Realistic Facial Animation During Speech

Arunachalam Somasundaram
Advisor: Dr. Rick Parent
CSE-ACCAD, The Ohio State University

1 Introduction

Realistic facial synthesis is one of the most fundamental problems in computer graphics and one of the most difficult. Both the facial geometry and associated motion convey very important information about the person. Face is one of the most important communication medium a person possesses. Such a magnitude of importance associated with the face brings along with it an equal magnitude of difficulty in modeling and animating it.

We propose to motion capture and analyze expressive facial and lip-sync animation to build a convincing lip sync animation model that can be used in an animation environment. We also plan to model the expressive visual features of expressive speech (for example as observed during angry, sad, happy and surprise speech). In reality, facial movements are produced by the movements of underlying skeleton, muscles and skin layers. Using motion capture we propose to understand and physically model the dynamics of facial muscles to produce realistic facial animation. In an animation environment different facial models that have varying geometry need to be animated. We propose to automatically transfer the motion captured facial movements onto face models that have different geometry while maintaining that model’s individual qualities as specified by an animator.

2 Capturing Data

We use an optical motion capture system at ACCAD using 14 cameras to capture the 3D positions of 85 markers on the face. These markers realistically reflect the movements of the lips, jaw, cheeks, nose, eyebrows, forehead. The marker positions can be seen in Figure 1. We also use markers close to the inner lip boundary to capture fine inner lip movements. We capture the speaker speaking various English sentences and words covering all the basic visemes and phonemes. We capture various facial expressions like anger, surprise, joy, sadness and other facial contortions. And finally, we capture speaking the English sentences with different emotional states such as anger, surprise etc.

Global head movements are significant in highly dynamic states such as anger. These global movements also seem to reflect the mood of the person. The overall global movement of the head of the speaker is not restricted and these global movements have to be subtracted to obtain the face deformations during expression and speech. We use 3 markers on the face placed at rigid points namely, the top forehead marker and one on each ear attachment, to extract the global head movement. For all the frames we match the rotation and translation of the triangle formed by these three markers to a reference frame of capture. We use this translation and rotation values to rotate all the other markers in all the frames. We noticed that there are significant global motions in strong expressive speech like anger. The expression is also reflected in the global head movements. These global transformation values are stored and later will be reapplied to demonstrate the significance of adding global motion along with facial deformations to increase expressive speech animation realism.

3 Expressive Speech Animation Using Constraints

A technique has been developed to automatically add the captured plain facial expressions to neutral speech to produce expressive speech while maintaining the speech constraints that the speech content might have and geometry constraints that the facial model might have. Facial motion during neutral speech can be acquired either using these techniques or with motion capture data. Basic expressions of the face like joy, anger, fear, surprise,
sadness etc are limited in number and can be captured using motion capture techniques. Just adding pure expressive motion on top of speech motion tends to produce motion that does not remain true to the generated speech. Expression blending must be performed while maintaining constraints. Speech constraints such as lip closure for example, while speaking 'p' and 'b' are considered. Geometry constraints such as maximum mouth opening are also considered. Simple ease-in and ease-out weight functions can cause not only errors in constraints but also can decrease (or increase) the speed of speech motion or expression motion so that it does not look realistic. For example, adding a surprise expression when the mouth is open to a word that has a /p/ causes the lip not to close, clearly producing an incorrect lip movement. However surprise expression can expand the lips while saying /ah/ or /i/.

Using an optical motion capture system we capture 3D positions of markers on face during neutral, expressive speech and plain expressions without speech. We divide the face into different regions - lips, jaw, nose, cheek, eyebrow, forehead and use a set of Facial Animation Parameters (for example jaw opening, lip opening, lip width, eyebrow displacement etc.) to study the influence of different expressions on the various regions of the neutral face. The values of some of these parameters during different expressions, along with speech constraints, are used to blend between corresponding regions of the neutral face during speech and the plain expressive face. This produces an approximation to expressive speech. The constraints are automatically detected from the 3D positions of speech markers (for example, lip closures correspond to minimum area between top and bottom lip). We also analyze the difference between the captured neutral speech and expressive speech. We first create a mapping between the captured sequences by computing the area between the lips and using Dynamic Time Warping. The Facial Action Parameters are then extracted and constraints are asserted by comparing the two time-warped captures. We analyze these Facial Action Parameters extracted during speech to compare the recorded expressive speech and calculated expressive speech. Thus, various facial expressions can be added to speech while automatically maintaining speech constraints to produce realistic expressive speech animation. This can be used to synthesize new expressive speech animation from existing captured or animated neutral speech. Figure 2 shows the facial expressions during neutral, surprised, angry and smiling faces. The corresponding movie can be seen at http://www.accad.ohio-state.edu/~asomasun/Face/Movies/SpeechAnim.mov

4 Animation Cloning

In an animation environment different models with varying geometry need to be animated including models with non-human face geometry. Animation of the models including speech animation and various expressions can be a very tedious task. 3D Motion capture can be used to drive the animation of these different models. Two problems need to be addressed here. First, the motion of the various models have be driven faithfully by the motion capture data. Second, the individual characteristics of the model must be maintained as specified by the animator. A technique for such an animation cloning is described below.

The animator designs a particular character with a certain quality. Many of these qualities are in general
Figure 2: Neutral, surprised, angry and smiling Faces

exaggerated to bring out the specific nature of the character. For example, a funny character can distort his facial expressions a lot and a serious character may underplay his happy expressions and exaggerate his angry expressions. Simply directly transferring the motion capture movements taking only the geometry, which is different from the motion captured person, into consideration will not bring out faithfully the animated model’s characteristics. In order to do so, the choses certain key poses from the motion capture face data and carves in 3D the corresponding poses for the animation model. Figure 3 shows some of the key motion captured poses and the corresponding carved poses of the animation model. The corresponding movie can be seen at http://www.accad.ohio-state.edu/~asomasun/Face/Movies/clone.mov

Figure 3: Motion captured and corresonding carved key poses

At exactly the key poses in a motion capture performance, the animated model assumes the corresponding carved posture. The posture of the animated model for all the inbetween transition frames is calculated from the key poses, currently using a naive search algorithm to demonstrate the feasibility of the cloning technique. For each inbetween motion capture pose, the two best motion capture key poses and the corresponding values to blend between them are searched for using a brute force technique. The best key poses and blending values correspond to the minimum sum of the Euclidean distances between the corresponding vertices of the actual and calculated motion capture keyframe poses. In order to calculate the inbetween poses for the animation model the two key poses of the animation model that correspond to those of the motion capture key poses and the calculated alpha blending value are used. This produces the transition frames that are not carved.
5 Future Work

5.1 Modeling Facial Muscles

In reality, facial animation is produced by the movement of the skeleton, muscles and skin. Physically modeling the muscles and their movements can help to control the facial movements as in reality and understand the biomechanical issues involved. Muscle dynamics, i.e., how muscles realistic move when a person speaks or blends between expressions, needs to be understood. Facial Motion capture can be very useful in the process. Since muscles are hidden behind facial skin, the physical properties and motion of the muscles would have to be inferred from the motion capture markers placed on the skin.

We propose to use a 3D muscle and skin model representation of the face. The face can be modeled using approximately 24 major muscles that are responsible for facial expressions and speech articulation. Figure 4 shows some of the major facial muscles.

![Muscles of Facial Expression](image.png)

Simulating the dynamics of muscles can help us understand how various complex expressions are blended and how speech animation including coarticulation is produced.

5.2 Lip sync animation using triphones

Lip sync animation can be done in one of 3 ways: drive the animation directly from a captured performance (mocap data); capture video of speech examples and process the video so that it can be recombined to produce novel speech; use a 3D facial model with associated generative algorithms to produce the visuals associated with speech driven, for example, by text. It is this third category that we address.

To produce visual speech, each phoneme is typically associated with a viseme. A naive approach is to assume that a viseme is a static position and that visual speech can be generated by interpolating (linear or otherwise) between these static positions. There are two problems with this. One is that phonemes (and visemes) change based on the context in which they are spoken. A phoneme, and its associated viseme, changes based on the context in which it is spoken (coarticulation). In speech research, the variations of a phoneme based on context are called its allophones. It is said in the speech literature that (I’m paraphrasing here) phonemes are never spoken, only allophones. The second problem with the naive approach is that a viseme is not a static pose - it is a ‘dynamic shaping’ of the vocal tract. Coarticulation has to interpolate, or blend, these dynamic shapings based on relative importance of what’s being said.

We propose to use to capture facial 3D motion while speaking a corpus of triphones. A triphone is the smallest phoneme-based chunk of forward and backward context for a given phoneme. The literature tells us that there
are about 10,000 triphones. There are approximately 40 phonemes with averages to 250 triphones per phoneme. By capturing all feasible (with respect to the English language) triphones, we capture all feasible allophones of the given phoneme. We expect that the approximately 250 triphones of a phoneme can be divided into a much smaller number of allophonic equivalence classes; the precise number would depend on the quality of speech desired. For each of these equivalence classes, a procedure can be constructed to modify the basic viseme to correspond closely with the captured motion. We hope these procedures to be computationally efficient so that calculated coarticulated speech can be performed at interactive rates. Performing this sampling and classification for each of forty phonemes would require extensive resources. However, it should be possible to perform the classification for a significant number of the most common phonemes and triphones in order to demonstrate the feasibility of the approach and to see if triphones are an effective way to chunk speech when handling coarticulation effects.