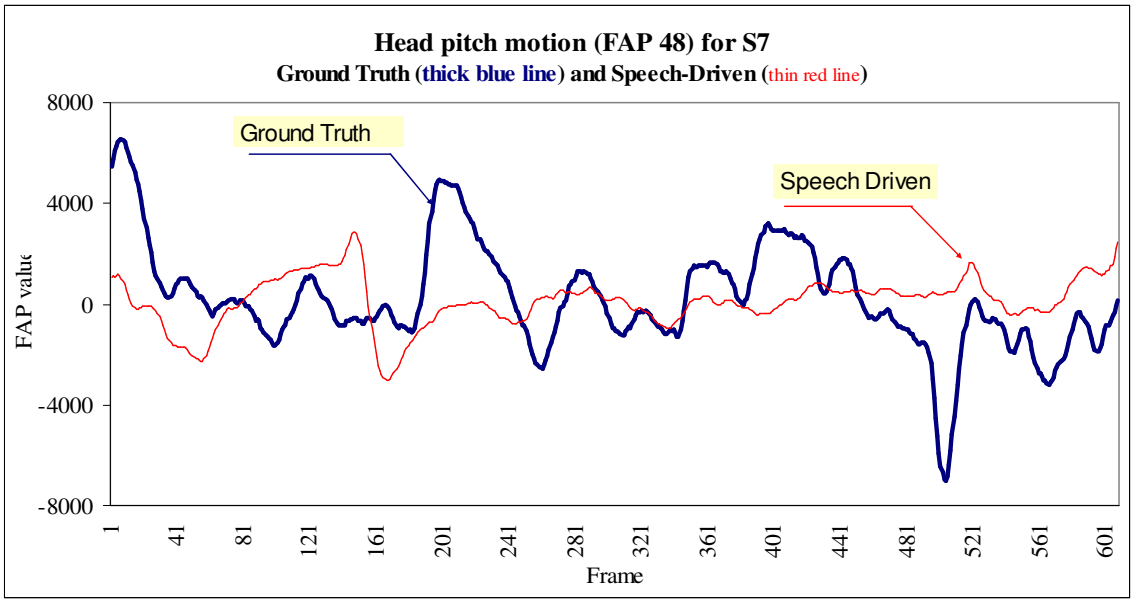
Head movements were generated using an autoregressive (ARX) model [65], which uses a linear combination of past input and output signals to compute the output signal at a later time. The model is specified by:

$$y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na) = \\ b_1 u(t-nk) + b_2 u(t-nk-1) + \ldots + b_{nb} u(t-nk-nb+1) + e(t) \quad (36)$$

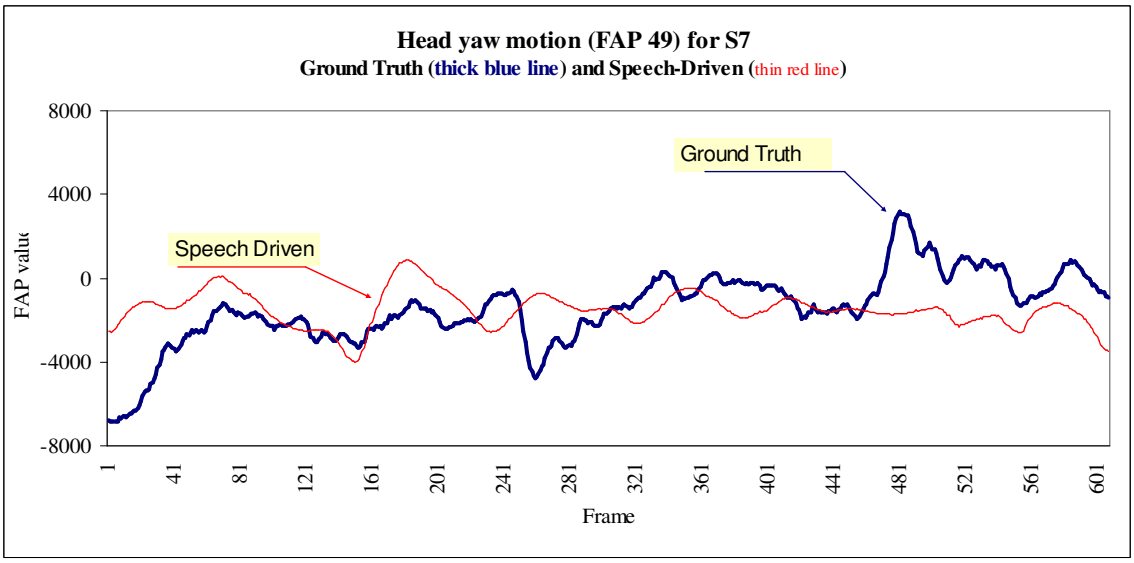
where *na*=4, *nb*=4, and *nk*=1 in our implementation. The input signals *u*, which serve as independent variables for the regression, were:

- previous predictions of head movements (pitch, yaw, and roll),
- energy level of the speech signal,
- *F0Conditioned*, as described in the previous section,
- the product of energy and *F0Conditioned*, to allow for simple non-linear effects
- mean-filtered energy and F0 contours (window width of 30 frames or 1 second), to allow the ARX model to operate at two different time scales.

The output consisted on the three desired head motion values: pitch, yaw, and roll movement. Fig. 26 shows a sample of audio-driven visual prosody generated for the video segment S7.

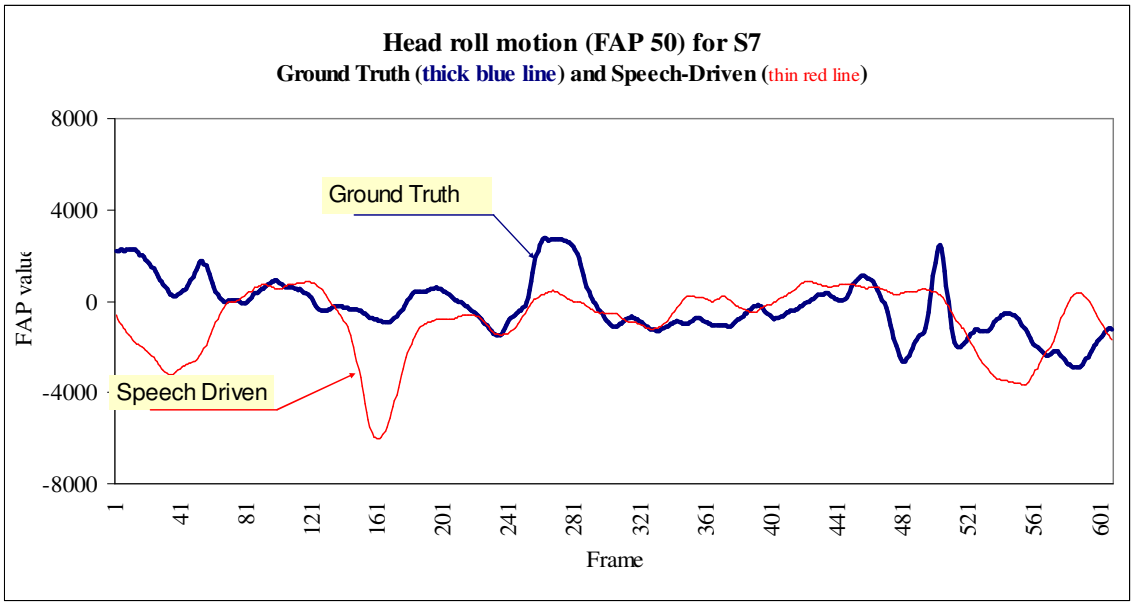


(a)



(b)

Fig. 26. Speech driven facial animation parameters generated for video segment S7 (first 600 frames, or 40 seconds). (b) Head pith; (c) head yaw; and (d) head roll.



(c)

Fig. 26. Continued.

5 PERCEPTUAL EVALUATION OF VISUAL PROSODY

This section describes the final perceptual evaluation of the four visual prosody models using a pool of subjects. The goal of the experiment was to determine whether there exists statistically significant differences between the models, and determine whether speech-driven visual prosody produced a more realistic, coherent and convincing animations than randomly generated movements, i.e., the main hypothesis of this work.

5.1 Stimulus presentation

Stimuli were presented in pairs with the same underlying audio track. For this purpose, a software interface was developed by the author to drive two instances of the Facial Animation Engine (FAE) [21] in synchrony with two separate motion data files. The basic code used to send commands to the FAE was taken from [66]. The perceptual experiments were carried on a Notebook Intel Pentium IV 3.08 GHz with 512MB RAM. The two FAE instances were run as separate high-priority threads, side by side, as shown in Fig. 27.

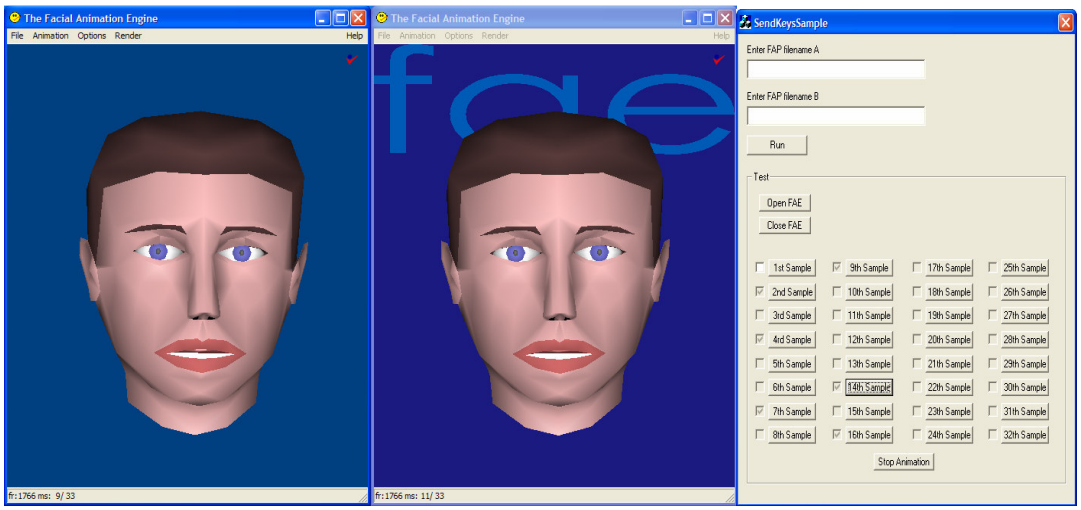


Fig. 27. Software interface based on [66] modified to drive two high priority instances of the Facial Animation Engine (FAE) [21] to play two animations in synchrony.

5.2 Experiment 1

The four stimuli (NO_PROSODY, RANDOM, SPEECH_DRIVEN, and GROUND_TRUTH) were presented to five subjects in a pair-wise fashion, for a total of 32 pairs (4x4 combinations, times two segments: S2 and S7). The subjects were asked the following question: "*Which animation do you consider to be more realistic motion-wise?*" The subjects were also instructed to dismiss any lip motion differences because the utterances for both talking heads were synchronized and the lip motion was the same. In order to discard any bias due presentation order or position (right or left), the stimulus pairs were presented in a random order, as shown in TABLE 3.

TABLE 3

Sample order for presentation of stimulus pairs

	Presentation of S2 segment	
Order	**Left Panel**	**Right Panel**
1	RANDOM	SPEECH_DRIVEN
2	SPEECH_DRIVEN	NO_PROSODY
3	GROUND_TRUTH	RANDOM
4	SPEECH_DRIVEN	SPEECH_DRIVEN
5	NO_PROSODY	RANDOM
6	RANDOM	NO_PROSODY
7	GROUND_TRUTH	SPEECH_DRIVEN
8	GROUND_TRUTH	NO_PROSODY
9	NO_PROSODY	NO_PROSODY
10	SPEECH_DRIVEN	RANDOM
11	RANDOM	RANDOM
12	NO_PROSODY	SPEECH_DRIVEN
13	NO_PROSODY	GROUND_TRUTH
14	RANDOM	GROUND_TRUTH
15	SPEECH_DRIVEN	GROUND_TRUTH
16	GROUND_TRUTH	GROUND_TRUTH

	Presentation of S7 segment	
Order	**Left Panel**	**Right Panel**
17	RANDOM	SPEECH_DRIVEN
18	SPEECH_DRIVEN	NO_PROSODY
19	RANDOM	GROUND_TRUTH
20	NO_PROSODY	GROUND_TRUTH
21	NO_PROSODY	NO_PROSODY
22	SPEECH_DRIVEN	RANDOM
23	NO_PROSODY	RANDOM
24	SPEECH_DRIVEN	GROUND_TRUTH
25	RANDOM	RANDOM
26	GROUND_TRUTH	NO_PROSODY
27	NO_PROSODY	SPEECH_DRIVEN
28	SPEECH_DRIVEN	SPEECH_DRIVEN
29	RANDOM	NO_PROSODY
30	GROUND_TRUTH	SPEECH_DRIVEN
31	GROUND_TRUTH	RANDOM
32	GROUND_TRUTH	GROUND_TRUTH

The survey results are presented in TABLE 4 for S2 video and TABLE 5 for S7 video in the form of a confusion matrix. A letter (A) in the cell means that the stimulus in the left panel was judged to be more realistic than the stimuli in the right panel. For instance, the comparison between GROUND_TRUTH and NO_PROSODY, located in the 1st column and 4th row, reads ‘AAABB’, which means that 3 out of 5 subjects preferred the GROUND_TRUTH animation. The results of this preliminary survey, collapsed by the number of ballots each model received, are shown below in TABLE 6, where the score equals the number of times a particular model was selected as more realistic motion-wise (regardless of whether it was displayed on the left or the right panel).

TABLE 4

Confusion matrix results for video S2 in experiment 1

S2 Video pair-wise comparison		Right panel animation			
		NO_PROSODY	RANDOM	SPEECH-DRIVEN	GROUD_TRUTH
Left panel animation	NO_PROSODY	BAABB	BBBBB	BBBBB	BBBBB
	RANDOM	AAAAA	BBBBA	AAAAA	ABBAB
	SPEECH_DRIVEN	AAAAA	ABBBA	BABBB	BBBAB
	GROUD_TRUTH	AAABB	AAAAA	AAAAA	ABAAA

TABLE 5

Confusion matrix results for video S7 in experiment 1

S7 Video pair-wise comparison		Right panel animation			
		NO_PROSODY	RANDOM	SPEECH-DRIVEN	GROUD_TRUTH
Left panel animation	NO_PROSODY	AAABA	BABAA	BBBBB	BBBBB
	RANDOM	AAAAA	ABABB	AABAB	BBAAB
	SPEECH_DRIVEN	AAAAA	BBABA	BBBBB	BBBBA
	GROUD_TRUTH	ABAAA	BABBB	BBAAB	BAAAB

TABLE 6

Collapsed number of ballots for experiment 1

Model	**S2 Score**	**S7 Score**	**S2&S7 Score**
NO_PROSODY	7	9	16
RANDOM	25	24	49
SPEECH_DRIVEN	18	23	41
GROUND_TRUTH	30	24	54
		Total	160

These results show that the NO_PROSODY model is perceived as the least realistic, whereas the GROUND_TRUTH model scores the highest, followed by RANDOM and SPEECH-DRIVEN models. A student's T-test was performed on the ratings to determine if these differences are statistically significant. The data was collapsed per person surveyed, i.e. five surveys (four degrees of freedom). The statistics in TABLE 7 confirm that for both videos (S2 and S7) the NO_PROSODY animation model is statistically different from RANDOM, SPEECH-DRIVEN, and GROUND_TRUTH models (e.g., refer to Pair 1, Pair 2, Pair 3, Pair7, Pair 8, Pair 9 significance). On the other hand, the difference between the remaining three models (RANDOM, SPEECH-DRIVEN, and GROUND_TRUTH) was found to be not statistically significant.

TABLE 7

T-test pair-wise mean comparison for experiment 1

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of Mean Difference Lower	Upper			
Pair 1	S2_NO_PR - S2_RANDO	-3.6000	.89443	.40000	-4.7106	-2.4894	-9.000	4	.001
Pair 2	S2_NO_PR - S2_SP_DR	-2.2000	.44721	.20000	-2.7553	-1.6447	-11.000	4	.000
Pair 3	S2_NO_PR - S2_GRD_T	-4.6000	1.67332	.74833	-6.6777	-2.5223	-6.147	4	.004
Pair 4	S2_RANDO - S2_SP_DR	1.4000	.89443	.40000	.2894	2.5106	3.500	4	.025
Pair 5	S2_RANDO - S2_GRD_T	-1.0000	1.73205	.77460	-3.1506	1.1506	-1.291	4	.266
Pair 6	S2_SP_DR - S2_GRD_T	-2.4000	1.67332	.74833	-4.4777	-.3223	-3.207	4	.033
Pair 7	S7_NO_PR - S7_RANDO	-3.0000	1.87083	.83666	-5.3229	-.6771	-3.586	4	.023
Pair 8	S7_NO_PR - S7_SP_DR	-2.8000	1.78885	.80000	-5.0212	-.5788	-3.500	4	.025
Pair 9	S7_NO_PR - S7_GRD_T	-3.0000	1.00000	.44721	-4.2417	-1.7583	-6.708	4	.003
Pair 10	S7_RANDO - S7_SP_DR	.2000	2.68328	1.20000	-3.1317	3.5317	.167	4	.876
Pair 11	S7_RANDO - S7_GRD_T	.0000	1.00000	.44721	-1.2417	1.2417	.000	4	1.000
Pair 12	S7_SP_DR - S7_GRD_T	-.2000	1.92354	.86023	-2.5884	2.1884	-.232	4	.828

5.2.1 Discussion

Surprisingly, the results in TABLE 6 show that RANDOM prosody scores as high as the GROUND_TRUTH model for S7 video, and appears more realistic than SPEECH-DRIVEN for both videos. Further analysis of the head and eyebrow motion trajectories reveals differences in mean and variance between each model, as shown in TABLE 8. It is interesting to note that the standard deviation of the head motion is largest in the RANDOM model for both videos (S2 and S7), which might explain the results in TABLE 6, where RANDOM received 49 ballots compared to 41 for SPEECH_DRIVEN. Thus, it appears that the subjects used the amount of head movements (i.e., standard deviation) as a strategy to select the preferred facial animation, rather than coherence between these movements and the speech track. For this reason, a new perceptual experiment was designed in which the mean and standard deviation of all the animations was normalized to the same values.

TABLE 8

Statistics for video snippets used in experiment 1

Subject	Variation	Statistic	Inner eyebrow	Middle eyebrow	Head pitch	Head yaw	Head roll
S2	GROUND_TRUTH	st dev	5.54	2.75	2803.69	3482.31	1743.25
		mean	-2.38	-0.96	-916.94	5061.64	-1741.52
S2	RANDOM	st dev	11.60	11.60	3644.63	4431.30	3052.26
		mean	5.32	5.32	-961.85	-2260.03	-814.10
S2	SPEECH_DRIVEN	st dev	21.78	21.78	1441.66	1573.01	1039.17
		mean	15.10	15.10	-1544.10	4613.23	-2722.44
S7	GROUND_TRUTH	st dev	17.05	6.51	1563.21	1560.67	1589.00
		mean	15.45	2.82	491.78	-1843.58	-865.86
S7	RANDOM	st dev	13.47	13.47	3534.11	4276.10	3047.89
		mean	6.75	6.75	-1648.90	-1524.91	-92.81
S7	SPEECH_DRIVEN	st dev	17.18	17.18	666.93	1049.36	665.11
		mean	10.70	10.70	2034.31	-1823.35	-1281.07

5.3 Experiment 2

For the second experiment, new animations were generated by scaling and adding an appropriate offset to the head motion parameters so that the FAPs for the three conditions (RANDOM, SPEECH-DRIVEN and GROUND_TRUTH) contained the same standard deviation and mean statistics. Fourteen (14) surveys were conducted. Six of them were conducted showing the 16 pairs for video S7 first, followed by the 16 pairs for video S2, whereas the remaining eight surveys were conducted in the opposite order. All viewers were instructed to rate the animations based on the following question: "*Which of the animations displays head motion and eyebrow motion that is more coherent/consistent with the spoken segment?*" As in the previous experiment, the

audience was informed that lip motion was the same for all models, and that it was the original motion captured from video. The survey is summarized in TABLE 9 for S2 video and TABLE 10 for S7 video. The collapsed results are shown in TABLE 11.

TABLE 9

Confusion matrix for experiment 2 using video S2

S2 Video pair-wise comparison		Right panel animation			
		NO_PROSODY	RANDOM	SPEECH-DRIVEN	GROUD_TRUTH
Left panel animation	NO_PROSODY	BBABBBB ABABBBA	BBBBBBB BBBBBBB	BBBBBBB BBBBBBB	BBBBBBB BBBBBAB
	RANDOM	AAAAABA BAAAAAA	BBBABAB AAAAAAB	BBAABAB BBBBABB	BBBBABB BBABAAB
	SPEECH_DRIVEN	AAAAAAA AAAAAAA	AABBBBB BBAABBB	AAAAABB AAABBBA	BAABBAB ABABBBB
	GROUD_TRUTH	AAAAAAA AAAAAAA	BAABBAA AAAABAB	ABABAAA BAAAAAB	BAABBAA ABABAAB

TABLE 10

Confusion matrix for experiment 2 using video S7

S7 Video pair-wise comparison		Right panel animation			
		NO_PROSODY	RANDOM	SPEECH-DRIVEN	GROUD_TRUTH
Left panel animation	NO_PROSODY	BABAAAB ABBBAAB	BBAABBB BBBBBBA	BBBBBBB BBBBBBB	BBBBBBB BBBBBBB
	RANDOM	AAAAAAA AAAAAAA	BAABABA ABBABBA	BBBABBA AAABAAB	BBBBBBB ABBBBBB
	SPEECH_DRIVEN	AAAAAAA AAAAAAA	AAABABA AAAAABB	AAAAABA ABBABBB	BAABBBB BABBABA
	GROUD_TRUTH	AAAAAAA AAAAAAA	AAAAAAA AAAABBA	ABBBABA AABBAAA	ABBBBAB ABABABA

TABLE 11

Collapsed number of ballots for experiment 2

Model	**S2 Score**	**S7 Score**	**S2&S7 Score**
NO_PROSODY	17	23	40
RANDOM	63	53	116
SPEECH_DRIVEN	65	70	135
GROUND_TRUTH	79	78	157
		Total	448

These scores consistently show that GROUND_TRUTH motion is more coherent than the other models, followed by SPEECH_DRIVEN, RANDOM and NO_PROSODY. In addition, the Student's T-test, shown in TABLE 12, reveals that the differences between all pairs are statistically significant, with the sole exception of RANDOM vs. SPEECH_DRIVEN for the case of S2. In that case, SPEECH_DRIVEN rates just a little bit better than RANDOM. This could be explained by the fact that the model structure used to generate head and eyebrow motion was not optimized for each subject separately but was identical for both. Could better performance be obtained for S2 if the model structure was optimized (e.g., through cross-validation)?

TABLE 12

T-test pair-wise mean comparison for experiment 2

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of Mean Difference Lower	Upper			
Pair 1	S2_NO_PR - S2_RANDO	-3.2857	1.38278	.36956	-4.0841	-2.4873	-8.891	13	.000
Pair 2	S2_NO_PR - S2_SP_DR	-3.4286	1.15787	.30945	-4.0971	-2.7600	-11.079	13	.000
Pair 3	S2_NO_PR - S2_GRD_T	-4.4286	1.01635	.27163	-5.0154	-3.8417	-16.304	13	.000
Pair 4	S2_RANDO - S2_SP_DR	-.1429	2.14322	.57280	-1.3803	1.0946	-.249	13	.807
Pair 5	S2_RANDO - S2_GRD_T	-1.1429	1.74784	.46713	-2.1520	-.1337	-2.447	13	.029
Pair 6	S2_SP_DR - S2_GRD_T	-1.0000	1.51911	.40600	-1.8771	-.1229	-2.463	13	.029
Pair 7	S7_NO_PR - S7_RANDO	-2.5714	1.28388	.34313	-3.3127	-1.8301	-7.494	13	.000
Pair 8	S7_NO_PR - S7_SP_DR	-3.7857	1.12171	.29979	-4.4334	-3.1381	-12.628	13	.000
Pair 9	S7_NO_PR - S7_GRD_T	-4.7857	.89258	.23855	-5.3011	-4.2704	-20.061	13	.000
Pair 10	S7_RANDO - S7_SP_DR	-1.2143	2.04483	.54650	-2.3949	-.0336	-2.222	13	.045
Pair 11	S7_RANDO - S7_GRD_T	-2.2143	1.25137	.33444	-2.9368	-1.4918	-6.621	13	.000
Pair 12	S7_SP_DR - S7_GRD_T	-1.0000	1.56893	.41931	-1.9059	-.0941	-2.385	13	.033

5.3.1 Discussion

The results of experiment 2 thus are quite promising yet inconclusive. As shown by the analysis of S7 (cf. TABLE 11), visual prosody driven by acoustic utterances show improvements over random prosody. However, the fact that our hypothesis can not be confirmed for the S2 animations indicates that this relationship might be subject-dependent or that it is somewhat more complex than the one assumed in the speech-driven model used for the experiments.

Additionally, the higher rating given to SPEECH_DRIVEN vs. RANDOM in the case of normalized variance across models, reaffirms our belief that experiment 1 biased the

viewers to rate as more acceptable the RANDOM animation model due to exaggerated motion (i.e., standard deviation, cf. TABLE 8) This brings another discussion topic for consideration: to what extend can exaggeration in visual prosody used to generate perceptually more acceptable human characterizations?

6 CONCLUSIONS AND FUTURE WORK

Facial avatars are a promising technology for future multimedia human-computer and computer-mediated interaction. In order for facial animations to gain broad acceptance, they have to display accurate visual speech (lip and tongue movements) but also exploit background channels that we employ during face to face communication, including head, eyebrow and eye motion, as well as facial expressions and hand gestures. Research on the use of these non-verbal movements for facial animation is, however, hampered by the lack of an underlying language model.

This thesis has explored the use of two of these channels (head and eyebrow motion) to improve facial avatars. Our main hypothesis was that visual prosody driven by speech acoustics produces perceptually more realistic, coherent and convincing facial animations. Our work has encompassed all aspects of the system, from audio-visual data-acquisition to perceptual evaluation, from speech processing to computer vision. To achieve our goal we have:

- developed a complete motion capture system from the grounds up using off-the-shelf equipment under $1,000, and substantially engaged undergraduate engineering students in the design and implementation process.

- designed an experimental protocol to elicit visual prosody from naïve subjects. Inspired from techniques used in gesture research [67], subjects are presented with a short cartoon and subsequently asked to narrate the story.
- implemented two different computational models of visual prosody, the first one driven by Perlin noise, and second one driven by acoustic features of the speech signal.
- developed an interface to perform pair-wise perceptual evaluations of the animation stimuli, and performed statistical analysis of these experiments.

Our results are quite promising: using very simple computational models for the prediction of visual prosody from speech (e.g., rule-based heuristics and linear autoregressive models) as well as simple acoustic features (e.g., fundamental frequency and energy contours), we show that speech-driven facial prosody is perceptually comparable and in some cases superior to movements generated with Perlin noise. We expect that improved speech-driven performance may be obtained by tuning the model structure individually for each subject by means of a cross-validation stage, and also by using more powerful prediction models. In addition, we showed that exaggerated visual prosody can bias the viewer to perceive the avatar motion as more realistic.

There exist several important directions for future work. First, this work has been limited by the spatial (17 markers) and temporal (30 fps) resolution of the acquisition system, which is unable to capture subtle or fast facial phenomena. This calls for the use of high-

end motion capture equipment capable of tracking more facial markers (up to 100 in some cases) at high frame rates (200 fps). In addition, the use of multiple cameras may allow us to recover 3D position of these markers and avoid the use of the head-mounted frame. Second, improved prediction results may be obtained by using more powerful prediction models for the audio-visual prosody mapping. In particular, nearest neighbor and input-output Hidden Markov Models have been shown to work well for the prediction of lip motion [1]-[2]. Further prediction improvements may be achieved by extracting more informative features from the speech acoustics, such as shape-based descriptors of the F0 and energy contours, rhythm and speaking rates, and segmental features (e.g., syllable boundaries). Third, the perceptual evaluations explored in this work have been of a subjective character. More objective evaluations are required to assess the benefits of visual prosody in facial animation, such as improvements in speech intelligibility or task-related performance.

It has been proposed that supra-segmental speech features are closely related to the syntax and semantics of sentences [68], thus indicating that these features could *in some cases* serve as an indirect measurement with which to articulate semantically correct visual prosody. However, it is important to realize that not all visual prosody can be predicted from the utterances of the speaker. This includes movements related to more complex semantics or affective state, head movements associated with emblems (nodding or shaking for agreement/disagreement), or those associated with maintaining the flow of conversation (turn taking system), to mention a few [42]. In these case “data-

driven" visual prosody models, such as the ones explored in this thesis, may have to be complemented with those already explored in the context of conversational agents [69].

REFERENCES

[1] S. Fu, R. Gutierrez-Osuna, A. Esposito, K. Praveen and O. N. Garcia, "Audio/Visual Mapping with Cross-Modal Hidden Markov Models", in *Proc. of IEEE Transactions on Multimedia*, 2004 (in press).

[2] R. Gutierrez-Osuna, P. K. Kakumanu, A. Esposito, O. N. Garcia, A. Bojorquez, J. L. Castillo, and I. J. Rudomin, "Speech-driven Facial Animation with Realistic Dynamics," *IEEE Transactions on Multimedia*, 2004 (in press).

[3] R. Bruyer, *The Neuropsychology of Face Perception and Facial Expression.* Hillsdale, NJ: Erlbaum Associates, 1986.

[4] D. Archer and J. Silver, *The Human Face: Emotions, Identities and Masks*. Berkeley, CA: University of California Extension Center for Media and Independent Learning, 1996.

[5] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY: Oxford University Press, 1997.

[6] P. Ekman, *Darwin and Facial Expression: A Century of Research in Review*, New York, NY: Academic Press, 1973.

[7] C. Katchen, *Painting Faces and Figures*. New York, NY: Watson-Guptill Publications, 1986.

[8] C. Crawford and R. Rouse III, "Artists against Anatomists". *Computer Graphics*, vol. 36, no. 1, pp. 8-10, 2002.

[9] P. Ekman and W. Friesen. *Manual for the Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press Inc., 1978.

[10] F. Parke and K. Waters. *Computer Facial Animation*. Wellesley, MA: AK Peters, Ltd., 1996.

[11] F. Parke, *A Parametric Model for Human Faces*. Tech. Report UTEC-CSc-75-047, University of Utah, 1974.

[12] A. Pearce, B. Wyvill, G. Wyvill, and D. Hill, "Speech and Expression: A Computer Solution to Face Animation", in *Proc. Graphics Interface '86*, Canadian Information Processing Society, Calgary, 1986, pp. 136-140.

[13] S. DiPaola, "Extending the Range of Facial Types", *Visualization and Computer Animation*, vol. 2, no. 4, pp. 129-131, 1991.

[14] I. Essa and A. Pentland, *Coding, Analysis, Interpretation, and Recognition of Facial Expressions*, Cambridge, MA: The Media Laboratory, MIT, 1995.

[15] N. Magnenat-Thalmann, H. Minh, M. deAngelis, and D. Thalmann. "Design, Transformation and Animation of Human Faces", *The Visual Computer*, vol.5, no 1-2, pp. 32-39, 1989.

[16] I. Pandzic and R. Forchheimer. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. Chichester, Hoboken, NJ: J. Wiley & Sons, Ltd., 2002.

[17] Haptek Inc., Homepage [Online]. Available: www.haptek.com/, (accessed on Jan. 12, 2005).

[18] DA Group, Homepage [Online]. Available: www.digital-animations.com/, (accessed on Jan. 12, 2005).

[19] Face2face Animation Inc., Homepage [Online]. Available: www.f2f-inc.com/, (accessed on Jan. 12, 2005).

[20] Famous3D Pty. Ltd., Homepage [Online]. Available: www.famous3d.com/index.html, (accessed on Jan. 12, 2005).

[21] F. Lavagetto and R. Pockaj, "The Facial Animation Engine: Toward a High-Level Interface for the Design of MPEG-4 Compliant Animated Faces". *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 2, pp. 277-289, 1999.

[22] M. Rydfalk, *Candide, A Parameterized Face*. M.S. thesis, Department of Electrical Engineering, Linkoping University, Sweden, 1987.

[23] Oregon Health & Science University, Center for Spoken Language Understanding, CSLU Toolkit Download Page [Online]. Available: cslu.cse.ogi.edu/toolkit/, (accessed on Jan. 12, 2005).

[24] Helsinky University of Technology, Laboratory of Computational Engineering, Research Area: Artificial Person Homepage [Online]. Available: www.lce.hut.fi/research/cogntech/artiper.shtml, (accessed on Jan. 12, 2005).

[25] D. Terzopoulus and K. Waters. "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569-579, 1993.

[26] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active Contour Models", *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321-331, Marr Prize Special Issue, 1987.

[27] C. Bregler, M. Covell, and M. Slaney. "Video Rewrite: Driving Visual Speech with Audio", in *Proc. ACM SIGGRAPH*, 1997, pp. 353-360.

[28] T. Ezzat, G. Geiger, and T. Poggio. "Trainable Videorealistic Speech Animation", in *Proc. of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, 2002, pp. 57-64.

[29] E. Cosatto, *Sample-Based Talking-Head Synthesis*. Ph.D. dissertation, Lausanne, Switzerland: Swiss Federal Institute of Technology, Signal Processing Laboratory, 2002.

[30] Vicon Motion Systems Ltd., Homepage [Online]. Available: www.vicon.com/jsp/index.jsp, (accessed on Jan. 12, 2005).

[31] Qualisys Medical AB, Homepage [Online]. Available: www.qualisys.com/, (accessed on Jan. 12, 2005).

[32] S. Baker and T. Kanade, *Super Resolution Optical Flow*. Tech. report CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, 1999.

[33] H. Ishikawa and D. Geiger, "Occlusions, Discontinuities, and Epipolar Lines in Stereo", in *5th European Conference on Computer Vision*, pp. 232-248, Freiburg-Germany, 1998, pp. 232-248.

[34] F. Quek, X. Ma, and R. Bryll, "A Parallel Algorithm for Dynamic Gesture Tracking", in *International Workshop on Recognition, Analysis, and Tracking of*

Faces and Gestures in Real-Time Systems, ICCV'99, Kerkyra, Greece, 1999, pp. 64-69.

[35] M. Cohen and D. Massaro. "Modeling Coarticulation in Synthetic Visual Speech". In *Models and Techniques in Computer Animation*, N. M. Thalmann & D. Thalmann (eds.), pp. 139-156, Berlin: Springer Verlag, 1993.

[36] D. Masaro, J. Beskow, M. Cohen, C. Fry, and T. Rodriguez, "Picture My Voice: Audio to Visual Speech Synthesis Using Artificial Neural Networks", in *Proc. of Auditory-Visual Speech Processing (AVSP'99)*, Santa Cruz, CA, 1999, pp. 133-138.

[37] P. Kakumanu, R. Gutierrez-Osuna, A. Esposito, R. Bryll, A. Goshtasby, and O. Garcia. "Speech Driven Facial Animation", in *Proc. of the Workshop on Perceptive User Interfaces*, Orlando, FL, 2001, pp. 1-5.

[38] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the Relationship between Eyebrow Movements and F0 Variations", in *Proc. of the ICSLP*, Philadelphia, PA, 1996, pp. 2175-2179.

[39] M. Dohen, H. Lœvenbruck, M. Cathiard, and J. Schwartz. "Audiovisual Perception of Contrastive Focus in French". International Conference on Visual Speech Processing, St. Jorioz, France, 2003, pp.245-250.

[40] B. Granström, D. House, and M Swerts. "Multimodal Feedback Cues in Human-Machine Interactions", in *Proc. of the Speech Prosody 2002 Conference*, Aix-en-Provence, 2002, pp. 347-350.

[41] E Krahmer, Z. Ruttkay, M. Swerts, and W. Wesselink. "Perceptual Evaluation of Audiovisual Cues for Prominence", in *7th International Conference on Spoken Language Processing*. Denver, CO, 2002, pp. 1933-1936.

[42] K. Pelachaud, N. Badler, and M. Steedman. "Generating Facial Expressions for Speech". *Cognitive Science*, vol. 20, no. 1, pp. 1-46, 1996.

[43] K. Munhall and J. Buchan. "Something in the Way She Moves". *Trends in Cognitive Sciences*, vol. 8, no. 2, pp. 51-53, 2004.

[44] K. Munhall, J. Jones, D. Callahan, T. Kuratate E. and Vatikiotis-Bateson, "Visual Prosody and Speech Intelligibility", *Psychological Science*, vol. 15 no. 2 pp. 133-137, 2004.

[45] H. Graf, E. Cosatto, V. Strom, and F. Huang, "Visual Prosody: Facial Movements Accompanying Speech", in *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, Washington DC, 2002, pp. 381-386.

[46] Z. Deng, C. Busso, S. Narayanan, and U. Neuman. "Audio-based Head Motion Synthesis for Avatar-based Telepresence Systems", in *Proc. of ACM SIGMM Workshop on Effective Telepresence*, New York, NY, 2004, pp. 24-30.

[47] I. Albrecht, J. Haber, and H. Seidel. "Automatic Generation of Non-Verbal Facial Expressions from Speech", in *Proc. Computer Graphics International CGI'02*, City University of Hong Kong, Hong Kong, 2002, pp. 283-293.

[48] S. Lee, J. Badler, and N. Badler, "Eyes Alive", in *Proc. of the ACM SIGGRAPH 2002*, San Antonio, TX, 2002, pp. 637-644.

[49] Z. Deng, J. P. Lewis, and U. Neumann, "Practical Eye Movement Model Using Texture Synthesis", in *Proc. of the ACM SIGGRAPH 2003 Conference on Sketches & Applications*, San Diego, CA, 2003, pp. 1-1.

[50] P. Ekman, "About Brows: Emotional and Conversational Signals", in *Human Ethology: Claims and Limits of a New Discipline*, M. von Cranach, K. Foppa, W. Lepenies, D. Ploog (eds.) pp. 169-202, Cambridge: Cambridge University Press, 1979.

[51] K. Grammer, W. Schiefenhövel, M. Schleidt, B. Lorenz, and I. Eibl-Eibesfeldt, "Patterns on the Face: The Eyebrow Flash in Crosscultural Comparison". *Ethology* vol. 77, pp.279-299, 1988.

[52] J. Cosnier, "Les Gestes de la Question", In *La Question*, Kerbrat-Orecchioni (ed.), pp. 163-171. Lyon, France: Presses Universitaires de Lyon, 1991.

[53] David Koons, "PupilCam Construction Instructions". *IBM Research Report, RJ 10212 (95086)* July 11, 2001.

[54] S. Zhai, "What's in the Eyes for Attentive Input", *Communications of the ACM*, vol. 46, no. 3 pp.34-39, 2003.

[55] K. Jablonsky, *Real-Time Audiovisual Speech Capture and Motion Tracking for Speech-driven Facial Animation*. Honor's thesis, Department of Computer Science, Texas A&M University, College Station, 2003.

[56] A. Criminisi, I. Reid, and A. Zisserman, "A Plane Measuring Device", in *Proc. of the 8th British Machine Vision Conference, BMVC'97* [Online]. Available:

http://www.bmva.ac.uk/bmvc/1997/papers/057/planedev.html, (accessed on Jan. 12, 2005).

[57] *Overview of the MPEG-4 Standard*, ISO/IEC JTC1/SC29/WG11 [Online]. Available: http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm, (accessed on Jan. 12, 2005).

[58] M. Maltese, T. Pierce, W. Foster, S. Marcus Writers, C. Jones, I. Freleng, R. McKimson Directors. *Looney Tunes: Golden Collection*. Warner Home Video, Burbank CA, 2003.

[59] K. Perlin, "An Image Synthetizer", in *Proc. of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, San Franciso, CA, 1985, pp. 287-296.

[60] K. Perlin, "Real Time Responsive Animation with Personality", *Visualization and Computer Graphics, IEEE Transactions on*, vol. 1, no. 1, pp. 5-15, March 1995.

[61] H. Elias, Personal Webpage: Perlin Noise Page [Online]. Available: freespace.virgin.net/hugo.elias/models/m_perlin.htm, (accessed on Jan. 12, 2005).

[62] K. Perlin, Personal Webpage: Responsive Face Page [Online]. Available: mrl.nyu.edu/~perlin/facedemo/, (accessed on Jan. 12, 2005).

[63] K. Perlin and A. Goldberg, "Improv: A System for Scripting Interactive Actors in Virtual Worlds", in *Proc. of ACM SIGGRAPH '96*, New Orleans, LA, 1996, pp. 205-216.

[64] Universiteit van Amsterdam, Institute of Phonetics Sciences (IFA), PRAAT Download Page [Online]. Available: www.fon.hum.uva.nl/praat/, (accessed on Jan. 12, 2005).

[65] Mathworks Inc., Helpdesk Article: System Identification Toolbox, ARX Models [Online]. Available: www.mathworks.com/access/helpdesk/help/toolbox/ident/ch2gui22.html, (accessed on Jan. 12, 2005).

[66] The Code Project, Sofware Article Page: SendKeys in C++ [Online]. Available: www.codeproject.com/cpp/sendkeys_cpp_Article.asp, (accessed on Jan. 12, 2005).

[67] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL: The University of Chicago Press, 1992.

[68] M. Steedman, "Structure and Intonation". *Language*, vol. 67, pp. 260–296, 1991.

[69] 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.

VITA

Marco Enrique Zavala Chmelicka
Address: Egas N35-45 y Mañosca, Quito-Ecuador
Phone: (5932) 245-4680
Email: marco_zavala@hotmail.com

Education
M.S., Computer Science, Texas A&M University, May 2005
B.S., Electrical Engineering, Army Polytechnic School, January 1996

Work Experience
Texas A&M University, September 2003 to December 2004
Job: Graduate Assistant in the Sponsored Student Programs office
Maintained a software for tracking sponsored international students
Xerox, June 2000 to April 2002
Job: XBS Specialist (Xerox Business Services)
Led a team of System Analysts, managed installation of outsourcing projects
Microsoft, March 2000 to May 2002
Job: Market Accounts Manager
Supported software sales through resellers
Tecmoware, August 1996 to February 2000
Job: Technical Department Manager, Projects Engineer
Organized technical department resources, installed and configured Windows NT Servers, planned and implemented network projects for several customers
Occidental, December 1995 to July 1996
Job: Maintenance Department Operator
Compiled technical equipment data for TMS program, planned preventive maintenance activities

Computer Skills
PLC programming, Visual C++, Visual Basic, Matlab, Verilog, SAS, OpenGL, Intel server assembly, IBM routers & switches configuration

Certifications & Honors
Fulbright Fellow, 2002
Microsoft Certified Systems Engineer (MCSE), 1999
Professional Server Expert, IBM, 1999