A Facial Animation System for Expressive Audio-Visual Speech

Arunachalam Somasundaram
and
Rick Parent

Computer Science and Engineering Dept.
The Ohio State University

Chapter 1

Introduction

Realistic facial motion synthesis is one of the most fundamental problems in computer graphics and one of the most difficult. Facial motion is primarily composed of expressions, speech and motion associated with biological needs such as eating. Combining both expressions and speech, face forms a powerful and one of the most important medium for communication. Such a magnitude of importance associated with the face brings along with it an equal magnitude of difficulty in animating it on the computer.

1.1 Applications of Expressive Facial Speech Animation

Expressive facial speech animation has a lot of applications in various fields such as character animation for films and advertising, computer games, user-interface agents and avatars, video-teleconferencing, human-computer interaction, education and speech therapy.

Various characters in films such as *Woody* in *Toy Story 2*, *Shrek* in *Shrek 2*, the superheroes in *The Incredibles*, *Yoda* in *Star Wars:Episode II - Attack of the Clones* and *Gollum* in *The Lord of the Rings: The Return of the King* have enchanted the viewers as any real actor in a movie. The digital actors play very important roles in many movies and believable facial animation for these characters is imperative. Dead actors can be also be brought to life using the digital medium. New audio can be recorded to mimic their voices or their original recorded voices could be modified to reflect different emotions. Their faces can be brought to life with all their expressive characteristics.

Realistic facial animation in computer game characters is needed to enhance the

interaction of the human player with the characters. The character's facial movements have to respond automatically to the player's actions to augment the experience of the player in the virtual world.

As a medium of communication, speech animation, can be used in user-interface agents or avatars on the computer or the internet. A speaking face can list out the options that can be chosen or explain situations. An avatar can be used to read news on the web with expressions reflecting the nature of the news.

Human-computer interaction can be enhanced by facial speech animation. For example, in an artificial intelligence environment, the face can be used to speak out answers to the user's questions. Interesting digital characters can be used to engage children for learning purposes. These characters would attract attention and make learning enjoyable.

One of the problems with fluency disorders like stuttering is *speed stress* that involves speaking at a rate faster than comprehension rate of the brain. *Delayed Auditory Feedback* [26], which introduces a delay into the feedback speech signal, reduces speech stress. *Delayed Visual Feedback*, which introduces a delay into the visual (facial) feedback signal, can also be tried to reduce speech stress. The stuttering person can also be asked to mimic a slow speaking digital character by looking into a computer screen to reduce the speech rate which in turn can reduce speech stress. Lip reading can also be enhanced by practicing using the speaking face of a digital character.

1.2 Report Organization

Chapter 2 describes our facial model that is used to create the animations. A muscle based animation model is used to animate the facial skin. The muscle geometry is derived from the skin geometry. The muscles are interconnected; the activation of a driving muscle affects the motion of the connected muscles. The motion of the muscles can either be obtained by keyframing the driving muscles or the motion of all the muscles can be obtained from motion capture data. The motion of the eyes is controlled by procedural techniques. Teeth and tongue motions are keyframed based on speech poses and are interpolated by setting acceleration limits on their movements.

Facial motion capture provides us the capability of tracking complex facial motion in 3-D during expressive speech. Chapter 3 discusses our facial motion capture process and discusses techniques to track facial muscle-skin motion and overall head motion from motion capture data. Our technique to morph an existing muscle-skin geometric model to fit the motion capture marker data is also described in this chapter.

Chapter 4 discusses the method we employ, based on speech literature, to modify neutral speech audio to produce expressive speech audio. Auditory prosodic elements such as pitch, duration and intensity are modified based on user input such as nature of emotion, fluency level and speech segments' emphasis levels. We produce emotional speech for basic emotions such as happiness, sadness, anger and disgust. We also synthesize non-fluent speech with stuttering effects.

Chapter 2

Face Model

The human face is a very complex object to create and animate. The facial skin can deform in a variety of ways and building articulatory controls to move the facial skin is a challenging task. This chapter describes the 3D facial model that we use to create expressive speech animations. A muscle-based facial animation model is used to drive the animation of the skin. Section 2.1 describes the motivation behind using a muscle model and the process of modeling and animating the muscles. Section 2.2 describes the tongue, teeth and eye model.

2.1 Facial Muscles

2.1.1 Motivation for a Muscle Model

The human face is composed of muscles, bones, skin and subcutaneous tissue. There are several hundred muscles that are present in the human face which are used for a variety of tasks like chewing, speech and facial expressions. However there are only eleven muscles that are primarily involved in facial skin expressions [15]. Complex facial skin motion can be broken down into individual or a group of simpler muscle movements. It is intuitive to create various poses by moving corresponding muscles. For example, the smiling pose is caused by the activation of the muscle *Zygomatic Major*. This muscle is connected to the lip muscle *Orbicularis Oris* and pulls the corners of the lip to produce a smiling pose. Understanding these muscle movements will help in understanding the dynamics of facial motion during speech and expressions. This will help in building a better expressive speech co-articulation model. The muscle movements can also be recombined in a variety of ways to produce different expressive animation results.

2.1.2 Facial Muscle Modeling

The jaw and eleven key facial muscles for skin movement that cause the various expressions and speech poses are modeled based on Gray's [20], Faigin's [15] and Ekman's [14] work. Gray [20] illustrates the facial muscle anatomy as shown in Figure 2.1. Figure 2.2 shows the facial muscles and skin model in the neutral pose. Figure 2.3 and Figure 2.4 show the muscles and facial skin during muscle activation poses when each of these muscles and jaw are individually activated. Table 2.1 describes the actions caused by the various facial muscles. Most of the muscles are modeled on both the right and left side of the face individually whereas muscles such as the nasal bridge muscle *Procerus*, lip muscle *Orbicularis Oris* and the pouting muscle *Mentalis* are modeled alone around the center of the face. Both the inner and outer *Orbicularis Oris* muscles are modeled for controlling the inner part of the lips and the outer region around the lips. In reality, most of the modeled muscles are attached at one end to the bone and to the skin at the other end. However, the lip muscle *Orbicularis Oris* is attached to the skin and to a lot of other muscles and is free to move around. It can assume a wide variety of complex shapes and is affected by other expression muscles.

Table 2.1: Actions caused by the various facial muscles.

Muscle	**Action**
Inner Frontalis	Raises inner eyebrow
Outer Frontalis	Raises outer eyebrow
Procerus/Corrugator	Lowers the brow
Orbicularis Oculi	Squints or squeezes the eye
Levator Labii Superioris	Raises upper lip and wrinkles nose
Zygomatic Major	Raises the corners of the lip
Risorius/Platysma	Stretches the lip
Triangularis	Depresses the corners of the lip
Depressor Labii Inferioris	Depresses the lower lip
Mentalis	Raises the chin
Orbicularis Oris	Funnels, sucks or tightens the lip

The muscles and jaw are modeled using a polygonal surface that is derived from the skin surface according to user specification. The user can select points on the skin mesh and a polygonal surface is created using these points as its vertices. The muscle vertices are offset at a user specified distance from the skin along the local skin normal. This mesh is subdivided if higher resolution is desired to capture more subtle motion and the new vertex positions are calculated again as an offset from the

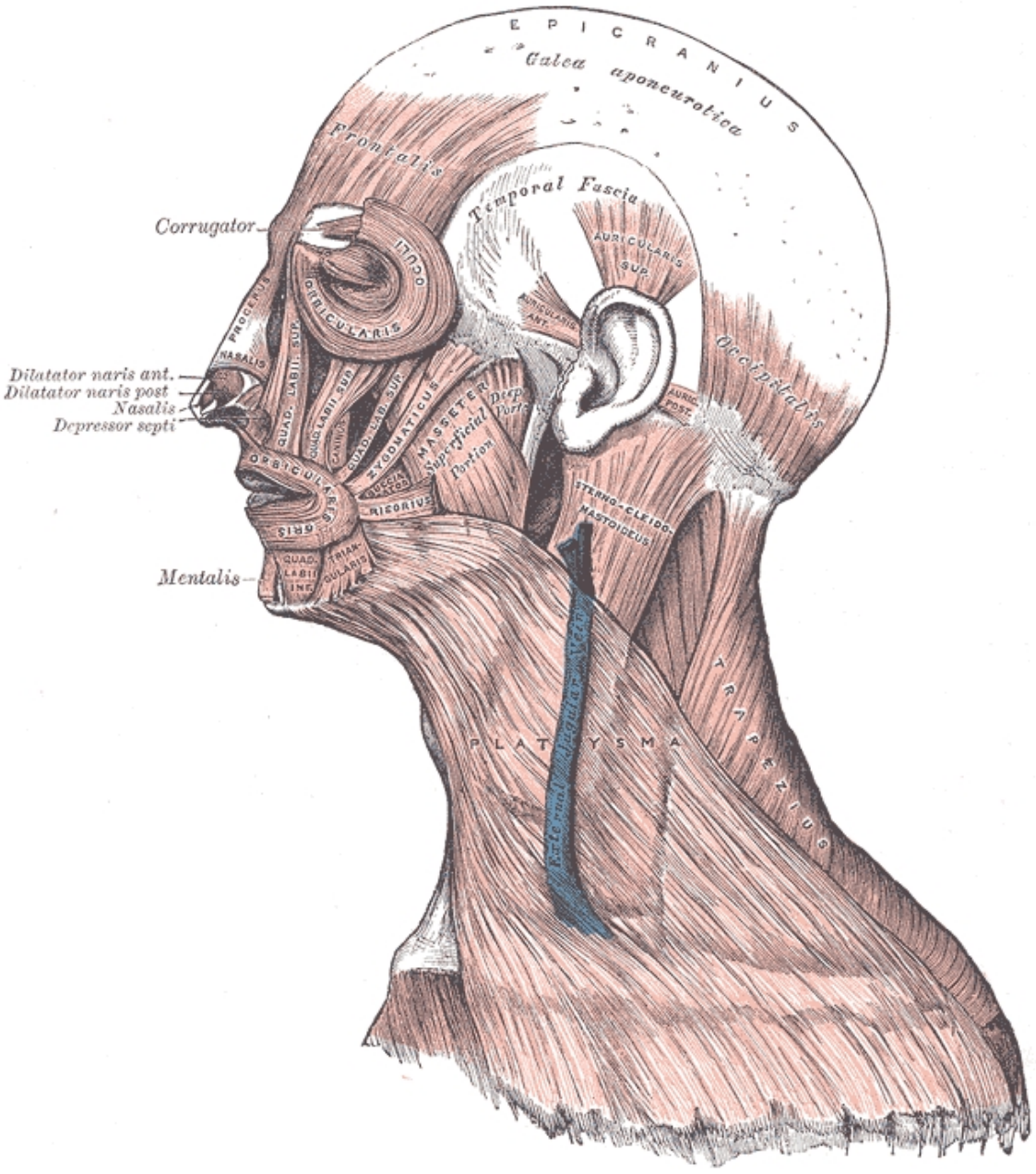


Figure 2.1: Facial muscle anatomy from the book *Anatomy of the Human Body* [20]

skin surface. The muscle thus takes the shape of the local skin surface. The muscle shape is fine tuned by hand.

Figure 2.2: Facial Muscles and Skin Model

2.1.3 Facial Muscle Animation

The movement of the muscles drive the movement of the skin. The movement of the muscles themselves can be acquired through motion capture which is described in Chapter 3. The muscle movements can also be described via muscle activations of the driving muscles which move the attached muscles. The technique for animation via muscle activations is described below.

The various modeled muscles are inter-connected and and the attachments between them are specified by the user. For example, the *Orbicularis Oris* is attached to all the muscles surrounding it like the lower lip stretcher *Risorius*, pouting muscle *Mentalis*, lower lip puller *Depressor Labii Inferioris*, smiling muscle *Zygomatic Major*, saddening muscle *Triangularis* and sneering muscle *Levator Labii Superioris*. Each muscle has a linear fall off influence region that affects the movement of the connected muscle vertices falling within that region. The user can activate the driving muscles by keyframing their movements including bulging when needed. The total number of vertices of all muscles is far lesser, by an order of of magnitude, than the total number of corresponding skin vertices in our examples and so driving muscle vertices can be keyframed faster than keyframing the corresponding skin vertices. The movement of the activated muscle is propagated to other connected muscles automatically through repeated steps. First, the muscles M_c connected to the driving muscles M_d

are moved. Then, the muscles connected to the muscles M_c are moved and so on until all the required muscles are moved. Each vertex of the muscle that is to be moved is influenced by the closest muscle that has been previously moved. Once a muscle is moved at one step, it is marked as moved and is not moved again at a later step. This avoids looping and in reality the driven muscles move in order of connectivity from the activated muscle. Care must be taken when simultaneously activating more than one muscle that have an overlapping influence region to avoid abnormal muscle distortions. The jaw is just modeled as another rigid muscle that can only be rotated. Various muscles below the lower lip attach into the jaw and are pulled by the rotation of the jaw.

The skin vertices are attached to the muscles and their movements are driven by the muscle movements. A polygonal skin model was obtained from *FaceGen* [23] software. Each skin vertex V_s is mapped to a triangle T_m of a muscle by cylindrical projection or proximity if no such projection exists. In the case of a successful projection, the barycentric coordinates $Coord_{bc}$ of the projected point V_p on the muscle triangle T_m is stored. Otherwise, the barycentric coordinates $Coord_{bc}$ of the point V_c on the triangle T_m that is closest to the skin vertex V_s is stored. Care is taken around the lip region to see to that there is no erroneous mapping between top lip muscle and bottom lip skin and vice versa. The top and bottom lip skin vertices are marked by hand and are not allowed to get mapped erroneously. Once the muscle-skin attachment is determined, each skin vertex V_s assumes the movement of the corresponding point V_p (or V_c) on the attached muscle triangle T_m. Muscles have a linear fall off influence region that affects the skin vertices. The further the skin vertex V_s is from the muscle attachment point V_p (or V_c), the lesser it is affected by the attached muscle.

Animating the facial skin via underlying muscles is intuitive and muscle movements can be combined logically to produce meaningful expressions and speech. Muscles provide us with a way to breakdown complex facial motion into individual components that, when combined in different ways, can produce a variety of complex facial motions.

2.2 Eyes, Teeth and Tongue

Along with the motion of the facial skin, the movements of the eye, tongue and teeth play an important part in the visual display of speech and expressions. Polygonal spheres are used to model the eyes. Tongue, teeth and inner mouth polygon models are acquired from *FaceGen* [23] software. The lower teeth, which is attached to the jaw, can be rotated. Various tongue poses corresponding to the different Visemes are created by manually moving the vertices of the tongue model from the initial default

pose. The inner mouth model represents the roof palate and visible parts of the inner cheek when the mouth is open. Figure 2.5 shows the eyes and various parts of the mouth.

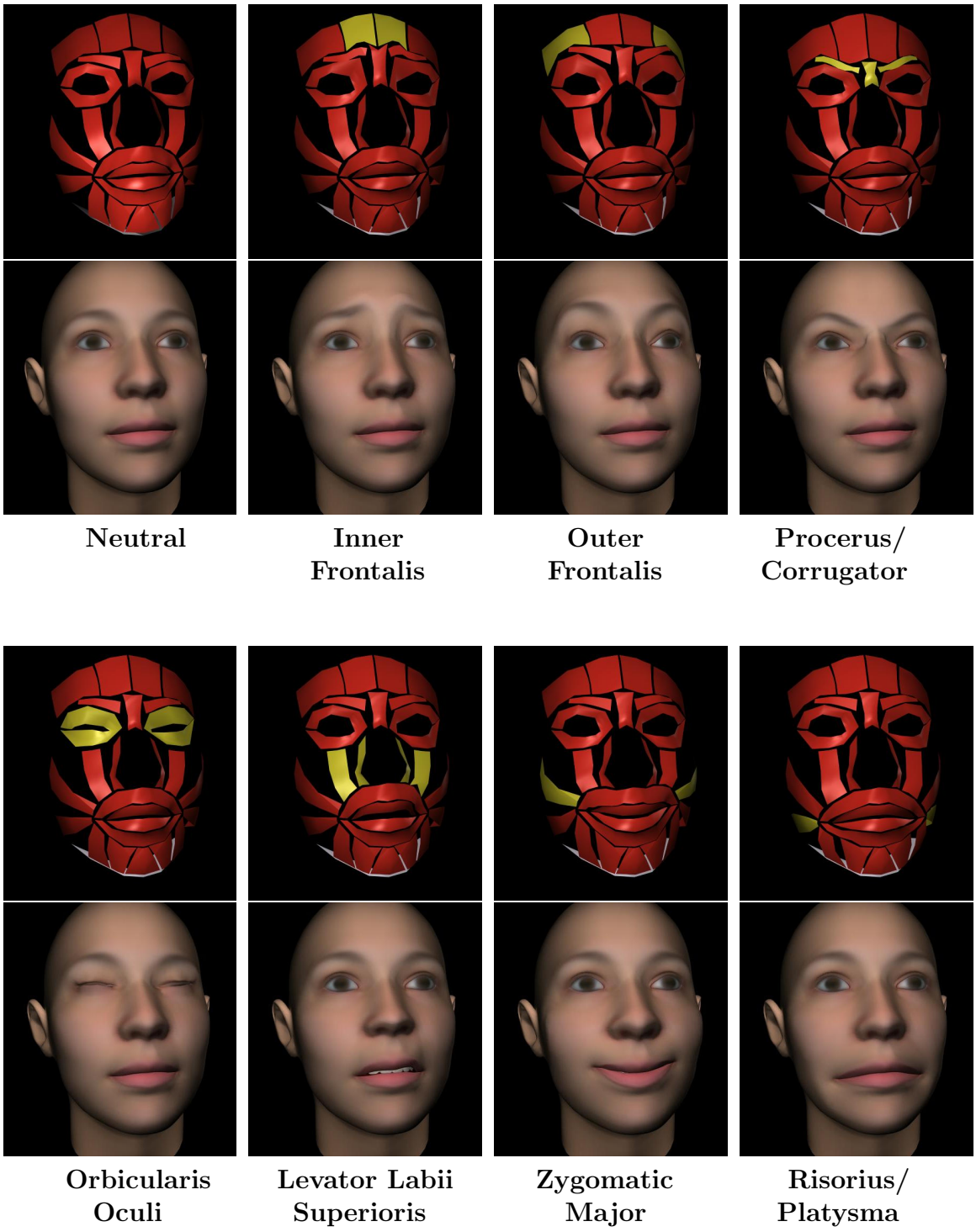
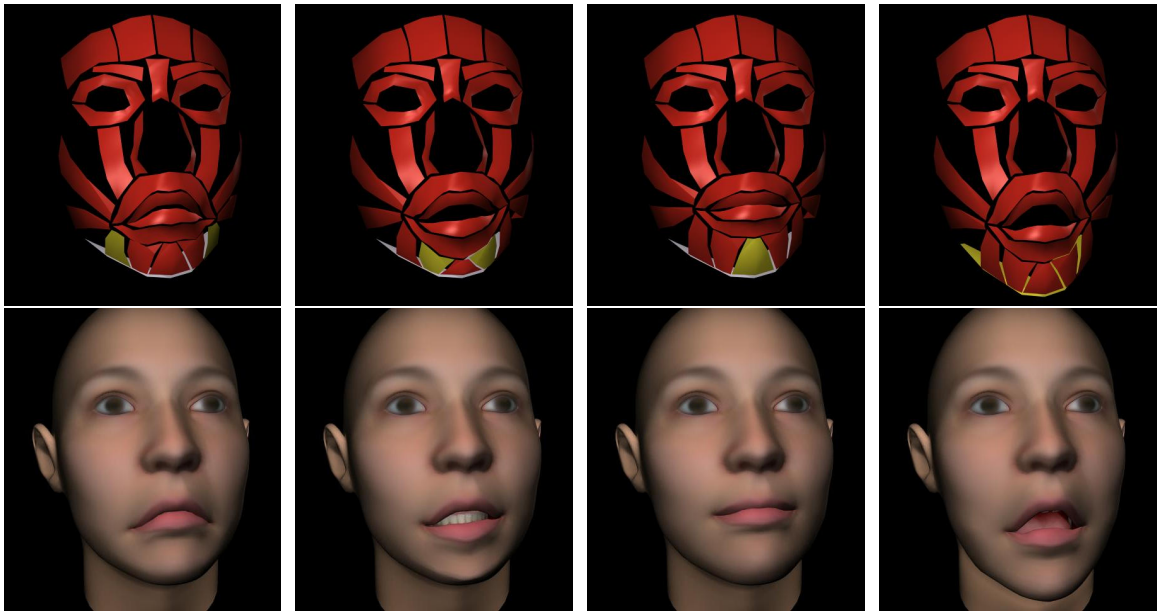


Figure 2.3: Face poses when each of the muscles and jaw are individually activated. Whenever posssible, corresponding muscles on both sides of the face have been activated. Activated Muscles are shown in yellow color.



Triangularis **Depressor Labii Inferioris** **Mentalis** **Jaw**

Figure 2.4: Face poses when each of the muscles and jaw are individually activated. Whenever posssible, corresponding muscles on both sides of the face have been activated. Activated Muscles are shown in yellow color.

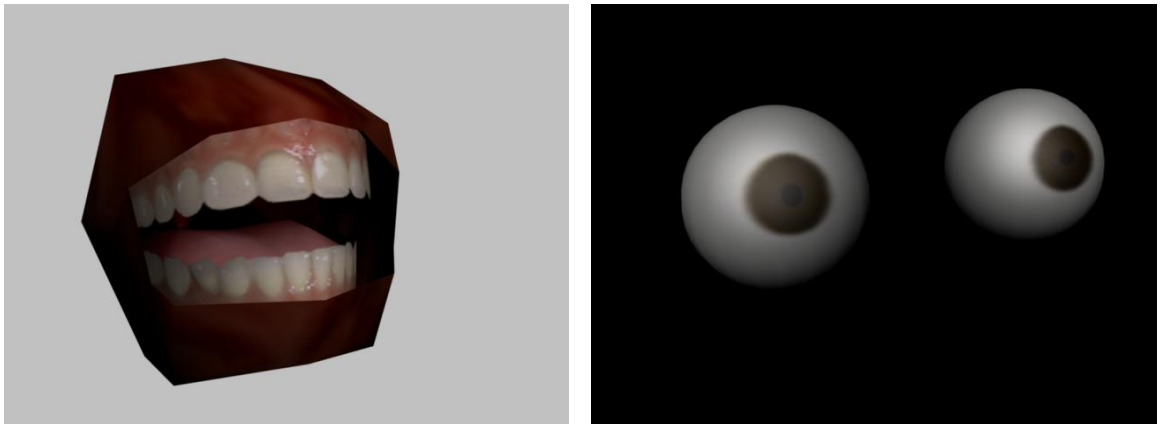


Figure 2.5: Eyes, Tongue, Teeth and Inner Mouth Models

Chapter 3

Facial Motion Capture

Facial skin motion during expressive speech is very complex. Motion capture technology using sparse markers, which are placed on the skin, allows the capture of skin motion in 3-D. Capturing expressive visual speech in 3-D gives us a wealth of data on facial poses and facial motion during speech under the influence of the different emotions. Moreover, overall head motion can also be tracked using markers placed on a rigid headband.

In this chapter, the techniques for animating the facial muscles and animating the head as a whole from motion capture marker data are described. Section 3.1 describes our motion capture process using sparse markers. The techniques for cleaning the marker data and extracting overall head motion are described in Section 3.2. The process of tracking muscle movements from marker data is described in Section 3.3.

3.1 Motion Capture Process

The facial motion capture was performed at the Motion Capture Lab [3] at The Advanced Computing Center for the Arts and Design (ACCAD). Ninety-nine small hemispherical reflective markers (about 3 mm in diameter) were placed on the facial skin of each speaker. Markers were also placed on the lip of the speaker to capture the complex lip motion. Seven more markers were placed on a rigid head band worn by the speaker. Figure 3.1 shows the face of a speaker with markers on the skin and headband. Vicon Optical Motion Capture System [7] using 12 cameras was used to capture the marker motion at 120 fps and reconstruct them in 3-D.

We captured the 3-D facial skin motion, video and audio of six different people speaking different sentences and speaking fourteen visemes (*/A/, /Ah/, /Aah/, /E-K/, /Oh/, /Ooh-Q/, /B-M-P/, /D-T/, /Ch-J-Sh/, /F-V/, /L-N/, /R/, /S/* and

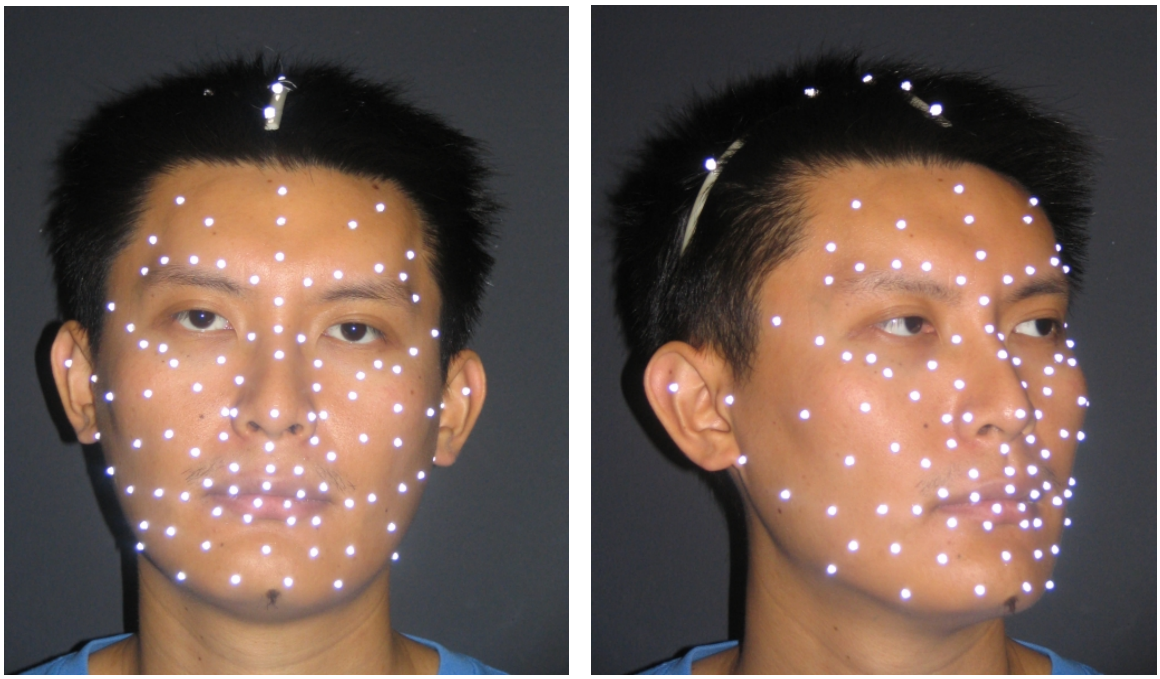


Figure 3.1: Front and side view of motion capture markers on skin and headband

/Th/). The visemes and sentences were captured for each of the six different basic emotions namely, happiness, sadness, disgust, anger, surprise and fear. We also captured the sentences asking the speakers to act out stuttering. Facial poses corresponding to the different Facial Action Units (FACS) [14] were also captured.

3.2 Mocap Data Processing

Once the data is captured, the data is cleaned and segmented. There are several issues involving the cleaning of facial motion capture data that are described below.

The actual number of markers that were used for the capture is 106. However, the number of marker trajectories identified by the mocap system is as high as 910. This is due to missing markers in some frames and incorrectly tracked markers. Once the system loses track of a marker at a particular frame, a new marker is initialized for the same physical marker causing a huge number of broken trajectories.

The first step to clean the data is to track corresponding markers from a reference frame and reduce the number of trajectories to 106. There were several overlap markers (as many as 200) that were captured. Overlap markers are those captured markers that correspond to the same physical marker at the same frame. Incorrect

3-D reconstruction causes extra markers to appear which lie very close to each other (less than 1 mm). The trajectories of overlap markers were identified based on their proximity and merged into one. One reference frame containing all the markers is identified and the markers are identified and labeled by hand. Global head motion is tracked and removed from the captured data to facilitate cleaning as described in Section 3.2.1. The markers for the rest of the frames are automatically tracked based on the proximity of the corresponding markers from the nearby frame. For frames that are before and after the reference frame, the markers are tracked backwards and forwards respectively from the reference frame. Some of the markers are not captured for certain frames. A simple linear interpolation of those marker positions is applied to those frames.

Some of the 3-D reconstructed motion capture markers exhibit small random high frequency jitters (about 3 mm every 30 frames). Those jitters are smoothed out using a N-frame averaging filter (default value N = 7) centered around each frame. The value for each frame of the averaging filter is $1.0/N$.

The automatic cleaning techniques significantly reduce the problem of cleaning. However a few markers (about 2-3) are still incorrectly tracked for a few frames (about a maximum of 20 frames). The markers are either swapped with other markers or the linear interpolation is not sufficient. These markers are manually corrected. Once the data is cleaned, the head motion is added back. C++ plugins were written in Maya [22] for cleaning the motion capture data.

The cleaned data is then segmented and stored. Facial poses of the various expressive visemes and individual Facial Action Units (FACS) [14] are stored. Each expressive viseme contains the pose of the entire face. During neutral and emotional speech, different head motions such as head nods, head sways, idle motion etc are extracted.

3.2.1 Tracking Global Head Motion

Global head motion during emotional speech adds to the realism of visual speech. Tracking global head motion helps us to identify different kinds of head motions associated with the emphasis and emotion during speech. Tracking the head motion also facilitates cleaning of motion capture data and helps in analyzing local facial skin motion. The technique for tracking head motion using the rigid head band is described below.

The head band is automatically located using the highest marker positions at a reference frame and using the rigid nature of the head band. Head translations and rotations are calculated for the captured motion by tracking the head band.

Three headband markers, namely, the markers on the extreme front, extreme left

and extreme right are chosen for the tracking. A rigid triangle is created with the positions of these markers as its vertices. In some cases, when headband markers were captured incorrectly, three facial skin markers, namely, the markers on the top middle of the forehead and one on top of each ear were chosen. These three skin markers show very little or no local motion and are considered to form a rigid triangle. Head motion can be tracked by tracking the motion of the rigid triangle. Figure 3.2 illustrates the process of tracking the rigid triangle. The steps for tracking the rigid triangle are listed below.

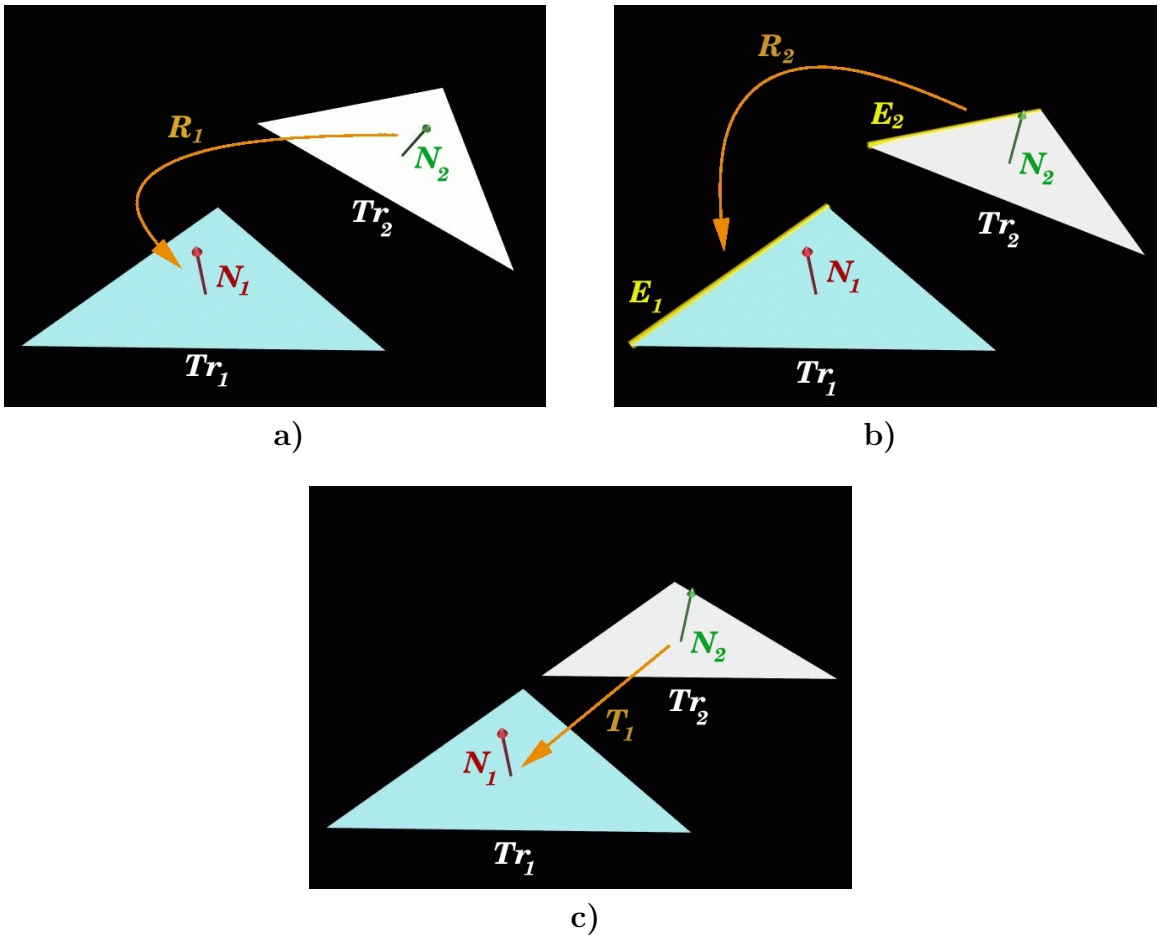


Figure 3.2: Process of calculating transformations to map triangle Tr_2 to triangle Tr_1. **a)** Calculate rotation R_1 to match normal N_2 to normal N_1. **b)** Calculate rotation R_2 to match edge E_2 to edge E_1. **c)** Calculate translation T_1 so that the vertices of triangle Tr_2 match the corresponding vertices of triangle Tr_1.

- Let Tr_2 be the rigid triangle from a reference neutral pose frame with neglible head transformations. Let Tr_1 be the corresponding rigid triangle from a frame of motion capture. To track the head motion, the rotation R_ang and translation T_1 that map the vertices of triangle Tr_2 to triangle Tr_1 are to be calculated. Let N_1 and N_2 be the normals of triangle Tr_1 and Tr_2 respectively. Let V_1 and V_2 be the vertices of triangle Tr_1 and Tr_2 respectively.
- Calculate rotation R_1 so that N_2 matches N_1 as given below
  - Calculate quaternion Q that represents the rotation between N_2 and N_1 as given in Equations 3.1 and 3.2

$$\begin{aligned} (3.1)\quad (x,y,z) &= normalize(N_1 \times N_2) \\ Q &= \begin{pmatrix} x*sin(\theta/2) & y*sin(\theta/2) & z*sin(\theta/2) & cos(\theta/2) \end{pmatrix} \\ (3.2)\quad &= \begin{pmatrix} A & B & C & D \end{pmatrix} \end{aligned}$$

  - Calculate the rotation matrix R_1 from the quarternion Q as described in Equation 3.3

$$(3.3)\quad R_1 = \begin{bmatrix} 1-2B^2-2C^2 & 2AB+2CD & 2AC-2BD \\ 2AB-2CD & 1-2A^2-2C^2 & 2BC+2AD \\ 2AC+2BD & 2BC-2AD & 1-2A^2-2B^2 \end{bmatrix}$$

- $V_2' = R_1 * V_2$. Now V_2' are the updated vertices of triangle Tr_2
- Calculate Rotation R_2 so that any one of the edges E_2 on triangle Tr_2 matches the corresponding edge E_1 on triangle Tr_1. This procedure is similar to the calculation of R_1 with normalized vectors pointing in the direction of the edges.
- $V_2'' = R_2 * V_2'$. Now V_2'' are the updated vertices of triangle Tr_2
- Calculate the combined rotation matrix R_c as $R_c = R_2 * R_1$. The Euler angles can be calculated from R_c as described in Equation 3.4.

$$R_c = \begin{bmatrix} R_{00} & R_{01} & R_{02} \\ R_{10} & R_{11} & R_{12} \\ R_{20} & R_{21} & R_{22} \end{bmatrix}$$

$$(3.4)\quad R_{ang}(R_x,R_y,R_z) = \begin{pmatrix} tan^{-1}(R_{21}/R_{22}) & -sin^{-1}(R_{20}) & tan^{-1}(R_{10}/R_{00}) \end{pmatrix}$$

- Calculate translation T_1 so that V_2'' matches V_1

The rotations and translations of the head for all the motion capture frames can be calculated as described above.

3.3 Tracking Skin and Muscle Motion from Mocap data

Muscles of the human face move in a complex fashion based on their own activation, their attachments to other muscles, skin and jaw and also due to collisions with adjacent bone structure. Facial motion capture data can be used to extract those complex muscle movements as described in this section.

Once the mocap marker data is cleaned, the motion is transferred to the facial muscle and skin model. Two issues arise in the motion transference process.

- The existing facial skin and muscle structure of the graphical geometric model are different from the actual skin and muscle structure of the captured speaker. So, the original skin and muscle models have to be morphed to fit the motion capture markers.

- Muscle motion has to be tracked from sparse motion capture marker data. Once the muscles are tracked the skin can be moved by techniques described in Chapter 2.

The technique for morphing the existing facial skin and muscle model to fit the captured markers is described below.

Let F_{source} be the source face mesh and M_{source} be the source muscle meshes. Let F_{target} and M_{target} be the corresponding target face and muscle meshes respectively that have be calculated. Figure 3.3 shows the source and target skin/muscle mesh models. Let C_{target} be the captured markers.

- A triangulated mesh T_{target} from the markers C_{target} is created manually as shown in Figure 3.4

- Corresponding to C_{target}, virtual markers C_{source} are placed on the mesh F_{source} and a triangulated mesh T_{source} from those markers is acquired. as shown in Figure 3.4

- The vertices of M_{source} are mapped to the triangles of T_{source} by cylindrical projection. For each vertex of M_{source}, the barycentric coordinates of projection $Coord_{bc}$ and the mapped triangle Tri_{hit} are recorded.

- The vertices of M_{target} are calculated from the triangles of T_{target} using the $Coord_{bc}$ and Tri_{hit} of the corresponding vertices in M_{source} as described below.

- Let V_T^n be the position of the nth vertex of a triangle. Let Mx_{source} be the matrix with rows V_T^1, V_T^2, V_T^3 of triangle Tri_{hit} in T_{source}. Correspondingly, let Mx_{target} be the matrix with rows V_T^1, V_T^2, V_T^3 of the corresponding triangle in the T_{target}.
- $\Delta M_{ts} = Mx_{target} - Mx_{source}$.
- Let Vm_{target} be the position of the vertex in the target muscle mesh that is to be calculated. Let Vm_{source} be the position of the corresponding vertex in the source muscle mesh.
- $Vm_{target} = Vm_{source} + Coord_{bc} * \Delta M_{ts}$.

The same procedure is applied to morph the existing facial skin model vertices to the motion capture data to obtain F_{target}. However vertices of F_{source} that could only be mapped to triangles in T_{source} were retained for F_{target}.

The muscle motion can be tracked from motion capture markers for all the captured frames using the same procedure. M_{target} at any given motion capture frame is calculated from the corresponding T_{target} mesh using the recorded $Coord_{bc}$ and Tri_{hit}.These muscle movements are then used to drive the motion of the facial skin mesh as described in Chapter 2

Figure 3.5 illustrates the entire process of transferring marker motion to skin motion via muscles.

3.4 Summary

Facial motion capture using sparse markers provides us with the capability of tracking the complex facial motion in 3-D during expressive speech. Facial Action Units (FACS) [14], emotional visemes and six sentences of emotional speech under the influence of each of the six basic emotions were all captured, cleaned and segmented. An emotional viseme captures the pose of the entire face.

Facial skin and muscle models for the captured markers are created by morphing the existing original models to fit the captured markers. The captured marker data is used to extract the facial muscle motion. The extracted muscle motion drives the movement of the facial skin as described in Chapter 2. Overall head motions are also extracted from the captured data and stored. The captured and segmented data is then used to synthesize new emotional speech.

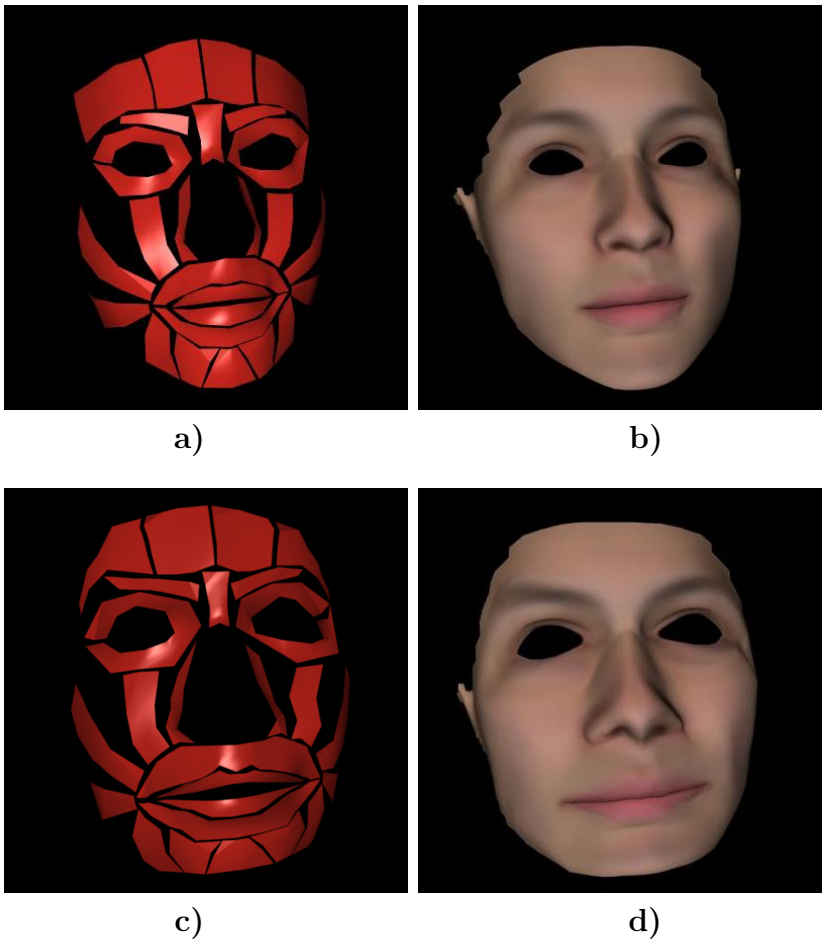


Figure 3.3: Original and morphed skin/muscle models. a) Source (original) muscle model M_{source}. b) Source (original) face model F_{source}. c) Target (morphed) muscle model M_{target}. d) Target(morphed) face model F_{target}

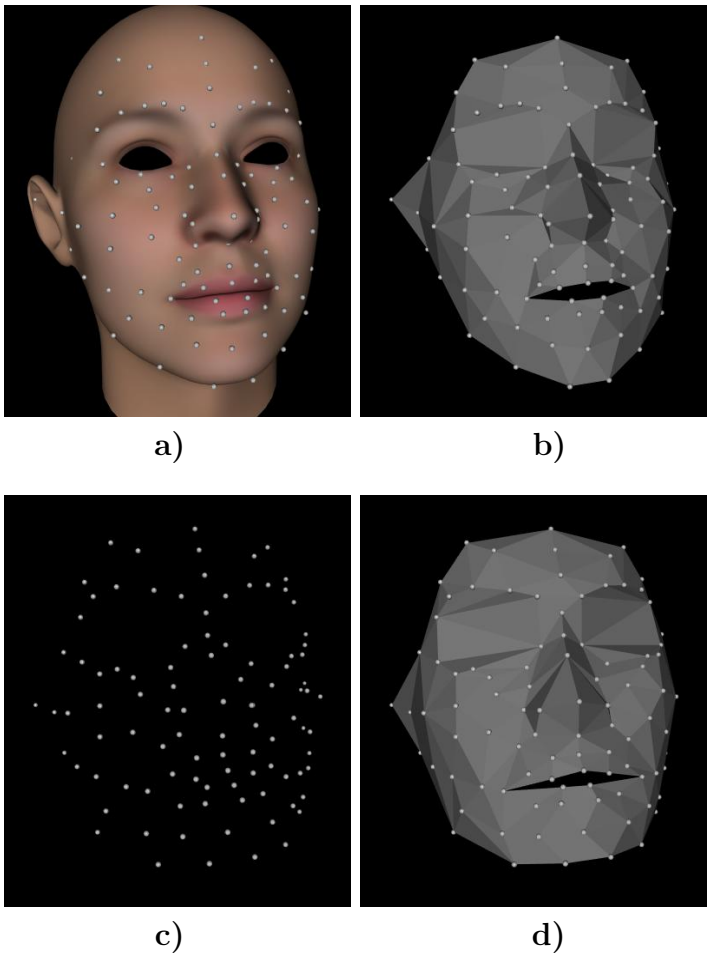


Figure 3.4: Mocap markers and their triangulation. a) Source model marked with virtual markers C_{source}. b) Triangulated mesh T_{source} of C_{source} markers. c) Captured markers C_{target}. b) Triangulated mesh T_{target} of C_{target} markers.

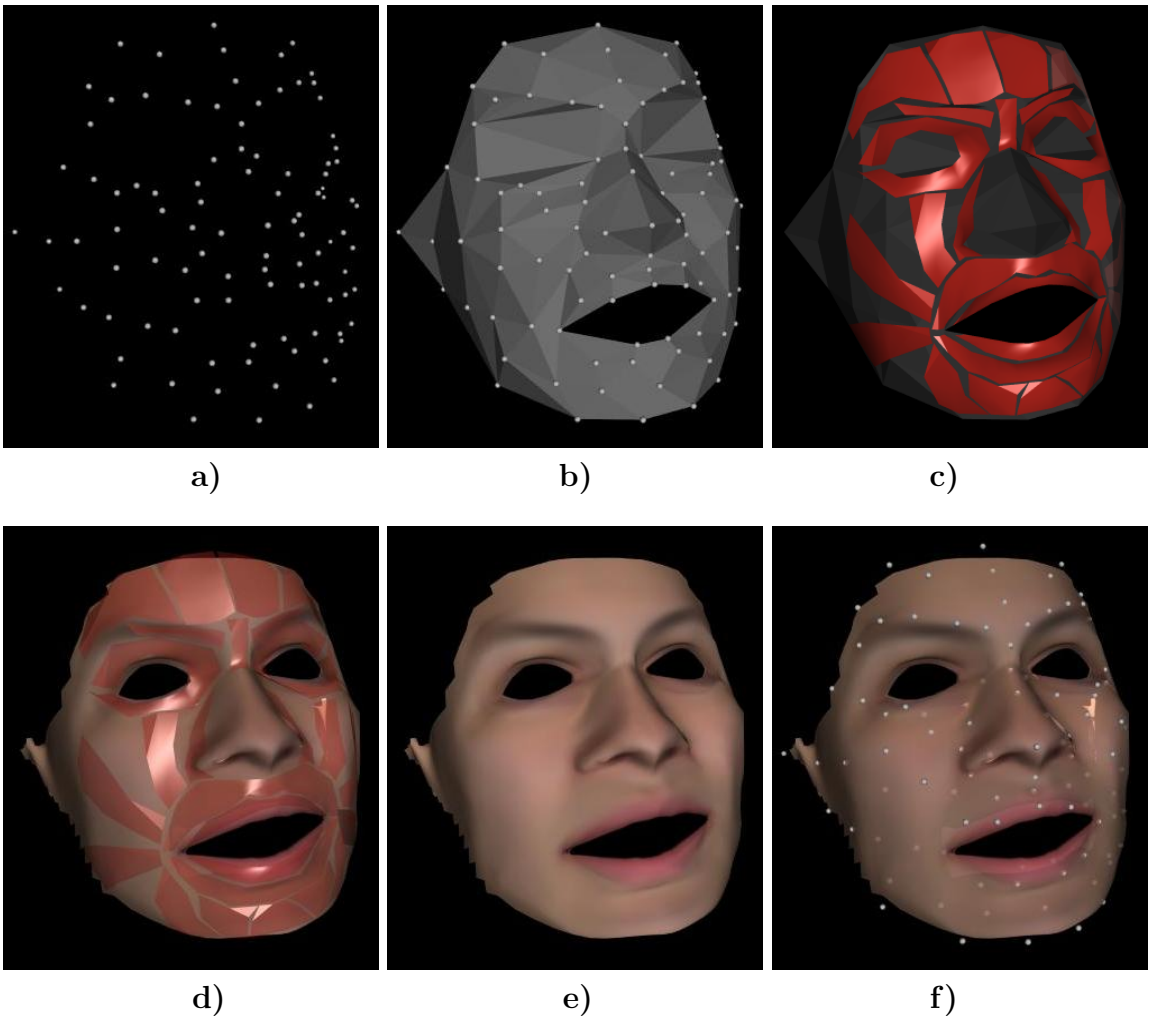


Figure 3.5: Motion transfer process from markers to facial skin a) Captured Markers. b) Triangulated mesh from markers. c) Muscles from triangulated mesh. d) Skin pose from muscle pose. e) Facial Skin. f) Skin pose with overlaid markers.

Chapter 4

Expressive Audio

The expressive qualities of human auditory speech gives it great potential to convey information effectively. Great orators, who inspire huge crowds, utilize the expressive power of the human voice. The emotion, mood, motive and nature of the speaker affect the audio of the spoken sentences in significant ways. Even intrinsically, without conscious effort, some of these factors are reflected in the audio. The same sentence, which is composed of the same sequence of words, can even imply different meanings based on the manner in which it is spoken.

In this chapter, expressive audio generation techniques are discussed. Synthesizing speech [10,13] including emotion or expression generation [21,12] is one way to produce speech. However, since human speech is more natural when compared to synthesized speech we have chosen an approach to modify recorded neutral speech to produce a variety of expressive speech samples. We consider modifying recorded speech based on the emotions such as happiness, sadness, anger, disgust and also based on the fluency of the speaker generating stuttering effects. Section 4.1 describes the auditory prosodic parameters that are modified. Section 4.2 and Section 4.3 discuss the changes that occur in speech during various emotions and during non-fluent speech. Audio modifications based on the expressive parameters are described in Section 4.4. Section 4.5 summarizes this chapter.

4.1 Prosody

Auditory prosody refers to the pitch, duration and intensity in speech that convey non-lexical information. The prosodic features are also called suprasegmental features because they can affect all the segments of a unit of speech (syllable, word or phrase). Prosody plays an important part in perceiving the meaning of the spoken sentence

[27]. Prosodic changes are also reflected in various emotions [12, 17, 16, 30]. The auditory prosodic features pitch, duration and intensity are described below.

The *pitch* of the speech is the fundamental frequency (f_0) at which the vocal folds vibrate. The emphasized harmonic frequencies are called *formants*; about 5 formants (f_1 to f_5) are generally needed for phoneme recognition. *Intonation* refers to the changes in pitch that contribute to the meaning of the spoken sentences.

Duration is the term used to refer to the length of speech segments such as phonemes or syllables. Speech rate and stress placement affect the duration of the segments. The combination of phonation and pause durations along with stress placement determines the *rhythm* of speech [12].

Intensity or *amplitude* of the speech is perceived as loudness. It is the relative amplitude, and not absolute amplitude, that is important in the perception of speech. The same sound heard at a lower volume still maintains the general characteristics of speech.

4.2 Emotional Audio

Emotion changes the auditory prosodic elements such as duration, pitch and intensity. Juslin and Laukka [25] provide a summary of 104 studies on the vocal expression of emotion. Banse and Scherer [9] provide the predicted changes in the auditory prosodic elements due to emotional effects by studying recorded emotional audio. Fairbanks and Pronovost [17] study the effect of pitch by having actors simulate emotions such as anger, fear, sadness and contempt. Fairbanks and Hoaglin [16] studied the effects of emotion on speech rhythm which included speech rate and pauses. The effects of emotions on speech prosodic parameters was analyzed by William and Stevens [30]. Below is a summary of some of those prosodic changes that happen during emotional speech when compared to neutral speech.

- **Sadness:** Speech rate is slower, loudness decreases, pitch is low, pitch transitions are slow and words are not stressed that much with long fluent pauses.
- **Anger:** Speech rate is faster, loudness increases, pitch is high, pitch transitions are fast and words are very much stressed with very few pauses.
- **Fear:** Speech rate is a slower than in anger but faster than in sadness, pitch is highest, pitch transitions are fast and there are very few pauses.
- **Disgust:** Speech rate is slow, pitch is low, pitch transitions are extreme and there are few pauses.

We captured the audio of six different people speaking different sentences. The audio captured for each of the six different basic emotions namely, happiness, sadness, disgust, anger, surprise and fear. We observed similar prosodic changes in the captured emotional audio. Also, during extremely sad speech, when on the verge of crying, the pitch of the voice increased. Speech rate was fast and pitch increased during happy speech.

Gay [18, 19] noticed that when speech rate is increased, vowel durations tend to be proportionately more compressed than consonant durations. When a word is emphasized, durations of vowels tend to be expanded larger than consonants in many cases. However, the durations of any of the phonemes that can be produced continously over a period a time can also be expanded while stressing. For example, consonants like /m/, /l/, /n/, /s/ and /f/ can also be stressed by expanding their duration. These consonants will be addressed as *stressable consonants* in this chapter.

The various changes brought in speech due to the addition of emotions makes the audio more expressive and interesting. The modifications that we make to the recorded neutral audio to produce emotional audio are described in Section 4.4.2.

4.3 Fluent and Nonfluent Audio

Fluent speech has at least three components that converge to give listeners the impression of fluency [6]. First, fluent speech is continuous or smooth without hesitations or without stopping unexpectedly. Second, rate of information flow is about 150-170 words per minute for fluent speech. The third component is the effort of the speaker. Fluent speech looks easy and effortless both physically and mentally.

Nonfluent speech is associated with increased pauses, interjections and stuttering. Pauses and interjections may be introduced due to a variety of reasons. Some of those include time for reflections or searching over the next words, mood of the person and inability to speak coherently.

Stuttering is more of a speech disorder and is a problem associated with the timing of speech. Riper [28,29] defined stuttering as "*when the forward flow of speech is interrupted by a motorically disrupted sound, syllable, or word, or by the speaker's reactions thereto*". The "Eight Danger Signs of Stuttering" are listed in [5, 4]. Some of those signs for audio include

- Part-word (syllable) repetitions such as "da-da-da-daddy"
- Substitution of weak vowel in a repetition such as saying "buh-buh-baby" instead of saying "bay-bay-baby"

- Prolongation of a sound such as “mmmmommy”
- Pitch and loudness rise whien repeating or prolonging sound
- Avoidance leading to unusually long or unusual number of pauses

There are two main causes for stuttering. One of them is *anticipatory stress* that leads to the habit of looking ahead for feared sounds and speaking situations. The second cause is *speed stress* that involves speaking at a rate faster than the comprehension rate of the brain. *Frequency auditory feedback* [8], which shifts the frequency of the speech and feeds it back to the speaker, simulates choral speaking and eliminates anticipatory stress. *Delayed Auditory Feedback* [26], which introduces a delay into the feedback speech signal, reduces speed stress. The time lag that is created between the speech articulators such as tongue, teeth etc and the audio feedback causes the brain to slow down the rate of speech. Devices such as *SpeechEasy* [24] and *Fluency Enhancer* [2] that aim at improving stuttering are based on these ideas on altered auditory feedback.

Synthesizing fluent and non-fluent speech can aid in improving the fluency of people who stutter.

4.4 Audio Processing

Our method to modify recorded neutral audio based on emotion and fluency parameters is described in this section. The differences between using synthesized expressive speech and modifying neutral speech to produce expressive speech are discussed in Section 4.4.1.

4.4.1 Synthesized vs Recorded Audio

> *When compared to human speech, synthesized speech is distinguished by insufficient intelligibility, inappropriate prosody and inadequate expressiveness.*
>
> - Janet E. Cahn [12]

The audio for speech can be either be produced synthetically or the voice of the speaker can be recorded. Text-to-speech synthesis systems like Festival [10] can be used to generate audio from text. Festival is used to segment the text into phonemes and durations. This information is used to create a audio waveform using a speech synthesizer such as a Festival voice or MBROLA [13].

In Hofer's work [21], Festival was employed to function as an emotion speech synthesiser. Cahn explored improvements to the affective (expressive) component of synthesized speech by manipulating 17 affect parameters in the areas of pitch, timing, voice quality and articulation [12].

Though the text-to-speech synthesis methods provide an easy way to synthesize speech, the synthesized speech lacks the naturalness of human speech. In creating expressive speech, we chose to record and modify neutral audio based on the emotion and fluency parameters as described in Section 4.4.2. Modifying recorded audio has certain advantages and disadvantages when compared to synthesized expressive audio which are described below.

Recorded audio is more natural when compared to synthesized speech. Recorded audio has the signature voice quality of the speaker embedded in it that distinguishes one speaker from another. Reasonable expressive modifications applied to the recorded neutral audio would likely maintain that signature voice quality. Archival audio of speakers, for example, audio of famous personalities, can be modified and used to create interesting expressive audio.

Synthesized audio has the advantage of portraying any sentence unlike recorded audio, which is limited to the expanse of recorded sentences. Recorded audio needs to be processed first and segmented into smaller units of speech before which they can be modified as described in Section 4.4.2. However, the naturalness of recorded human speech overweighs these factors in producing better expressive auditory speech.

4.4.2 Audio Modification

Neutral audio of different sentences are recorded and are modified to produce emotional audio as described in this section. Praat software [11] is used to analyze the recorded neutral audio and modify it to produce expressive audio. The modifications are based on observations reported in the speech literature as described in Section 4.2 and Section 4.3.

First, the recorded audio is manually broken down into phonemes, syllables, words and durations. The manual segmentation of speech provided us with more accuracy compared to existing speech recognition systems such as CMU Sphinx [1]. Festival [10] is used to obtain the list of phonemes using the text of the spoken sentence. Emotion and fluency based audio modifications are described below.

Emotional Audio Modification

Based on user input of *emphasis level* ranging from 0 to 1 (1 being the maximum emphasis) of words and the nature of the emotion, the duration, pitch and loudness

of the neutral speech segments are modified to produce emotional speech. By default, the duration of the phonemes in the emphasized words that correspond to vowels and stressable consonants are modified based on the emotion as described in Section 4.2. The loudness and pitch of such phonemes are also modified. The factor by which these values are increased or decreased for the different words during a specific emotion is based on their emphasis level. An emphasis level of 0 would lead to no changes in the prosodic elements whereas an emphasis level of 1 would lead to user specified extreme values. The user can override the word emphasis levels by specifying emphasis at the phonemic level.

An example audio modification for angry emotion is illustrated in Figure 4.1 and Figure 4.2. Figure 4.1 shows the waveform and corresponding phonemes and words for the sentence "Do not impose so many rules on me." spoken neutrally. Figure 4.2 shows the modified audio waveform and corresponding speech segments of the same sentence for anger emotion. An emphasis level of 1 was specified for the word "rules" in the sentence. The resulting audio has the word "rules" emphasized with larger intensity and expanded duration.

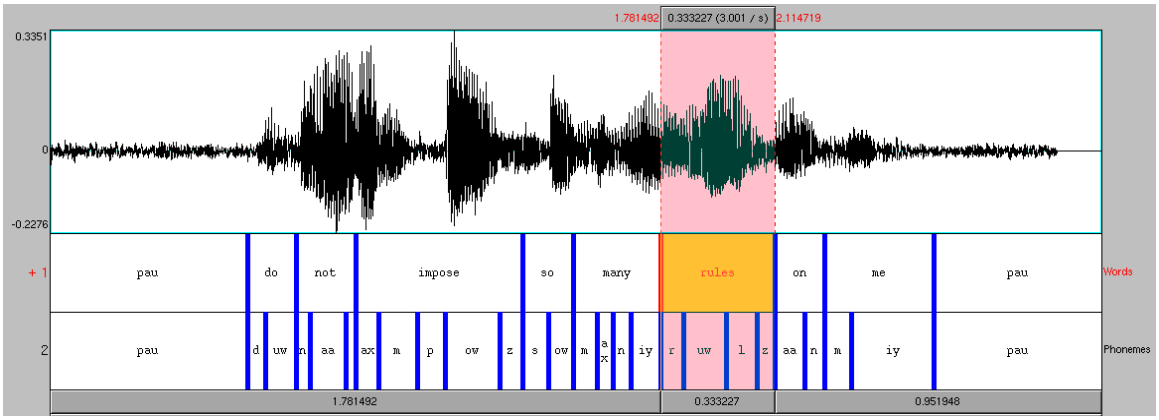


Figure 4.1: Screen capture from Praat software, showing the speech signal and the phoneme-word audio segmentation, of the sentence "Do not impose so many rules on me.", spoken neutrally.

4.5 Summary

Generating expressions for auditory speech is very important to improve speech intelligibility and make speech more natural. The human voice has tremendous potential in portraying emotions, reflecting mood and conveying information effectively.

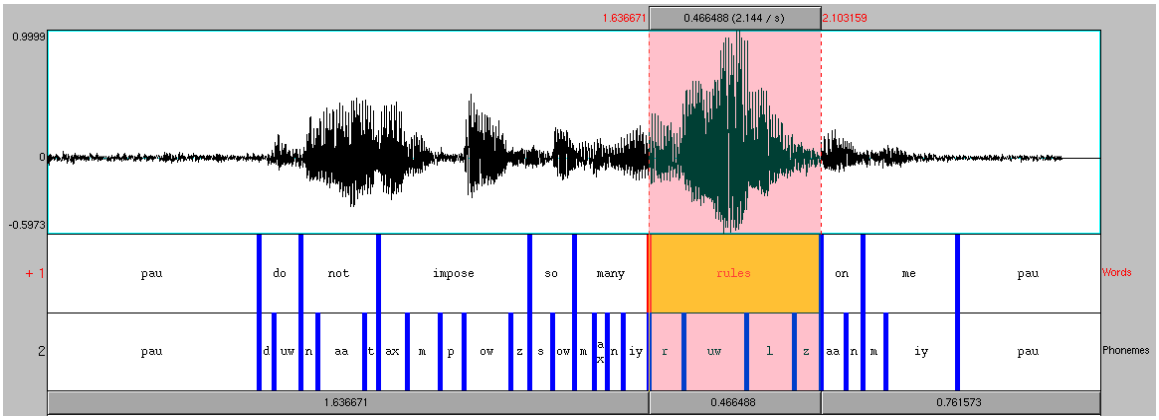


Figure 4.2: Screen capture from Praat software, showing the modified speech signal and the modified phoneme-word audio segmentation, of the sentence "Do not impose so many rules on me.", spoken with anger.

While speech synthesizers are attempting to match the quality of the human voice, synthesized speech lacks the expressiveness of natural voice. We modify neutral recorded audio to produce expressive audio based on the emotion that needs to be portrayed and based on the fluency of the speaker. Auditory prosodic features such as pitch, duration and loudness are modified to add expressiveness to neutral speech. We can generate a variety of emotional, fluent and non-fluent audio examples from a single neutral speech recording.

Bibliography

[1] Cmu sphinx project.
http://cmusphinx.sourceforge.net/.

[2] *The Fluency Enhancer*.
http://www.fluencyenhancer.com.

[3] *Motion Capture Lab, The Advanced Computing Center for the Arts and Design (ACCAD), The Ohio State University*.
http://accad.osu.edu/research/mocap/index.html.

[4] Prevention of stuttering: Identifying the danger signs (video).
http://www.stuttersfa.org.

[5] Stuttering - onset and development.
http://www.d.umn.edu/~cspiller/stutteringpage/onset.htm.

[6] Stuttering - phenomenology.
http://www.d.umn.edu/~cspiller/stutteringpage/phenomenology.html.

[7] *ViconPEAK*. http://www.vicon.com.

[8] J Armson, J. Kalinowski, S Foote, C Witt, and A Stuart. Effect of frequency altered feedback and audience size on stuttering. *European Journal of Disorders of Communication*, 32:359–366, 1997.

[9] Rainer Banse and Klaus R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.

[10] Alan W. Black, Rob Clark, Korin Richmond, Simon King, Heiga Zen, Paul Taylor, and Richard Caley. The festival speech synthesis system. *The Centre for Speech Technology Research, The University of Edinburgh*.
http://www.cstr.ed.ac.uk/projects/festival.

[11] Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 4.3.05) [computer program]. http://www.praat.org.

[12] Janet Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, May 1989.

[13] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van Der Vreken. The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of ICSLP*, volume 3, pages 1393–1396, 1996. http://tcts.fpms.ac.be/synthesis/mbrola.html.

[14] Paul Ekman and Wallace V. Friesen. *Manual for the Facial Action Coding System.* Consulting Psychologists Press, Inc., Palo Alto, California, 1978.

[15] Gary Faigin. *The Artist's Complete Guide to Facial Expression.* Watson-Guptill Publications, NY, 1990.

[16] G. Fairbanks and L. W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. In *Speech Monographs*, volume 8, pages 85–90, 1941.

[17] G. Fairbanks and W. Pronovost. An experimental study of pitch characteristics of the voice during the expression of emotion. In *Speech Monographs*, volume 6, pages 87–104, 1939.

[18] Thomas Gay. Effect on speaking rate on vowel formant movements. In *Journal of the Acoustical Society of America*, volume 63, pages 223–230, 1978.

[19] Thomas Gay. Mechanisms in the control of speech rate. In *Phonetica*, volume 38, pages 148–158, 1981.

[20] Henry Gray. *Anatomy of the Human Body.* Philadelphia: Lea & Febiger, 1918; Bartleby.com, 2000. http://www.bartleby.com/107/.

[21] Gregor O. Hofer. Emotional speech synthesis. Master's thesis, School of Informatics, University of Edinburgh, 2004.

[22] Autodesk Inc. *Maya 6.5 Unlimited Software.* http://www.alias.com/eng/products-services/maya/index.shtml.

[23] Singular Inversions. Facegen software. http://www.facegen.com.

[24] NC Janus Development Group, Inc. Greenville. *SpeechEasy.* http://www.speecheasy.com.

[25] Patrik N. Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814, 2003.

[26] J Kalinowski, A Stuart, S Sark, and J Armson. Stuttering amelioration at various auditory feedback delays and speech rates. *European Journal of Disorders of Communication*, 31:259–269, 1996.

[27] A Mehrabian. Communication without words. *Psychology Today*, 2:53–56, 1968.

[28] Charles Van Riper. *The Nature of Stuttering*. Englewood Cliffs, NJ: Prentice-Hall, Inc, 1971.

[29] Charles Van Riper and L. Emerick. *Speech Correction (7th Edition)*. Englewood Cliffs, NJ: Prentice-Hall, Inc, 1984.

[30] C.E. Williams and K.N. Stevens. Emotions and speech: Some acoustical correlates. In *Journal of the Acoustical Society of America (JASA)*, volume 52(4), pages 1238–1250, 1972.