A Facial Model and Animation Techniques for Animated Speech

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Scott Alan King, B.S., M.S.

* * * * *

The Ohio State University

2001

Dissertation Committee:

Richard E. Parent, Adviser

Wayne E. Carlson

Han-Wei Shen

Approved by

Adviser

Department of Computer and Information Science

ABSTRACT

Creating animated speech requires a facial model capable of representing the myriad shapes the human face experiences during speech and a method to produce the correct shape at the correct time. We present a facial model designed to support animated speech. Our model has a highly deformable lip model that is grafted onto the input facial geometry providing the necessary geometric complexity for creating lip shapes and high-quality lip renderings. We provide a highly deformable tongue model that can represent the shapes the tongue experiences during speech. We add teeth, gums, and upper palate geometry to complete the inner mouth. For more realistic movement of the skin we consider the underlying soft and hard tissue. To decrease the processing time we hierarchically deform the facial surface.

We also present a method to animate the facial model over time to create animated speech. We use a track-based animation system that has one facial model parameter per track with possibly more than one track per parameter. The tracks contain control points for a curve that describes the value of the parameter over time. We allow many different types and orders of curves that can be combined in different manners. For more realistic speech we develop a coarticulation model that defines visemes as curves instead of a single position. This treats a viseme as a dynamic shaping of the vocal tract and not as a static shape.

This work is dedicated to Tamara, my wonderful spouse, who sacrificed much during the years of research, and to our son Graham whose arrival created the urgency needed for me to finally decide to finish.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr Richard Parent. The commitment that an advisor takes on when he agrees to be your mentor is tremendous, and in my case, extremely appreciated.

I thank the members of my dissertation committee, Dr Wayne E. Carlson and Dr Han-Wei Shen, for their time, valuable comments, and careful reading of this dissertation. I also thank Dr Jacqueline C. Henninger for her careful reading of this thesis and her valuable comments during the oral defense. And I thank Dr Roger Crawfis for his generous support of me and the lab, and for our collaboration.

This work was made possible by the help and support of many people. I would like to thank Barbara Olsafsky for her work on the procedural shaders. I thank Dr Osamu Fujimura for sharing his knowledge of human speech. I would like to thank Texas Instruments Inc. for their financial support of this research, particularly Bruce Flinchbaugh. I would like to thank Dr. Maureen Stone and Andrew Lundberg for their time and data from their work on tongue surface reconstruction.

Like most big software projects I used the work of many people out in the free software community. Their enormous efforts and generosity in releasing their software has saved countless people years of work, and allowed access to cutting edge solutions to numerous problems. In my case I would like to thank in particular: the Festival team (Alan W Black, Paul Taylor, Richard Caley and Rob Clark) [BTCC00];

Jonathan Richard Shewchuk for his work on Triangle [She96a]; the MBROLA project [MBR99]; the Visible Human Project [NLH99] sponsored by the National Library of Medicine; Viewpoint Digital, Inc. for their free scan of my head at SIGGRAPH; and Cyberware, Inc. for their release of 3D models on the net, particularly that of teeth.

Our lab has a great working environment with lots of talented students who are always willing to talk about ideas, give advice on a problem or read a rough draft. I give thanks to all of the members of our lab and department that have shared ideas, support and an ear with me or just made my experience at OSU a better one. Particularly I'd like to thank Frank Adelstein, Mowgli Assor, Kirk Bowers, Paolo Bucci, Tamera Cramer, Steve Demlow, Sandy Farrar, Tom Fine, Mark Fontana, Margaret Geroch, Sandy Hill, Leslie Holton, Yair Kurzion, Matt Lewis, Nathan Loofbourrow, Marty Marlatt, Steve May, Torsten Möller, Klaus Mueller, Barbara Olsafsky, Elizabeth O'Neill, Johan Östmann, Eleanor Quinlan, Kevin Rodgers, Steve Romig, Ferdi Scheepers, Naeem Shareef, Po-wen Shih, Karansher Singh, Don Stredney, Brad Winemiller, Suba Varadarajan, Lawson Wade, and Pete Ware for not only helping me get through the research, but for also giving me a life during my time at OSU.

VITA

1988	B.S. Computer Science, Utah State University
1994	M.S. Computer and Information Science, The Ohio State University
Sep, 1988 - Sep, 1989	Tandy Corp - Programmer
Sep, 1989 - Jun, 1991	General Dynamics - Software Engineer
Jun, 1991 - Sep, 1992	Harris Methodist Health Systems - Programmer Analyst II
1992-2000	Graduate Assistant, Department of Computer and Information Science, The Ohio State University
Jan 1995-Jan 1996	Graduate Research Assistant, Ohio Supercomputer Center, The Ohio State University
1995-1997	Summer Intern, Texas Instruments
1997-2000	Graduate Research Assistant, Department of Computer and Information Science, The Ohio State University
Sep 2000 - present	Lecturer, Department of Computer and Information Science, The Ohio State University

PUBLICATIONS

Research Publications

Scott A. King, Roger A. Crawfis and Wayland Reid, "Fast Volume Rendering and Animation of Amorphous Phenomena", chapter 14 in *Volume Graphics* edited by Min Chen, Arie E. Kaufman and Roni Yagel, Springer, London, 2000.

FIELDS OF STUDY

Major Field: Computer and Information Science

Studies in:

Computer Graphics	Prof. Richard E. Parent
Software Methodology	Prof. Spiro Michaylov
Communications	Prof. Dhabaleswar K. Panda

TABLE OF CONTENTS

Page

Abstract ii

Dedication iii

Acknowledgments iv

Vita vi

List of Tables xiii

List of Figures xv

Chapters:

1. Introduction 1

    1.1 Applications of Facial Modeling and Animated Speech 3
    1.2 Motivation For Using Computers To Animate Speech 4
    1.3 Current Solutions for Speech-Synchrony 6
    1.4 Thesis Overview 10
    1.5 Organization of Thesis 12

2. The Facial Model 15

    2.1 Skin 18
        2.1.1 Characteristic Points 19
        2.1.2 Cylindrical space 21
    2.2 Skull 22
        2.2.1 Collision Detection. 24
    2.3 Lips 27
    2.4 Tongue 27

2.5 Eyes . . . 28
2.6 Other facial parts . . . 29
2.7 Summary . . . 30

3. The Lip Model . . . 31

3.1 Introduction . . . 31
3.2 Previous Work . . . 34
3.2.1 Speech Reading . . . 34
3.2.2 Computer Models . . . 36
3.2.3 ICP Lip Model . . . 37
3.3 Lip Anatomy . . . 38
3.3.1 The Mandible . . . 40
3.4 Lip Parameterization . . . 41
3.5 Implementation . . . 45
3.5.1 Grafting . . . 48
3.6 Rendering . . . 49
3.7 Results . . . 51
3.8 Summary . . . 53

4. The Tongue Model . . . 56

4.1 Introduction . . . 56
4.2 Previous Work . . . 57
4.3 Tongue Anatomy . . . 60
4.4 Tongue Model . . . 62
4.4.1 Geometry . . . 63
4.4.2 Tongue Parameterization . . . 65
4.4.3 Implementation . . . 66
4.5 Rendering . . . 69
4.6 Results . . . 69
4.7 Summary . . . 72

5. Speech-Synchronized Animation . . . 76

5.1 Coarticulation . . . 76
5.1.1 Previous Work . . . 77
5.1.2 Our Coarticulation Model . . . 80
5.2 Results . . . 92
5.3 Summary . . . 92

6. TalkingHead: A Text-to-Audiovisual-Speech System 94

6.1 Introduction 94
6.2 System Overview 95
6.3 Text-To-Speech Synthesis 96
6.4 The Viseme And Expression Generator 97
6.5 The TalkingHead Animation Subsystem 100
6.5.1 Different Characters 100
6.6 Summary 102

7. Results 103

7.1 Results From Our Facial Model 104
7.1.1 The Tongue Model 104
7.1.2 The Lip Model 109
7.2 Animation Results 112
7.3 Summary 125

8. Future Research 130

8.1 Future Modeling Research 130
8.1.1 Hair 131
8.1.2 Rendering Skin 131
8.2 Future Animation Research 132
8.2.1 Syllables as Concatenative Units 133
8.2.2 Cineradiographic Database 134
8.2.3 Prosody 135
8.3 Improvements to TalkingHead 136
8.4 Validation Study 137
8.5 Summary 137

Appendices:

A. Facial Animation Overview 139

A.1 Background 139
A.1.1 Definition of Facial Animation 141
A.2 Anatomy 142
A.2.1 Skin 143
A.2.2 The Skull 145
A.2.3 The Tongue 149
A.2.4 The Lips 150

A.2.5 The Cheeks . . . 151
A.2.6 Vocal Tract . . . 152
A.2.7 Muscles . . . 152
A.2.8 Articulation of Sound . . . 159
A.3 Modeling . . . 160
A.3.1 Geometric Representations . . . 160
A.3.2 Instance Definition . . . 163
A.3.3 Geometry Acquisition . . . 168
A.3.4 Rendering . . . 176
A.3.5 Famous Models . . . 181
A.3.6 Other Facial Parts . . . 193
A.4 Parameterizations . . . 204
A.4.1 Empirical Methods . . . 205
A.4.2 Muscle Based . . . 206
A.4.3 FACS . . . 207
A.5 Animation Techniques . . . 214
A.5.1 Image Based Animation . . . 215
A.5.2 Hierarchical Control of Animation . . . 217
A.5.3 Animation Using Scripting . . . 218
A.5.4 Procedural Animation . . . 219
A.5.5 Animating Expressions . . . 219
A.5.6 Physically Based Animation . . . 221
A.5.7 Keyframing . . . 221
A.5.8 Performance Animation . . . 223
A.5.9 Displacement . . . 224
A.5.10 Natural Language Animation . . . 225
A.6 Animating Speech . . . 225
A.6.1 Contribution of the Visual Signal in Speech Perception . . . 226
A.6.2 Head Movement During Speech . . . 228
A.6.3 Lip Synchrony . . . 228
A.6.4 Coarticulation . . . 233
A.6.5 Emotion in Speech . . . 236
A.6.6 Animating Conversations . . . 237
A.7 Surgery . . . 241
A.8 Vision . . . 243
A.8.1 Tracking . . . 244
A.8.2 Face Detection and Recognition . . . 245
A.8.3 Compression . . . 246
A.8.4 Speech Reading . . . 246
A.8.5 Expression Detection . . . 247
A.8.6 Interfaces . . . 247

A.9 Standards 248
A.9.1 MPEG-4 SNHC FBA 249

B. Medical Appendix 252

B.1 Terms of location and orientation 252
B.2 Body tissues 254
B.3 Anatomical Terms of the Face 256

Bibliography 257

LIST OF TABLES

Table **Page**

3.1 Muscles of the lips 39

A.1 Bones that comprise the skull. 146

A.2 The muscles of the mandible and their actions. 154

A.3 The muscles of the scalp and their actions. 154

A.4 The muscles of the eyes and their actions. 155

A.5 The muscles of the nose and their actions. 156

A.6 The muscles of the mouth and their actions. 156

A.7 The muscles of the tongue and their actions. 157

A.8 The muscles of the soft palate and their actions. 158

A.9 The muscles of the pharynx and their actions. 158

A.10 FACS Action units for the Upper Face group. 208

A.11 FACS Action units in the Lower Face Up/Down group. 209

A.12 FACS Action units in the Lower Face Horizontal group. 210

A.13 FACS Action units in the Lower Face Oblique group. 210

A.14 FACS Action units in the Lower Face Orbital group. 211

A.15 FACS Action units in the Lower Face Miscellaneous group. 212

A.16 FACS head and eye positions. 212

A.17 Translation of FACS AUs into emotions 214

LIST OF FIGURES

Figure **Page**

2.1 Characteristic points 19

2.2 Barycentric coordinates 20

2.3 Slices used to create the 3D geometry of the skull. 23

2.4 The generic skull we fit to the input geometry. 24

2.5 An uncapped cylindrical arc. 26

3.1 The temporomandibular joint. 41

3.2 Lip model geometry 45

3.3 The grafting process 50

3.4 High quality rendered lips 50

3.5 Face model saying "Oh" 52

3.6 Frames from an animation rendered using our lip shaders. 52

3.7 Frames from an animation rendered using our lip shaders. 53

3.8 Our lip model in the position of /aw/, a happy /aw/ and a half smile. 53

4.1 Tongue images from the Visible Human Project 61

4.2 The tongue geometry and its control grid. 64

4.3 Samples from the four categories of tongue shapes from Stone 68

4.4 Our facial model in the viseme position for the phoneme /d/ 70

4.5 The tongue model in the position of the phoneme /ay/. 70

4.6 A side-by-side comparison of our facial model with and without the tongue model 71

4.7 Frames from an animation of our facial model licking its lips. 72

4.8 Frames from an animation of our facial model sticking out its tongue. 73

5.1 Cohen and Massaro Coarticulation 83

5.2 Our version of coarticulation based 84

5.3 Comparison between our coarticulation and Cohen and Massaro coarticulation 85

5.4 Curves for the orbicularis oris measured by hand from video. 86

5.5 The curve generated by our coarticulation model for the orbicularis oris parameter. 87

5.6 Our coarticulation model versus linear interpolation. 88

5.7 Our coarticulation model versus cubic interpolation. 89

5.8 Our coarticulation model versus Bezier approximation. 90

5.9 Our coarticulation model versus rational Bezier approximation. . . . 90

5.10 Our coarticulation model versus B-spline approximation. 91

5.11 Our coarticulation model versus rational B-spline approximation. . . 91

5.12 A curve for the Jaw_Open parameter 92

6.1 TalkingHead System Overview 96

6.2 An example viseme definition for the phoneme /p/ 98

6.3 Screen shot of TalkingHead 101

7.1 Samples from the four categories of tongue shapes from Stone 105

7.2 A side-by-side comparison of our facial model with and without the tongue model 106

7.3 Frames from an animation of our facial model licking its lips. 107

7.4 Frames from an animation of our facial model sticking out its tongue. 108

7.5 Our facial model sticking out its tongue. 108

7.6 Our facial model with its lips puckered as if to kiss. 109

7.7 Our facial model is happy. 110

7.8 Our facial model expressing sadness. 111

7.9 Our facial model in a very happy mood, with a very big grin. 111

7.10 Our facial model in the position of a wry smile. 112

7.11 A comparison of the waveforms for the synthetic and natural speech. 113

7.12 Animation of speech compared to recorded speech. 115

7.13 Animation of speech compared to recorded speech. 116

7.14 Animation of speech compared to recorded speech. 117

7.15 Animation of speech compared to recorded speech. 118

7.16 Animation of speech compared to recorded speech. 119

7.17 Animation of speech compared to recorded speech. 120

7.18 Animation of speech compared to recorded speech. 121

7.19 Animation of speech compared to recorded speech. 122

7.20 Animation of speech compared to recorded speech. 123

7.21 Animation of the phrase “Hello world.” 126

7.22 Animation of the phrase “Hello world.” 127

7.23 Animations produced by the techniques described in this dissertation 128

7.24 Animations produced by the techniques described in this dissertation 128

A.1 A schematic of a facial animation system. 142

A.2 A cross section of the skin. 144

A.3 The temporomandibular joint. 148

A.4 Movement of the temporomandibular joint. 148

A.5 The vocal tract. 153

A.6 Waters vector muscle model 186

A.7 Waters sheet muscle model. 187

A.8 A typical design for speech synchronized animation. 229

B.1 The planes that describe location and orientation in human anatomy. 253

CHAPTER 1

INTRODUCTION

Humans communicate primarily with the spoken word and have a lifetime of experience both in speaking and in listening. Talking heads, therefore, make an excellent choice for communicating with humans. Motion pictures allow an event or performance to be recorded and replayed at a later time and place making film an attractive choice for delivering communication. Not all events can be recorded on film. An animator can create a motion picture of a performance that otherwise may not be possible to obtain. Combining talking heads with animation, allows an animator to communicate with her audience.

Speech is a multi-modal signal with an audial and visual component. The amount of information contained in each mode is debated, but clearly speech can be comprehended using either of the two alone, so there is a great amount of information in both. Using the entire signal increases the comprehension rate. Using the redundant information is important when part of the signal is noisy. Cohen and Massaro [CM93] add poor quality video (only one of every three frames displayed) to audio and report phoneme[1] recognition rates of 55% with sound only, 4% with visual only, and 72%

[1]A phoneme is a phonetic element of speech. There is a lot of debate as to what constitutes the basic elements of speech, those that can be concatenated together to create an utterance. Phonemes (phonetic elements) are a popular choice, with the number of phonemes per language varying widely.

with both. Their results show that the two modes combine in a super-additive fashion and that the visual cues are important in recognizing speech. Similar results from other researchers [BS83, ESF79, GMOB95, GMAB97, MC83, McG85] have also been reported.

When creating animated speech, an animator is generally concerned with synchronizing the lips to the sound. The motion of the lips is extremely complex and difficult to reproduce by a human at a frame-by-frame basis. If an attempt is made to animate each frame, the resulting animation generally has too much motion and does not look natural. Instead, a keyframing approach that uses just a few different lip positions [Mad69] is often more successful. The resulting animation is not always realistic, but still aids the listener in understanding the speech. Fortunately for the animators, the audience often dispels their disbelief rather quickly and just concentrates on understanding the speech.

Computer graphics has been successfully used to reduce the effort and required skill level of animators in many areas of animation, such as crowds, physics, 3D set layouts, inbetweening, etc. Naturally, attention was turned to reducing the difficult task of creating animation of speech using computers. For nearly three decades, researchers [BL85, CPB+94, deG89, Ess94, EP98, Gue89a, GGW+98, McG85, Pat91, Par72, Par74, Par75, PW96, Pel91, PHL+98, SMP81, Pla85, Wat87, Wat87, Wil90a] have actively pursued techniques to create computer animations of talking heads. For the purpose of communication, the important areas of research have been concerned with facial modeling, animating facial expressions, and lip synchronization.

1.1 Applications of Facial Modeling and Animated Speech

Facial modeling and animation have diverse uses in a wide range of fields including the entertainment industry, education, computer vision, surgery, human-computer interaction, and virtual environments to name a few.

The entertainment industry, which includes film, television, commercials, theme rides, and games has produced some of the most realistic facial animations. They use facial animation to tell us a story, make us laugh, make us cry, make us spend our money, and help us immerse ourselves in a new world, if only for a little while. They use computer graphics to create scenes that would otherwise be impossible, or prohibitively expensive to create. Examples include scenes of talking Martians, animals, John F. Kennedy, John Wayne, etc. They achieve these results at great costs, in talent, equipment, and time.

Because speech is a natural method of communication, a talking head is a wonderful user interface. Instead of a person reading an encyclopedia, having a talking head recite the encyclopedia may make it more exciting for a child to learn possibly leading to easier or quicker understanding. Additional uses of speech include a talking catalog that lists the products of a company, a talking kiosk that greets visitors to a city and tells them where to find fun and excitement, or a greeter for a web site. Take for example, the talking catalog for a company that sells thousands of products. To record an actor discussing the particulars for all of the products is prohibitively expensive in time and storage. In addition, product updates need to be considered. If the company is international, or a country with a multi-lingual culture, the products would need to be described in different languages. Animation of talking heads is a great solution, especially if it can be derived from text. The company already

updates the text in the catalogs, the animation just needs to be regenerated, and if the generation is fast, only storage for the text is required.

Teleconferencing is a popular and productive method for getting groups of people together that are physically separate. The cost in equipment, cabling, and bandwidth of sending video can be enormous. Teleconferencing also lacks important eye-to-eye contact; one cannot look directly at the camera and screen at the same time. Virtual teleconferencing systems attempt to overcome these drawbacks with compression to reduce bandwidth and immersive virtual reality techniques to allow participants to make eye contact and also to interact with the same objects. In such a system, a digital duplicate, called an avatar, represents a participant, generally with similar behavior but not necessarily similar appearance.

One of the challenges of teaching is to attract and maintain the student's attention. A possible method to garner attention is to use animations of characters that interest the student. Imagine that the student would like to learn physics from Albert Einstein or maybe Donald Duck. Other educational applications include teaching the deaf to read speech and also to speak, and teaching foreign languages.

1.2 Motivation For Using Computers To Animate Speech

Speech is important as a method of communication to a human audience. Animation makes it possible to create verbal communication that otherwise would not be possible, such as individuals no longer living, aliens, animals, and fictitious characters. Computer graphics has the ability to decrease the costs of creating animations of talking characters. The costs include time, skill, equipment, personnel, and storage. Other advantages of computer graphics over hand drawn animation are that with

computer graphics it is easier to produce photo-realistic animation and with true 3D representation changing camera views is simplified.

Creating convincing computer-animated speech requires a highly deformable facial model capable of shaping the visible vocal tract and facial surface in a realistic manner along with convincing animation of that model. The problem of creating computer-animated speech can be split into modeling and animating aspects. Modeling involves creating deformable geometry to represent the visible portions of the head. The role of the underlying tissues in shaping the surface still needs to be considered even if not explicitly modeled. Animation of a facial model involves manipulating the geometry over time to create realistic motion.

In general, a facial animation solution should have the following properties:

- Static realism.
- Dynamic realism.
- Individual realism.

Creating *Static realism* is a modeling problem and involves creating a realistic face for a single frame, that is, the geometry and the rendering of the face look real. Realistic geometry requires a medium-resolution to high-resolution surface. Creating such a surface is a well-understood problem. Realistic rendering is also a well-understood problem and the most successful solutions involve texture mapping. Static realism can be achieved by any number of methods that capture geometry, such as from photographs or using a laser to create a distance map. For a model to have static realism, realism for a single neutral pose, or a small set of poses, is not sufficient; realism must exist for any natural shape the face may take during the animation.

Dynamic realism requires that the motion be realistic. Dynamic realism is independent from static realism. Dynamic realism can be achieved using just outlines or wire-frame rendering. However, static realism is generally more effective when combined with dynamic realism. Dynamic realism in facial animation, particularly speech, is difficult because the face is highly deformable, small deformations can be significant, and the audience members are experts, having seen real speech their entire lives.

Individual realism requires that the facial model not only look, but also act, like a particular person. Static and dynamic realism are necessary but not sufficient for individual realism. Often only *static individual realism* is sought, where the synthetic character looks like the correct person, but does not necessarily move or act like that person. *Dynamic individual realism* is when the synthetic character moves or acts like a particular person but does not look like that person. This is similar to doing an impersonation.

1.3 Current Solutions for Speech-Synchrony

Most facial modeling research to date has not concentrated on the specific aspects required for realistic speech-synchrony. Meanwhile, the research into speech-synchrony has mainly focused on animation and not on modeling. Facial modeling to support speech must concentrate on the mouth, which includes the lips, the tongue, the skin surrounding the mouth and the jaw.

Many different computer animation techniques have been used for facial animation with limited success. Keyframing [BL85, deG89, deG90, EP97, EP98, Gou89, Kle89, MAH90, PW91, Pat91, Par75, PBS91, Pel91, PVY93, PALS97, PHL$^+$98, PB95,

Wai89, WL93] has been by far the most common method. Keyframing produces adequate results for facial expressions but unsatisfactory ones for speech-synchronized animation. Keyframing is a technique borrowed from the craft of hand drawn animation. Keyframes ease the effort of the animator and produce smoother animations as compared to attempting to create each frame separately.

Using linear or spline interpolation, which works well for most animation, falls short when animating speech. The primary problem with animating speech is the difficulty of creating the keyframes. The common approach is to break the speech into phonemes and then convert them into visemes[2] which become the keyframes. But, a phoneme does not always look the same. Its appearance depends on its emphasis, duration and the neighboring phonemes. These effects are known as coarticulation and prosody. *Coarticulation* is due to the fact that the vocal tract is made up of tissues that have finite acceleration and deceleration. Often, there is insufficient time between phonemes for an articulator to reach the next ideal position or hold it long enough. The articulator must therefore either start its motion earlier or lag behind blurring the lines between phonemes. *Prosodic* features of speech include length, accent, stress, tone, and intonation[Fox00] and effect the timing and position of the vocal tract parts. Coarticulation is physical, while prosody is systemic.

Coarticulation methods have been employed to increase the realism of animated speech, generally in combination with keyframing. Pelachaud et al. [Pel91, PBS91, PVY93] handle coarticulation using a look-ahead model [KM77] that considers articulatory adjustment on a sequence of consonants followed or preceded by a vowel.

[2]A viseme is a visual element of speech. Phonemes correspond to the audio portion of speech and visemes to the visual part of speech. Visemes grew out of speechreading education. Generally, there are far fewer visemes than phonemes since many different sounds do not look visually distinct since the articulators of the vocal tract that make them different are not visible.

Cohen and Massaro [CM93] use the articulatory gesture model of Löfqvist [Löf90], which uses the idea of dominance functions. Each segment of speech has a dominance function for each articulator. The dominance functions overlap and are blended together to determine the correct position for the articulator. Waters and Levergood [WL93] use cosine acceleration and deceleration and assign masses to the targets with attached springs to simulate the elastic nature of tissue.

Prosodic effects are rarely considered. Work at the University of Pennsylvania [CPB+94, Pel91, PBS91, PVY93] on automated generation of conversations considers intonation. Their system automatically generates animated conversations between multiple human-like agents, with speech, intonation, facial expressions and hand gestures. However, intonation is handled with facial expressions and does not modify the appearance of the viseme.

Some of the best results of speech synchronization to date involve some sort of motion capture [deG89, deG90, PLG91, SVG95, Stu98, Wil90b, Wil90a]. Motion capture is an animation method that can drive any type of model if a correspondence between the motion capture data and deformations of the model exists. The motion of the lips and key facial feature points are tracked and the audio is recorded while the actor is reciting the dialog. The tracked feature points are used to drive the computer generated character while the recorded dialog is played. The resulting speech-synchronized animation is of extremely good quality if good motion capture equipment is combined with a high quality facial model. This solution has two main problems, namely, expense and the speech must be previously recorded. The expense of the method has many facets. First, the equipment is very expensive. Talent can also be expensive especially when using a famous actor for the voice talent. Time is an

additional expense to consider. Once the data is captured it must be post processed to make it usable. This requires time and experience. Since motion capture is just capturing a specific performance, any animation that is to be created must previously be performed. That means no novel speech animation can be created, only speech previously recorded. A minor change in the dialog requires that new motion capture data be acquired.

2D image processing techniques [EP97, EP98, PB95] have also been used with good results for speech synchronization. The character is filmed speaking a corpus that includes all the necessary phonemes or triphones (a combination of three phonemes). Triphones are popular because most of the coarticulation effects on the middle phoneme are present in the triphone. Blending the triphones together with image processing techniques gives very realistic motion. These systems give very good quality and allow novel speech to be constructed. However, the character to be animated must be available to speak the corpus. For non-human characters or historical figures this may not possible. In addition, these methods are 2D techniques so lighting changes and novel camera views are problematic.

Some of the first uses of a tongue in computer facial animation were in the creation of animated shorts [Kle89, Ree90]. Although the tongue exists in these shorts, it is considered secondary. Parke [Par74] uses 10 parameters to control the lips, teeth, and jaw during speech synchronized animation. The parameters are chosen empirically and from traditional hand drawn animation methods. These parameters give control over the lip motion, but have limited control over lip shape. Cohen and Massaro [CM93] modify the Parke model to include a simple tongue and new lip parameters. The tongue animates stiffly and the limitations of a simple, rigid tongue are quite

apparent during excited conversation when the mouth is opened wide and during non-speech-related animation such as licking the lips or sticking out the tongue.

Pelachaud et al. [PvOS94] develop a method of modeling and animating the tongue using soft objects, taking into account volume preservation and penetration issues. This is the first work on a highly deformable tongue model. The major drawbacks of this method are the computationally expensive rendering and processing plus the large number of parameters.

Research at the Institut de la Communication Parlée [Adj93, GM92, GMAB94] in Grenoble France resulted in the ICP lip model. The ICP lip model is driven by five parameters determined by regression analysis on lip contours. The ICP lip model was designed by analyzing lip contours during speech for use in a tracking system. It is quite capable in tracking applications, however, in animation systems the model suffers from interior shape inconsistencies and does not provide enough variability in shape.

1.4 Thesis Overview

This research tackles the problem of creating realistic animated speech by attacking both modeling and animating. Without a highly deformable model, the best animation method will fall short, and conversely without a good animation method a highly deformable model will not matter. Since the information in the visual part of speech is concentrated in the mouth area we create highly deformable lips and tongue plus add teeth and an inner mouth. To create photo realistic images we propose a method to render surface detail of the lips and tongue. We consider the underlying tissue since it plays an important role in the final shape of the surface. To create

realistic motion we develop a coarticulation model that treats visemes as a dynamic shaping of the vocal tract instead of as a single keyframe.

Our research allows us to achieve individual realism. To achieve static individual realism our facial model accepts a facial mesh and texture map as input. This method is chosen over fitting a prototype because:

- When adapting a prototype the final mesh can retain traits of the prototype.
- If the data being adapted to is of higher resolution than the prototype, particularly in areas of high curvature, information will be lost.
- A prototype is limited in the possible facial geometries it can represent.

To create convincing speech, a facial model must be able to represent the myriad different shapes of the lips, inner mouth and tongue, the movement of the skin around the mouth due to lip movement, the deformations due to movement of the jaw, and the deformations in the surface of the face due to the contractions of the muscles. In addition, the facial model must be animated properly to achieve intelligible visual speech.

Our facial model considers the surface features and also the subsurface features and was designed to support speech synchronization as well as facial expressions. The facial model contains:

- a realistic skull layer with a movable jaw that affects the surface shape,
- a highly deformable lip model that considers the full lips as well as the mucous membrane and is grafted onto the input geometry,
- a parametric deformable tongue model capable of realistic tongue shapes,

- an inner mouth with teeth, gums, upper palate and uvula,
- and movable eyeballs and eyelids.

The Facial Model is used in a text-to-audiovisual-speech (TTAVS) system showing the model's utility.

In order to achieve dynamic realism, we propose a method of animating speech that treats the production of phonemes as a dynamic process instead of as a static one. This allows for better motion of the lips, tongue and jaw. We also modify the coarticulation model of Cohen and Massaro [CM93], which is based on the articulatory gesture model of Löfqvist [Löf90]. We modify the shape of the dominance functions to give better velocity and acceleration properties. We also modify the blending of the dominance functions to take away global control over movement of the articulators.

1.5 Organization of Thesis

Chapters 2-4 discuss modeling issues. Chapter 2 presents the facial model we have developed to support animated speech. Our facial model concentrates on the mouth, which provides most of the information from the visual part of speech. The facial model considers the internal tissues of the face as well as the surface to give more realistic motion. The facial model accepts input surface geometry and modifies it to include lips, tongue, teeth, gums, eyes, and eyelids.

Chapter 3 describes a new parametric lip model capable of producing the lip shapes necessary for speech as well as lip shapes for general facial animation. The lip model is grafted onto the input geometry to guarantee full lip geometry and the necessary complexity. A method of rendering the lips to include surface detail and lighting effects is also presented.

Chapter 4 presents a new deformable tongue model to support speech animation in detail. The tongue model is capable of all tongue shapes necessary for English speech as well as most tongue shapes required during facial animation. The tongue model is composed of a B-spline surface and a parameterization to define tongue shapes. A method of rendering the tongue for photorealism is also discussed.

Chapter 6 is an overview of TalkingHead, our text-to-audiovisual-speech (TTAVS) system. This system accepts text and produces an animation of a talking head reciting the text. Both the audio and video are created from the text allowing for the generation of vast amounts of novel animations with minimal storage requirements.

Ultimately, the animation method determines the quality of the animation, just as the skill of the master animator determines the quality of a hand drawn animation. Chapter 5 describes our method of animating speech using a modification of the coarticulation model by Cohen and Massaro [CM93].

Chapter 7 reports on results we have obtained using our system. This work is about creating speech synchronized animation, and the only way to truly appreciate the results is to see and hear the animation. Unfortunately, technology does not currently allow animations in a document such as this. Instead, a comparison between actual recorded speech and synthetic speech is shown using snapshots and motion curves.

This thesis gives preliminary results of a partial solution that extends the state of the art in a positive and important way. Chapter 8 sketches possible extensions to this work as well as new applications of this work.

Appendix A contains an extensive overview of computer facial animation. The overview gives a broad base of the knowledge necessary for facial animation including

anatomy, facial modeling, animation methods, parameterizations, vision applications, standards and surgery. The overview covers around 200 of the more than 400 research papers written on facial animation to date.

Because of the complexity of human anatomy, descriptive anatomical terms were devised to decrease confusion when discussing anatomical structures, their motion and their location. Appendix B lists many of the anatomical terms that are likely to be encountered when studying or describing the anatomy of the face and speaking apparatus.

CHAPTER 2

THE FACIAL MODEL

Speech is multi-modal with a visual and audial component. The listener uses the sound along with the motion of the lips to determine which words are being spoken. The listener also looks for nonverbal cues to get information on the speaker's intent and emotion. The nonverbal cues come from hand gestures, body language and facial expressions. To achieve communication with facial animation, a model must be capable of synchronizing the lips and displaying expressions. Realistic facial animation is necessary for convincing the viewer that they are viewing a real human being, but it is not necessary for communicating with the listener. Cartoons with rudimentary lip synchronization and simple expressions are capable of conveying what the artist wants. Systems such as that of Cohen and Massaro [CM93] that do not try to achieve total realism are still effective since the listener often dispels her disbelief after a short time of listening to the message and instead concentrates on comprehending the message.

Creating realistic animation of a deformable object starts with a good model of the deformable object. The object must be capable of adequately portraying any shape needed during the animation. Creating animated speech requires an adequate facial model that can represent the large space of possible mouth shapes that includes not

only the lips but also the inner mouth. The inner mouth includes the tongue, teeth, gums, upper palate and inner cheeks. To be able to create realistic animated speech we first need a highly deformable facial model, which we describe in this chapter.

The terms model and modeling have diverse meanings in computer graphics. Here we use *modeling* to mean the process of defining the shape of an object, that is creating the model. We define the term *model* as a description of an object that includes geometry, surface properties, animation controls, and rendering; basically everything required to create a single image of the object. The geometry can be described in any number of ways such as with triangles, implicitly, with splines, etc. and can be described in more than one way. At the minimum, the visible surface needs to be described but often subsurface information is also included. Surface properties include color, texture, and lighting information. Animation controls are a method of describing the shape of the model, which may use deformations from a base shape or may generate a complete description of the geometry. Animation controls generally tend to have fewer degrees of freedom than the description of the geometry. A popular choice is a high-level parameterization such as facial expressions. A model can have any number or type of animation controls.

Our model consists of:

- Geometry of the surface of the face that is input to the model. This geometry can come from diverse sources such as a laser scanner, from photographs, or a human modeler. We anticipate laser scanners or photographs to be the most common sources and therefore use points and triangles internally to represent the surface.

- A highly deformable model of the lips that is grafted onto the input geometry to guarantee acceptable complexity. This lip model was developed as part of this research.

- A highly deformable tongue capable of the shapes necessary for speech. The tongue model was created during our research to support more realistic speech.

- Full mouth geometry, including teeth, gums, upper palate and mucous membrane.

- Articulable eyes.

- A parameterization for control over the tongue, lips, jaw, eyes, and head rotation. This parameterization gives control to the animator over the shape of the geometry.

- Consideration of the subsurface tissues with a realistic skull fitted to the surface geometry based on soft tissue depths.

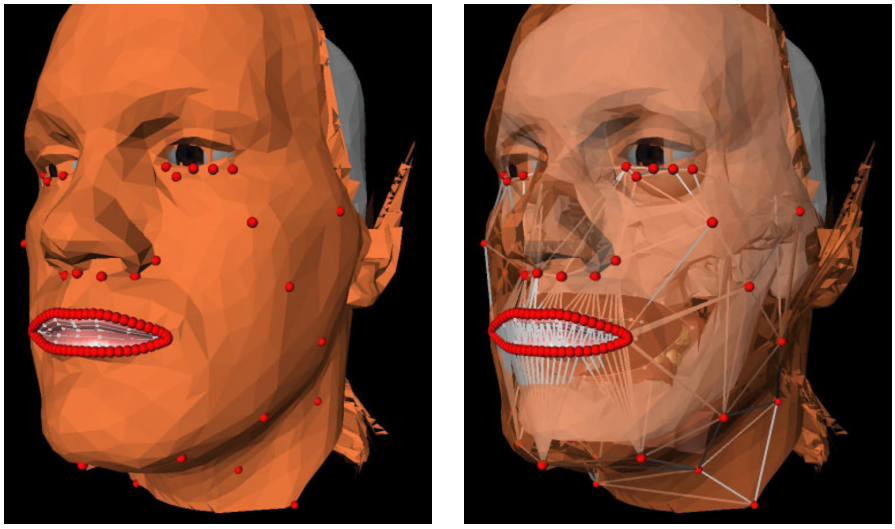
To achieve static realism we require a model that is highly deformable and capable of representing the huge space of possible shapes of the face. To reach our goal of animated speech, we create a facial model specifically designed for animated speech with a highly deformable mouth. This is in contrast to most facial modeling, which is designed for facial expressions. We concentrate here on solving the problem of speech-synchronized animation and ignore facial expressions. Facial expressions are very important in communicating meaning and emotion and need to be included in a final solution. However, the synchronization of speech is the focus of this research, and we leave the animation of expressions for future work.

In order to create static individual realism, accurate geometry and good texture information is required. A common solution is to use a laser scanner to get high-resolution geometry along with surface color information. Since laser scanners generate a depth map in a regular grid, points and polygons are a natural representation for the surface. Triangles are also fast to render in hardware. In addition, a point on the surface can be placed at a specific location in $\Re^3$ by a translation. We thus use triangles to represent the model geometry. Other types of representations for the input geometry are acceptable, they just need to be converted to polygons. Spline representation could be used directly with some algorithmic and procedural changes.

Dynamic realism is achieved with an animation technique in conjunction with a highly deformable model. To help the animation method the facial model has a high-level parameterization to control the lips, tongue, jaw, eyes and head rotation. The high-level parameterization reduces the complexity of the animation method and simplifies the description of an animation, thereby reducing the effort of the animator or director.

2.1 Skin

The skin is a protective covering of the body and defines its surface. In modeling, the skin is generally idealized as a 2D membrane. The skin of the face is loosely attached to the underlying tissues with direct muscle attachments that move the skin. These muscles allow for very subtle movements that convey important information. During speech the skin of the face is very mobile. This mobility is due to stretching, direct muscle action, and by movement of the underlying tissues. In this section the movement of the skin is discussed.



a) b)

Figure 2.1: Characteristic points used to deform the facial surface. Image a) shows the characteristic points while b) shows the control network formed by triangulating the characteristic points.

2.1.1 Characteristic Points

Characteristic points are used to reduce the complexity of the geometry. They are placed in areas that move in unison, have a high contrast in behavior, and along the borders of the articulators. A patch of skin, such as the cheeks, move very similarly so the movement of a single point can be used to represent the entire patch. The borders of these areas of similar movement must be marked, and this is done by placing characteristic points on the borders. Points that move due to movement of articulators, such as along the jawline and around the mouth are also marked as characteristic points. Figure 2.1a shows a model with its characteristic points and in Figure 2.1b the control net formed by triangulating the characteristic points can be seen. Characteristic points are interactively selected by the user by clicking

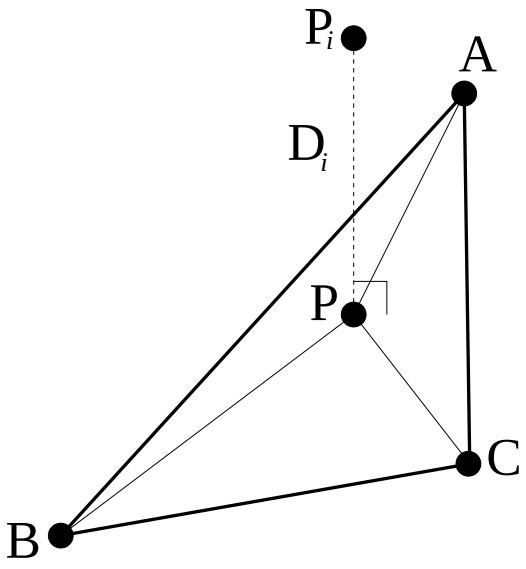


Figure 2.2: A vertex, P_i, is projected onto the plane of the triangle $\triangle ABC$ as P and the barycentric coordinates for P are determined. D_i is the offset of P_i from the surface of the plane for $\triangle ABC$.

with the mouse on their locations. The border vertices of the grafted lip model are automatically added to the set of characteristic points.

A fast method to relate changes in the characteristic points to changes in the other vertices is needed. We choose to define each vertex as a weighted combination of three characteristic points, known as barycentric coordinates. To define the barycentric coordinates, the three characteristic points must be chosen. This is done by triangulating the set of characteristic points, C_i, in cylindrical space, which is described in Section 2.1.2. The barycentric coordinates, $w_{i,j}$, for each vertex, P_i, are calculated using the control points that define the triangle that P_i intersects in cylindrical space. i is the index of vertex, and j is the index of the control points that define the triangle. Figure 2.2 shows a triangle of characteristic points A, B and C

and point P_i. P_i is projected onto $\triangle ABC$ as P. The barycentric coordinates for P are then calculated. Barycentric coordinates are weights for each vertex of the triangle and are the ratio of the area of the triangle formed by P and the other two vertices and the area of $\triangle ABC$. That is, the weight for vertex A is the area of $\triangle BCP$ divided by the area of $\triangle ABC$. The position of a non-characteristic point P_i is calculated by:

$$P_i = \sum_{j=1}^{3} w_{i,j} C_{i,j} + D_i$$

where $D_i = P_i - P$, is the displacement or offset of P_i from the plane of $\triangle ABC$. The displacement is needed since barycentric coordinates only define points in the plane defined by the three points of $\triangle ABC$. With the displacement any point in $\Re^3$ can be defined.

The control mesh of the characteristic points is determined in cylindrical space, and does not guarantee that a perpendicular projection of P_i onto the plane of $\triangle ABC$ is inside of $\triangle ABC$. Therefore, care must be taken when calculating the area of the triangle to determine the barycentric coordinates. A fast method can be found in Graphics Gems II [Gol91].

By using characteristic points the deformation of the skin surface is done in a hierarchical manner, thereby reducing the computational complexity to calculate the shape of the model. Applying barycentric coordinates to a vertex is much faster than determining the effect of all the parameters on the vertex.

2.1.2 Cylindrical space

Most problems, such as triangulation, are easier to solve in 2D than in 3D. If a mapping from $3D \rightarrow 2D$ exists then it can be used to simplify such problems. The face is shaped like an ellipsoid so a spherical mapping is possible. However, the

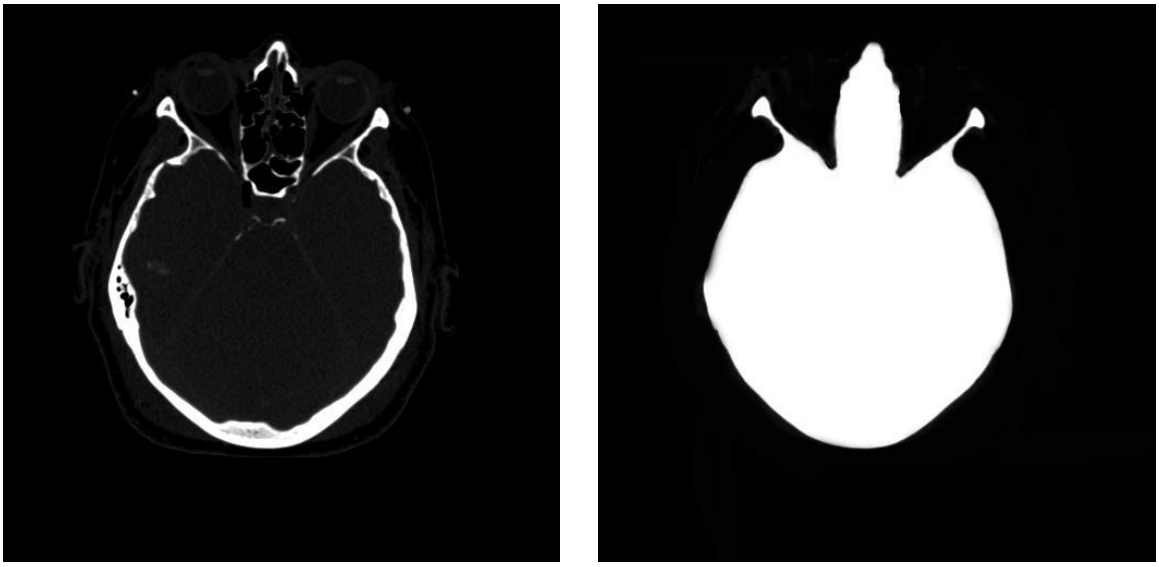
majority of the features we are interested in are in the body of the ellipse and not the ends. The top of the head and the neck are of less concern to us, allowing the use of a cylindrical mapping. This is indeed how many laser scanners work. The cylindrical mapping can fail to give us the desired results in the nose and ear area, where the surface is along the normal to the center of projection. However, for models obtained from a laser scan using a cylindrical projection, such points would not exist. Each vertex is projected into cylindrical space, (i, j), with:

$$(i, j) = \left(y, tan^{-1}\left(\frac{x}{z}\right)\right)$$

2.2 Skull

Although the skull is not visible, it has a profound influence on the shape of the skin at the surface. As the skin and muscles are stretched over the skull, they retain the skull's shape. During speech and mastication, the articulation of the mandible causes deformation of most of the face. To create realistic deformations the skull must be considered.

When animating a real person, the ideal skull would be one identical to that person's skull. Although it is possible to obtain geometry of the skull with medical imaging techniques, in most situations, a less invasive technique is required. Instead, we fit a generic skull by hand using tissue depth information [RC80, RI82] in a method similar to that of forensic facial reconstruction [KI86]. The resulting skull shape is not accurate for a specific individual, other than the person the data was taken from (in this case the Visible Human Female), but it does gives acceptable results. Using the generic skull gives acceptable results since we are only interested in



a) b)

Figure 2.3: CT images from the Visible Human Project used to create the generic skull. To the left is the original CT image and to the right is the hand-modified slice.

realistic animation, and not accurate simulation, however morphometric techniques [Boo91] could be used to achieve a more realistic skull fit.

The generic skull was created by applying marching-cubes on hand-modified CT images of the Visible Human Female [NLH99] resulting in a polygonal surface. In Figure 2.3 the original CT image is on the left and the hand-modified image is on the right. The images were filled in to achieve a one-sided surface and to ignore any of the interior features of the skull. The polygonal surface was then decimated, teeth and gums were added and the mandible was separated and made to articulate. The resulting generic skull is shown in Figure 2.4

Figure 2.4: The generic skull we fit to the input geometry.

2.2.1 Collision Detection.

The facial model includes the skull because of its effect on the resulting surface. A method to determine just when the skull will affect the surface is needed. One method is to attach the skin to the bone directly using springs to keep the skin vertices from penetrating the skull [KGC+96, LTW93]. This simplified skin model does not allow for tissue interaction with bone or with other tissue. Stretching of the skin is also restricted, as skin nodes tend to hover over the same area of bone.

General collision detection between two surfaces can be very expensive and difficult. Instead, we choose a method to avoid collisions that still allows interactions between the various tissues to occur. We use an implicit function to keep the skin at least a minimum distance away from the skull without restricting any other movement of the skin. This minimum distance is based on tissue depth information [RC80, RI82]

from the field of forensic facial reconstruction. This allows us to simulate the tissue (muscle, fat and connective tissue) between the skin and the skull. We therefore do not need to calculate actual collisions and we do not need to preserve the volume of the tissue. Our method is not accurate, but we do achieve realistic results, which is our goal.

The skull is approximated by an implicit function that is a combination of implicit primitives. The implicit function is then used to determine the distance to the surface of a vertex. Each skin vertex has a limited range of skull it can intersect and using this knowledge of the facial structure can significantly increase performance. For instance, a cylindrical arc, described below, can approximate the teeth and only those vertices of the upper chin need be tested for collision. When a collision is detected, the vertex is moved away from the bone along the normal of the control triangle by modifying the offset D_i.

An open cylindrical arc with boundary $A(E_1, E_2, r, \theta_1, \theta_2)$ is defined by the two endpoints, E_1 and E_2, of the spine of the cylinder, the radius r of the cylinder, and the beginning and ending angles, θ_1 and θ_2 of the arc as shown in Figure 2.5. The points a distance r from the spine with an angle from the spine within the range of the arc define a spherically capped cylindrical arc. By removing those points that do not intersect the line of the spine between the two endpoints, an uncapped cylindrical arc is produced. The same shape could be defined with only one angle, but may then need to be rotated about the axis of the spine.

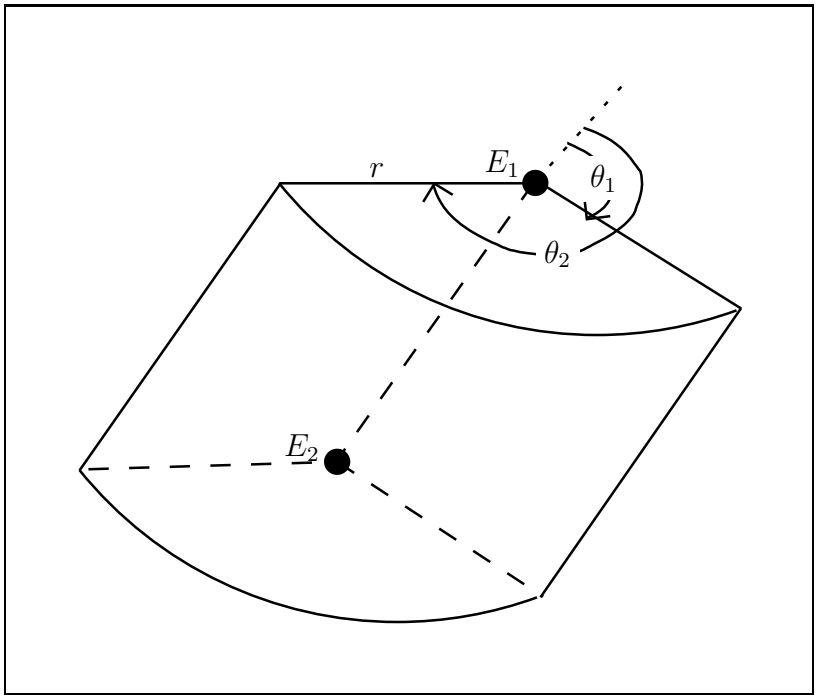


Figure 2.5: All the points a distance r from the spine defined by the endpoints E_1 and E_2 and between the angles θ_1 and θ_2 define a cylindrical arc.

2.3 Lips

Of the visual part of speech, the lips convey the most information. It is very important for both intelligibility and realism that the lips be capable of realistic shapes. To create realistic animated speech the lips must not only move realistically they must also be shaped correctly.

Additionally, when the lips part, the inner portion of lips[3] must exist. A common source of facial geometry is a laser scan of the subject. Generally, the face is scanned in a neutral position with the mouth closed. Such a scan misses the inner portion of the lips. When the mouth is opened the lack of flesh is extremely noticeable. Sometimes the subject will be scanned with the mouth open giving some of the internal lip geometry, but it cannot be guaranteed that enough geometry exists. Even if it exists there is no guarantee the lip geometry will be complex enough to create the necessary shapes for speech and facial expressions involving the lips.

To solve these problems we created a parametric model of the lips that is highly deformable, has a muscle-based parameterization, and has full lip geometry for both deformations and high quality rendering. This lip model is presented fully in Chapter 3.

2.4 Tongue

The tongue is an important articulator of the vocal tract and helps create the correct shape to produce the desired speech. Visually, it is only critical in distinguishing

[3]The inner portion of the lips is the portion inside the rima oris. The *rima oris* is the point of contact of the lips during closure.

a few sounds, such as /l/ and /th/. Although the tongue is not critical for distinguishing between all sounds, it still conveys information and the lack of a tongue is easily noted. Therefore, visually realistic speech is difficult to achieve without a tongue. Our goal of realistic animated speech requires the existence of a highly deformable tongue. To that end, we developed a parametric model of the tongue capable of representing the shapes necessary for speech as well as general facial expressions. Our tongue model is discussed in detail in Chapter 4.

2.5 Eyes

A facial model must have articulating eyes or the model will not look alive. An eyeball is a sphere with an elliptical protrusion for the lens. Unless extreme closeups are required, a sphere will suffice. The necessary motions for the eye are rotations about its center.

We model the eye as a sphere that can be rendered using a procedural shader with or without a texture map. A full texture map of an eye is difficult to obtain since most of the eyeball is not visible at any one time. In addition, since the eye is kept moist with tears, it has a very high specular component making it difficult to obtain a texture map with only ambient lighting characteristics. The shader allows for the eye to be rendered realistically regardless of lighting conditions and regardless of the direction the eye is pointing.

Our facial model has three parameters for each eye, one for rotation about each axis. In practice, a more useful parameter is a point in space that the eyes are fixated on and calculate the rotations using inverse kinematics.

2.6 Other facial parts

As stated earlier, laser scanners often miss geometry. The geometry is missing mostly due to the geometry not being visible to the sensor, except for hair, which actually dissipates the laser instead of reflecting it to the sensor. Sometimes the geometry can be specifically targeted or acquired with multiple scans. This is the case to get both the eyelids and the eyeballs. The missing geometry can be necessary to convince the audience they are looking at a human and not a computer generated image. The most common missing geometry is from the mouth, ears, nose, eyes and hair. Since we anticipate laser scanners as a likely source of input geometry it is assumed that some geometry may be missing. Since we are concerned with speech, we have concentrated on the mouth so we add a complete inner mouth to guarantee its existence. We also have added eyeballs. However, a complete solution requires the missing geometry.

Hair is still an open problem due to its complexity. There could be in the neighborhood of 100,000 individual hair strands that need to be modeled. These hair strands have anisotropic lighting characteristics, they occlude each other, and cast shadows on each other. As well, collisions between the strands and between the hair and the face must be considered. Hair also has mass and follows the laws of physics, but has highly varying physical attributes causing numerous differences in its behavior. If that is not enough, humans also use chemicals to enhance its behavior. A general solution is extremely complex and computationally intensive. Facial hair (eyebrows, mustaches, and beards) on the other hand, tends to behave very stiffly and some simplifying assumptions can be made.

2.7 Summary

To achieve realistic, intelligible, animated speech a highly deformable facial model is required. We have developed a model that supports the shapes needed for speech. The model concentrates on the mouth, as that is where the majority of the information in the visual portion of speech resides. To guarantee a mouth with sufficient geometric complexity for realistic rendering and deformations, we develop deformable parametric models of the tongue and lips and add geometry for the rest of the inner mouth.

The mouth is also capable of facial expressions such as smiling, frowning, pouting, sticking the tongue out, licking the lips, etc. The eyes are fully articulable and the head is able to rotate. Other facial expressions such as moving the eyebrows, nostrils, and eyelids can be added and are the subject of future work.

CHAPTER 3

THE LIP MODEL

3.1 Introduction

Facial animation is becoming more important as a communicative technique between man and machine. In addition, it is pivotal in the development of synthetic actors. The lips play an extremely important role in almost all facial animation. They are a significant component of expressing emotion as well as being instrumental in the intelligibility of speech. Therefore, in order to achieve realism and effective communication, a facial animation system needs extremely good lip motion with the deformation of the lips synchronized with the audial portion of the speech.

In order to animate a pair of lips, a mapping between the desired motion and lip deformations is needed. For example, a mapping between speech segments and lip shapes must exist. The mapping must then be fit to the geometry in order to be used. Possible methods to fit the mapping are: fit the mapping directly to the input geometry, fit a generic facial model with an embedded mapping to the geometry, or fit a generic lip model with an embedded mapping.

Fitting the mapping directly to the lip model requires that the input geometry has enough resolution to adequately use the mapping and must be done for every input

geometry. This can be quite time consuming. Creating a generic facial model with the mapping already embedded and then changing the shape of the prototype to match the input is a popular solution. A prototype model has the required resolution and fitting the geometry of the prototype to the desired character is less work. However, a drawback is that the resulting model retains some characteristics of the prototype after fitting.

This led us to develop a generic lip model with an embedded mapping. Using a generic lip model guarantees required resolution for both deformation and rendering, fitting the generic lip model is easier than fitting the mapping, and the traits of the generic lip model that are retained are less noticeable than those of a complete model.

Our lip model can be used with any human-like facial model. It provides:

- a sufficiently controllable model to support lip synchronization as well as supporting other motions used in expressing emotions,
- a sufficiently smooth model to support quality rendering,
- internal geometry (the part of the lips in the oral cavity not visible when mouth closed) usually not provided in digitized facial models,
- support for procedural texture maps for high quality rendering.

The lip model consists of a B-spline surface and high-level parameters which control the articulation of the surface. The high-level controls strive to be intuitive, precise, and as complete as possible. The controls are intuitive so an animator or director can easily and effectively use them. The controls are precise because of the desire to support lip-synced animation. The controls are complete in the sense that most useful positions and motions of the lips can be specified.

We use a B-spline surface for its c^2 continuity and the ease of deforming the surface by simply moving vertices of the control mesh. The drawbacks of B-splines include difficulty in placing a part of the surface exactly in $\Re^3$, preserving volume, detecting collisions and rendering. Fortunately, by polygonalizing the model, post-processing after deformations can achieve volume preservation and collision detection while rendering the polygons is straightforward. Polygonalization loses the c^2 continuity of the B-spline surface, but the quality is controllable and with Phong shading, the impact is minimal.

The lip model is fitted to the input geometry as a preprocessing step with a user guided process, shown in Section 3.5.1. This process replaces the lip region in a given facial model and grafts the generic lip model onto the rest of the facial geometry. The lip model is parameterized based on anatomy using the muscles that cause the lips to change shape as the basis. The anatomy of the lips is discussed in Section 3.3 and the parameterization in Section 3.4. As the parameters change, the lips deform which also drives deformation in the surrounding area. The formulas for calculating the change in the lip shapes are given in Section 3.5.

Our application is a text-to-audiovisual-speech system (TTAVS) that creates lip-synchronized speech from text input. Input facial geometry is transformed into a facial model that can be animated by our system allowing for the animation of any character. Animation is achieved by converting the input text into phonemes and using a mapping from phonemes into visemes. The mapping, which is a datafile, consists of keyframe poses of the facial model to achieve the desired phoneme. The model is rendered with procedural textures, described in Section 3.6, that create realistic surface detail and lighting.

3.2 Previous Work

Over the last three decades, many techniques have been used in an attempt to create convincing speech-synchronized facial animation. It has proven a difficult task due to the complexity of the system and the low tolerance for inconsistencies in the animation from a human audience. Concentration on the lips for the synchronization has been a theme, but only one research team has created a separate lip model. Generally, the concept of visemes, the visual elements of speech, is used under the guise of phonemes, the phonetic elements of speech. In this method, the facial model is deformed into a shape that represents a phoneme, which by consequence deforms the lips.

Early work in speech-synchronized facial animation involved creating animation using traditional hand-drawn animation techniques. Meanwhile, early work in the speech and hearing community involved the use of oscilloscopes to generate lip shapes. Later work involved automated techniques to synchronize the lips with the audio. In this section, we discuss the early methods of generating lips and lip synchronized animation.

3.2.1 Speech Reading

Fromkin [Fro64] reports on a set of lip parameters that characterize lip positions for American English vowels using frontal and lateral photographs, lateral x-rays, and plaster casts of lips. The lip parameters identified are:

1. Width of lip opening.

2. Height of lip opening.

3. Area of lip opening.

4. Distance between outer-most points of lips.

5. Protrusion of upper lip.

6. Protrusion of lower lip.

7. Distance between upper and lower front teeth.

This parameterization of the lips is very good for speech but it does not allow for other lip motions, such as expressive ones. We instead base our parameterization on muscle actions.

Research by the speech community on lip reading involved drawing lip outlines to represent speech. The results from these works are not as good as we would like, and our work is concerned with visual realism instead of just speech intelligibility. Boston [Bos73] reports that a lip reader was able to recognize a small vocabulary and sentences of speech represented by mouth shapes drawn on an oscilloscope.

Erber [Erb77] also uses an oscilloscope to create lip patterns and claims to create motion that is more natural. Their studies show that lipreading the display gives similar results to those of reading a face directly. Erber and De Filippo [EF78b] drew eyes, nose and a face outline on cardboard, cut out the mouth area and placed it over the oscilloscope. Erber et al. [ESF79] uses high-speed cameras to capture speech and determine lip positions by hand.

Brooke [Bro79] draws outlines of the face, eyes and nose with movable jawline and lip margins. Positional data is hand-captured from a video source. Brooke and Summerfield [BS83] report on a perception study to determine if a hearing speaker

can identify the utterances. The natural vowels were identified 98% of the time and there was good identification of the synthetic vowels, /u/ (97%) and /a/ (87%.) The vowel /i/ (28%), which is usually confused with /a/, and medial consonants were not identified.

Montgomery [Mon80] draws lip outlines on a CRT from data hand captured from video frames in a system designed to test lip reading ability. They augment by adding nonlinear interpolation between frames as well as forward and backward coarticulation approximation.

3.2.2 Computer Models

In Parke's [Par74] ground-breaking work, he uses 10 parameters to control the lips, teeth, and jaw during speech synchronized animation. The parameters are chosen empirically and from traditional hand drawn animation methods. These parameters give control over the lip motion, but have little consideration over lip shape, and the lip motion is limited.

Bergeron [BL85] reports that for the film short "Tony De Peltrie" keyframes were defined and then interpolated using curves. The keyframes included those for phonemes. This method of shaping the lips for each phoneme is a common approach. However, it is difficult to create a complete set of poses that encompass the entire space of motion desired. For example, to have a pose of smiling and saying /a/ at the same time requires having to create that specific pose. By defining shapes that differ from the neutral in a single way, such as smiling, these shapes can be used to define a space of possible shapes. Each base shape is calculated as a displacement [Ber87] from the neutral. By defining a percentage of each base shape, animations

that are more complex can be achieved. Our work is similar to this concept; however, we choose our bases using muscle displacements instead of phonemes and expressions, which we believe gives a larger space of possible lip shapes. Another difference is the method we use to combine the displacements.

3.2.3 ICP Lip Model

Guiard-Marigny [GM92] measures the lip contours of French speakers articulating 22 visemes in the coronal plane. Assuming symmetry, the vermillion region (the red area of the lips) of the lips is split into three sections and mathematical formulas are created to approximate the lip contours. From polynomial and sinusoidal equations, the 14 coefficients are reduced to 3 using regression analysis. The three parameters are internal lip width, internal lip height and lip contact protrusion. With the same technique on lip contours in the axial plane, Adjoudani [Adj93] identifies two extra parameters to extend the lip model to 3D. The two new parameters are upper and lower lip protrusion. Guiard-Marigny et al. [GMAB94] describe the 3D model in English. This work was carried out at the Institut de la Communication Parlée in Grenoble, France and we refer to the model as the ICP lip model.

Guiard-Marigny et al. [GMTA+96] replace the polygonal lip model with an implicit surface model that uses point primitives to achieve fast collision detection and exact contact surfaces. For the polygonal model, to increase the speed of computation they define a lip shape as the barycenter of a set of extreme lip shapes. The parameters correspond to the weights and each parameter interpolates two extreme shapes. They build the implicit surface from point primitives for each of the 10 key (extreme) shapes. Any lip shape is found by interpolating the point primitive positions and

field functions. Implicit surfaces give an exact contact surface [Gas93], allowing the interaction of the lips with other objects, for example a cigarette.

We originally used the ICP lip model using interpolation of the 10 extreme shapes, in our TTAVS system. The ICP lip model was designed by analyzing speech and is only capable of representing lip shapes used during speech production. We desire a model capable of expressing at least simple emotion such as smiling. The ICP lip model also lacks visual realism while producing rounded lip positions, such as for the phonemes /o/ and /r/, because the corners do not move correctly. While it was possible to significantly modify this model, it seemed more prudent to start from scratch, creating an anatomically based model that was parameterized and deformed based on muscle actions.

3.3 Lip Anatomy

The face is a biological system and the lips are deformed because of muscle contraction. We look at the anatomy of the lips and the underlying muscles to define possible lip motion. Any good general anatomy reference (e.g. Gray [Gra77]), speech and hearing anatomy reference (e.g. Dickson and Maue [DM70] or Bateman and Mason [BM84]) or facial anatomy reference (e.g. Brand and Isselhard [BI82]) is a beneficial source of information. In this section, we give a description of the muscles that affect the lips as well as a description of the motion of the mandible.

Table 3.1 lists the muscles that affect the lips with a brief description of their actions. All muscles that affect the lips, except the orbicularis oris, are paired, that is, there is a corresponding muscle on the right and left side of the face with the same name.

Name	**Action**
Buccinator	Compresses the cheek against the teeth, retracts the corner of the mouth.
Depressor anguli oris	Draws the corner of the mouth downward and medialward.
Depressor labii inferious	Depresses the lower lip.
Incisive inferior	Pulls the lower lip in towards teeth.
Incisive superior	Pulls the upper lip in towards teeth.
Levator anguli oris	Moves the corner of the mouth up and medially.
Levator labii superioris	Raises the upper lip and carries it a little forward.
Levator labii superioris alaeque nasi	Elevates the upper lip and the wing (ala) of the nostril, while aiding in forming the nasolabial groove.
Mentalis	Raises and protrudes the lower lip.
Orbicularis oris	Closes the lips, compresses the lips against the teeth, and protrudes the lips.
Platysma	Pulls the corner of the mouth down and back.
Risorius	Pulls the corner of the mouth back.
Zygomaticus major	Draws the corner of the mouth laterally and upward.
Zygomaticus minor	Draws the outer part of the upper lip upward, laterally and outward.

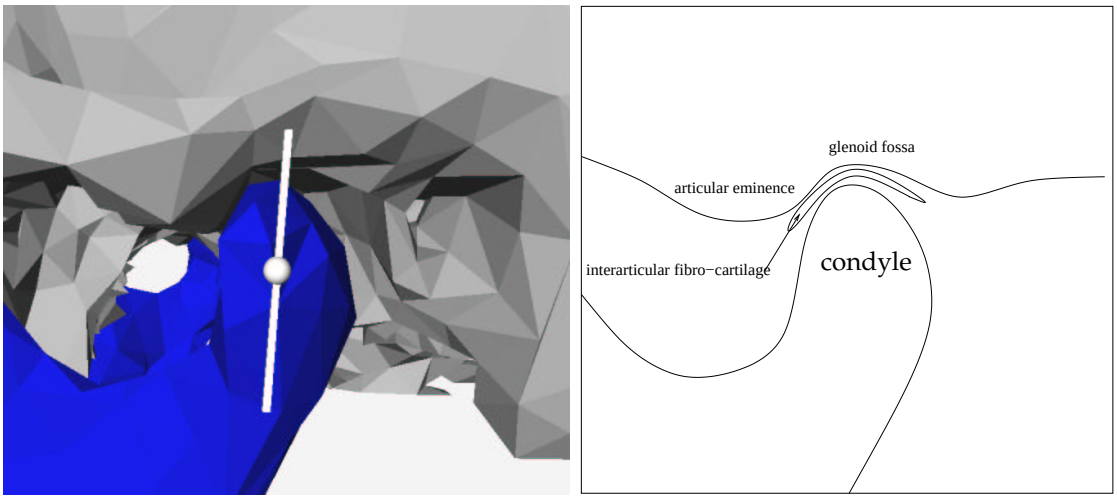
Table 3.1: Muscles that directly affect the lips and their actions.

3.3.1 The Mandible

The lips are tightly connected to muscle, with some muscles attached to the mandible, such as the depressor labii inferioris, so when the mandible moves so do the lips. In order to properly control lip deformations, the movement of the mandible must also be considered.

The temporomandibular joint is a diarthrodial ginglymous [NS62] (sliding hinge) joint that allows the mandible a large scope of movements. The mandible acts mostly like a hinge, with two separate joints acting together, and each of the joints having a compound articulation. The first joint is between the condyle and the interarticular fibro-cartilage while the second is between the fibro-cartilage and the glenoid fossa. The condyle of the mandible is nested in the mandibular (or glenoid) fossa of the temporal bone as seen in Figure 3.1. The upper part of the joint enables a sliding movement of the condyle and the articular disk, moving together against the articular eminence. In the lower part of the joint, the head of the condyle rotates beneath the under-surface of the articular disk in a hinge action between the disk and the condyle.

The mandible can be thought of as a 3 DOF joint, with the jaw moving in (retraction) and out (protraction), side-to-side (lateral movements), and opening and closing. The mandible lowers in a hinge-like manner except in extensive openings when the movement is downward and forward. During protraction and retraction, the condyles slide on the articular eminences and the teeth remain in gliding contact. Lateral movement is achieved by fixing one condyle and drawing the other condyle forward.



a b

Figure 3.1: The temporomandibular joint a) shown on our 3D geometry of the skull and b) a cross-sectional drawing of the joint showing the bony parts of the joint as corresponding to a). The mandible is shaded differently from the skull and the condyle on the right side of the skull can also be seen.

3.4 Lip Parameterization

Parameterizing the motion of the lips allows us to reduce the number of degrees of freedom of the system. The goal is to minimize the number of degrees of freedom while still providing flexibility and generality. Besides a minimal set, we need a parameterization for the lip motion that is intuitive to use; easily defined and modified for different mouths; and supports speech synchronization and the wide range of other lip motions needed for facial animation.

The lips deform due to the contraction of the connected muscles and the movement of the mandible. We use the muscles that affect the lips as the basis for our parameterization resulting in anatomy-based deformations. Movement of the mandible is followed closely by movement of the lower lips causing the upper lips to also change

shape. Therefore, the parameterization must also include mandible movements. The muscles are a good choice because their action is mostly along a vector, so their effect on the lips can be easily defined. This works for all muscles but the orbicularis oris, which actually constricts and protrudes the lips. We have a parameter for each muscle that inserts into the lips. For muscles that are paired we have a separate parameter for the left and right counterparts. Exceptions are made for the depressor inferioris, depressor oris and mentalis since individual control is rare. The incisive inferior and incisive superior are also treated as a single muscle. Some texts consider these part of the orbicularis oris, but the action of pulling the lips in tight against the teeth is needed as a separate degree of freedom from that of the orbicularis oris. Lastly, we treat the levator labii superioris and the zygomaticus minor as a single muscle since the zygomaticus minor is usually not well developed and their actions are very similar.

The 21 parameters we use for our lip model along with their definitions are:

- **Open Jaw** - The mandible rotating open which lowers the lower lip and the corners of the mouth causing the lower lip to rotate out slightly.
- **Jaw In** - Movement of the mandible superiorly and inferiorly causing the lower lip to move inward or outward.
- **Jaw Side** - Lateral movement of the jaw causing the lower lip to skew to one side or the other.
- **Orbicularis Oris** - Contraction of the orbicularis oris muscle causing the lips to pucker and protrude.

- **Left Risorius** - Contraction of the left risorius muscle pulling the left corner of the mouth back.
- **Right Risorius** - Contraction of the right risorius muscle pulling the right corner of the mouth back.
- **Left Platysma** - Contraction of the left platysma muscle pulling the left corner of the mouth down and back.
- **Right Platysma** - Contraction of the right platysma muscle pulling the right corner of the mouth down and back.
- **Left Zygomaticus** - Contraction of the left zygomaticus muscle pulling the left corner of the mouth up and back.
- **Right Zygomaticus** - Contraction of the right zygomaticus muscle pulling the right corner of the mouth up and back.
- **Left Levitator Superior** - Contraction of the left levator labii superioris muscle raising the left part of the upper lip.
- **Right Levitator Superior** - Contraction of the right levator labii superioris muscle raising the right part of the upper lip.
- **Left Levitator Nasi** - Contraction of the left levator labii superioris alaeque nasi muscle raising the left part of the upper lip as well as the wing of the left nostril.

- **Right Levitator Nasi** - Contraction of the right levator labii superioris alaeque nasi muscle raising the right part of the upper lip as well as the wing of the right nostril.
- **Depressor Inferious** - Contraction of both depressor inferious muscles depressing the lower lip.
- **Depressor Oris** - Contraction of both depressor anguli oris muscles drawing the corners of the mouth downward and medial-ward.
- **Mentalis** - Contraction of both mentalis muscles raising and protruding the lower lip, a paired muscle but treated as a single muscle.
- **Left Buccinator** - Contraction of the left buccinator muscle retracting the left corner of mouth and keeping cheeks taut against teeth.
- **Right Buccinator** - Contraction of the right buccinator muscle retracting the right corner of mouth and keeping cheeks taut against teeth.
- **Incisive Superior** - Contraction of both upper incisive muscles pulling the upper lip in towards teeth.
- **Incisive Inferior** - Contraction of both lower incisive muscles pulling the lower lip in towards teeth.

An added benefit of using a muscle-based parameterization is that the muscles also affect other parts of the face and the parameters can be used to deform these other parts as well. Examples are nose wrinkling, platysma affecting the neck, mentalis affecting the chin, the zygomaticus affecting the lower eyelid, and so forth. As well, when the muscles contract they bulge, which affects the surface of the face.

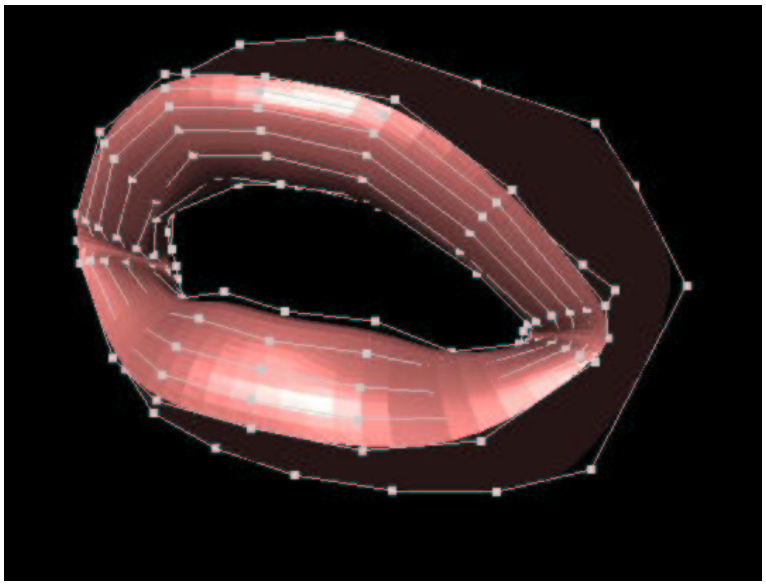


Figure 3.2: The B-spline control mesh and the surface used for the geometry of our lip model.

Higher-level parameterizations that use the basic parameters can be built to allow for easier use by an animator or director, such as smile, frown, pout, etc.

3.5 Implementation

We represent the lips as a B-spline surface with a 16x9 control grid. The parameters itemized above are mapped to changes in the positions of the control grid vertices. The geometry contains all of the vermillion zone (the red area of the lips) as well as the part of the mucous membrane that covers the lips internally. The geometry also contains a little extra of the mucous membrane to avoid observing an edge when looking at the lips from the outside. Figure 3.2 shows the control points of the lip model along with a polygonalization of the B-spline surface.

All of the muscles except the orbicularis oris are treated as vector displacements acting upon its insertion points. The orbicularis oris constricts the shape of the lips into an oval while also extruding them. The parameters for the jaw articulate a virtual mandible based on the three jaw parameters and the resulting transform is used to move the lower lips. The lower lip is rotated outward with the opening of the jaw as well.

For each control point, we calculate its position based on the parameters by the following:

$$p_i = O_i(\hat{p}_i + L_i + J_i + A_i)$$

where $\hat{p}_i$ is the starting value for control point i, L_i is the contribution of the linear muscles, J_i is the contribution of the jaw, O_i is the contribution of the orbicularis oris and A_i is the adjustment due to the interconnection of the control points.

The contribution from the linear muscles involves summing the displacements from all of the individual muscles and is calculated by

$$L_i = \sum_{j=0}^{m} \rho_j M_j \delta_{ij}$$

where ρ_j is the parameter value for muscle j, M_j is the maximum displacement for muscle j, and δ_{ij} is the influence of muscle j on control point i. $\delta_{ij} = 0$ when the muscle has no influence, $\delta_{ij} = 1$ when the muscle is inserted very near the control point, and $0 < \delta_{ij} < 1$ is used to create a zone of influence for the muscle. This is used for the upper lip only, as the lower lip moves more as a unit. δ comes into play particularly in the middle of the upper lip with low values for a stiff upper lip, as in most males, and high values to show more gum for a feminine appearance.

When the jaw moves it also moves the lower lip due to direct muscle attachment. The effect of jaw movement on the lips is calculated by

$$J_i = J_{open} + J_{in} + J_{side}$$

where J_{open} is the rotation about the axis through condyles, J_{in} is the movement of jaw in or out, and J_{side} is the lateral movement of the jaw.

The control points represent locations on the lips. The lips are made of muscle fibers that can stretch slightly but will maintain a mostly constant circumference. To keep the lip model from violating this property the control points are adjusted after muscles forces using

$$A_i = LD\alpha_i + \rho_{open}\gamma_i$$

where LD is the motion vector for the lower lip, α_i is how much lower lip affects the upper lip, ρ_{open} is the parameter value for the jaw begin open and γ_i is the affect of tightening lips. The lower lip moves mostly uniformly and individuals rarely have control over it. LD is the lower delta and represents the movement of the lower lip. As the lower lip moves, it pulls on the corners of the mouth and therefore the upper lips. The α weights take care of this affect. As the mouth opens, the lips stretch and tighten. As they tighten, they move medially toward the mouth center. The γ weights allow for this medial motion.

The orbicularis oris muscle constricts and protrudes the lips as it contracts. This effect is handled after all the other displacements are taken into account. It is done this way to make combining the muscle displacements less complex. The linear displacements are additive with maximum values for the displacements. However, the

orbicularis oris causes complex motion and does not simply add to the other displacements. The contribution of the orbicularis oris is calculated as

$$O_i(p) = R(\rho_{oris}\theta) + \rho_{oris}[e_i(p) + \chi_i]$$

where ρ_{oris} is the parameter value for the orbicularis oris, θ is the maximum angle of rotation from puckering the lips, $R(\theta)$ is the rotation due to contraction of the orbicularis oris, $e_i(p)$ keeps the point p on the ellipse created by the lips, and χ_i is the maximum extrusion from contraction of the orbicularis oris.

The weights and muscle displacement vectors are data to the lip model allowing the behavior of the lips to be changed by simply changing data files. Besides the different geometry, a different character will potentially have a different datafile for the lip model behavior. It may also be desirable to change the lip behavior for the same character such as modeling a muscular dysfunction on one side of the face.

An option would have been to calculate the forces of each muscle and using a Newtonian physics model, numerically solve the ODEs to find the new locations of the control points. This would have allowed us to constrain the lip shape using springs, but we would have had to solve the ODEs. We wanted a closed-form solution that would avoid the rubbery look of spring-based systems. Moreover, using implicit integration would add unnecessary complexity.

3.5.1 Grafting

Grafting of the lip model geometry onto the input face geometry is done interactively. First, an interactive tool is used to align the lip model with the input geometry. The tool allows the lip model to be shaped like the input lips in a neutral position. Figure 3.3b shows the lips fitted to the input geometry.

A convex outline surrounding the lips of the input geometry is created. After applying a cylindrical projection, all vertices, and thus all triangles, inside the convex hull are removed thereby removing the input lips shown in Figure 3.3c.

The fitted lip model is polygonalized and the outline of the lip model is triangulated along with the remaining input geometry using a Delaunay triangulation method [She96b, She96a] as shown in Figure 3.3d.

The new triangles are added, along with the lip model, to the input facial geometry, effectively replacing the input lips with the lip model geometry. This can be seen in Figure 3.3e and Figure 3.3f.

3.6 Rendering

In order to create realism, the rendering of the lips is important. A common method to improve realism is the use of texture maps. The same problems associated with gathering the geometry of the lips also exist for gathering color information. Incomplete texture information will leave visible artifacts. We could use methods to warp what texture information is obtained, but there is no clear-cut way to do this. This would also exacerbate the problems associated with texture maps, such as limited resolution and lighting inherent in texture acquisition.

We instead choose a different approach. We use a procedural texture shader to increase realism. Besides color information, we can also add surface detail with a bump shader. Figure 3.4 shows examples of wrinkled lips, both dry and wet. This method only works for off-line generation of animations since it is too slow for our real-time version, where we choose a single color for the lips.

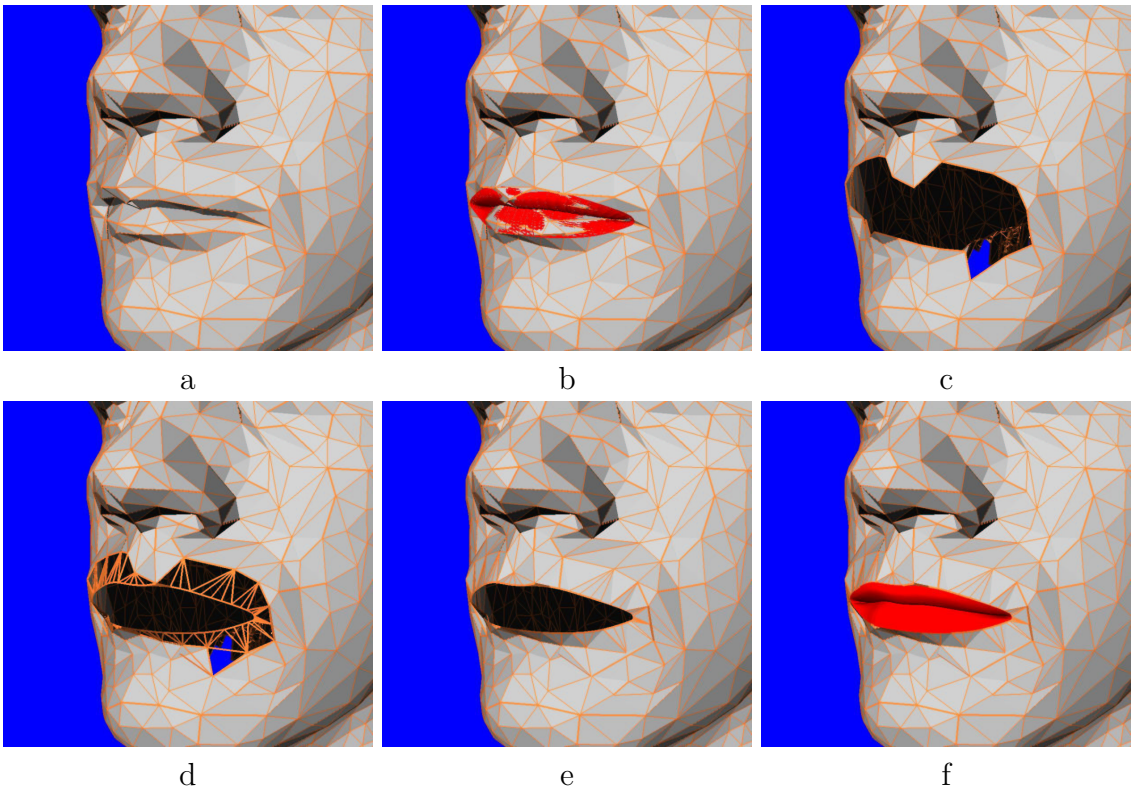


Figure 3.3: The grafting process. The input geometry (a) is used to fit the lip model (b). Overlapping triangles are then removed (c). Then the boundary of the lip model and the boundary of the removed triangles are retriangulated (d) and the new triangles are added (e). Finally the lip model geometry is added to the input geometry (f).

Figure 3.4: Rendering of the lip model using our custom shaders. The left image is of wrinkled dry lips, and the right of wrinkled wet lips.

Lips are covered with very thin skin that tends to wrinkly easily. Besides the constant fine to medium wrinkles, when the lips are compressed (as in a pucker) there are large undulations of the surface. We currently ignore the finer wrinkles and instead concentrate on the larger wave-like wrinkles created during compression. Wrinkles are implemented as a bump shader.

Another shader determines the color of the lips. We can simulate natural lip colors as well as lipstick and lipgloss. When the lips are licked, differing depths of saliva are deposited across the lips. We model this affect by creating a second layer, using a noise function, which represents the wetness pattern. This pattern is then mixed with the current lip color to increase the specular component. Lipstick and lipgloss are implemented as a uniform color change across the lips with transparency and glossiness components controlling matte versus glossy. Flecked lipstick is modeled by adding a flecked silver pattern to the lipstick color.

3.7 Results

We have successfully incorporated this lip model into the facial model used by our TTAVS system. Our TTAVS system creates animations from text creating a stream of visemes, or keyframes, that are interpolated. Figure 3.5 shows our lip model grafted onto an input geometry and displayed in our real-time system. Flat shading is used to more easily see the graft.

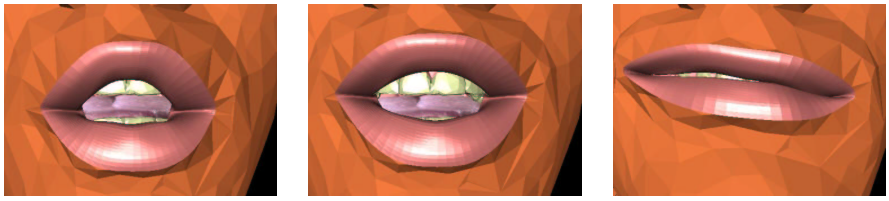
Figure 3.6 made up of frames from an off-line rendering using our rendering process for the lips. With our rendering technique, we can achieve wrinkled and wet lips for increased realism. Figure 3.7 depicts frames from an off-line rendering and demonstrate motion blur of the lips, which can move extremely fast during speech.

Figure 3.5: The lip model grafted onto face geometry. This is from our real-time TTAVS system using OpenInventor to render the frames.

Figure 3.6: Frames from an animation rendered using our lip shaders.

Figure 3.8 shows closeups of the mouth area from our TTAVS system. Figure 3.8a is the viseme /aw/, while Figure 3.8b is the viseme /aw/ while also activating the zygomaticus major muscle creating a happy /aw/. Figure 3.8c is a half smile, created by activating only the right zygomaticus major.

Figure 3.7: Frames from an animation rendered using our lip shaders.



(a) (b) (c)

Figure 3.8: Our lip model in the position of /aw/, a happy /aw/ and a half smile.

3.8 Summary

Our anatomically based lip model improves our ability to create realistic speech-synchronized facial animation with more realistic deformations of the lips. Because it is muscle-based, the effects of contraction of the muscles that affect the lips on other parts of the face are more easily calculated. Our lip model has both internal and external lip geometry, and by replacing the input lip geometry with the lip model's geometry, we are guaranteed to have the internal geometry. The internal geometry is often missing, especially when the input geometry is acquired via a laser scan of the subject. This internal geometry is important to have when the mouth opens to avoid

loss of realism. With our generic lip model that is fitted to the subject, we also do not need to redefine the insertion of the muscles for each new subject.

Our current focus has been on the motion of the lips due to muscle contractions, however, we also need to consider deformations due to collisions between the lips and other parts of the face. The lips must flow around the teeth and not penetrate them. Furthermore, when the tongue presses against the lips for creating sounds or when wetting them, a slight deformation is needed to improve realism. Finally, when the upper and lower lips come into contact there are subtle changes that need to be shown. However, these deformations can be done without collision detection between the lips. In addition, the spatial relationship between the upper and lower lips makes interpenetration hard to notice.

Our lip model does not have a concept of state, that is to say, they do not know what came before. Therefore, certain positions cannot be distinguished without further information. For example, to rotate the lower lip outward into a pout, the lower lip is pushed upward toward the upper lip, which is tensed, causing the lower lip to slide over the upper lip and outward. However, if the upper lip is not tensed it will be pushed upward by the lower lip. These two distinctly different positions can have the same parameter values and cannot be distinguished without the concept of state. We want the parameter set to map to a single shape so we do not use state. Adding new parameters can substitute for state, but requires a parameter to handle each of the special cases. For speech this does not come into play. As future work, the best of the two options needs to be determined.

Our rendering technique gives us improved realism by allowing control of the surface detail as well as lighting, particularly highlights due to wetness of the lips.

Unfortunately, lip wetness is currently applied to the entire lip as we do not have a way to specify to the shader how far the tongue has moved across the lip, how high it has gone on the lip as it moved, and in what direction it is moving. Changing this parameter to a function which tracks the aforementioned state changes and leaves a trail of saliva behind would allow better control. We would also like to be able to represent chapped lips as well as finer wrinkles. A better model of the mucous membrane is also desirable. This section of the lip is constantly moist and has significantly fewer wrinkles than the external part. The rima oris (point of contact of the lips during closure) in our model is simply a line in texture space, but in reality should be a curve over which there is a smooth transition between the two sections.

CHAPTER 4

THE TONGUE MODEL

4.1 Introduction

For nearly thirty years, computer graphics researchers [Par72, SMP81, Wat87] have been working on synthetic talking heads. Computer generated talking heads are useful for human-computer interfaces, entertainment, training, automated kiosks, and many other applications. During this period, the tongue has received very little attention. A common justification for this oversight is that the tongue has limited visibility, only plays a role in distinguishing a few speech segments and it is an extremely complex organ. However, our experience during development of a text-to-audiovisual-speech (TTAVS) system [KP00] has indicated that the tongue is very important and should be modeled.

Initially, we implemented a simple tongue model and the addition of this rigid box-like tongue increased the intelligibility of the speech, but the unnatural appearance of the tongue was quite noticeable. To achieve increased intelligibility and to have natural looking animation, a physically realistic tongue model was needed. We required a tongue model that could represent the myriad different shapes of the tongue during speech, was computationally efficient, and had a parameterization with intuitive

control. To that end, we developed a geometric model of the tongue, discussed in Section 4.4.1, capable of representing a highly deformable structure. We also devised a parameterization for the tongue, presented in Section 4.4.2, that requires only six parameters to describe the tongue shape and can be used with any tongue model. We implement a tongue model, discussed in Section 4.4.3, that allows realistic animation of the tongue and uses our parameterization. For improved visual realism of the tongue, surface and lighting detail are added with a procedural shader, which is described in Section 4.5.

4.2 Previous Work

Several researchers recognized the need to represent the tongue in a facial animation system. These systems have used a range of methods to simulate the tongue including a simple geometric tongue with rigid motion, a human sculpted tongue in keyframe positions, finite element modeling, and a highly complex model using soft objects.

Some of the first uses of a tongue in computer facial animation were in the creation of animated shorts. Reeves [Ree90] describes the use of a teardrop shaped collection of 12 bi-cubic patches to model the tongue in "Tin Toy", Pixar's Academy Award winning animated short film. Although the tongue was modeled, it was usually left in the back of the mouth. Klieser [Kle89] sculpts a face in phonemic positions and then interpolates between them, paying particular attention to the tongue, as well as the eyes, teeth, eyebrows and eyelashes.

In their work with the deaf and speech reading, Cohen and Massaro [CM93] modify the Parke model [Par74] to work with phoneme input and add a simple tongue model

to increase intelligibility of the generated speech. The tongue model they use is quite simple and animates very stiffly with parameters for tongue length, angle, width, and thickness. The limitations of a simple, rigid tongue are quite apparent during excited conversation when the mouth is opened wide and during non speech-related animation such as licking the lips or sticking out the tongue.

In studying the tongue as part of the vocal tract, Stone [Sto91] describes problems with a 3D model of the tongue and proposes a preliminary model. The model divides the tongue into five lengthwise segments: anterior, middle, dorsal, posterior, and root; and five crosswise segments: medial, left lateral1, left lateral2, right lateral1 and right lateral2. This gives the model 25 functional segments to be able to reproduce 3D tongue shapes.

Later, Stone and Lundberg [SL96] reconstruct 3D tongue surfaces during the production of English consonants and vowels using a developmental 3D ultrasound device. Electropalatography (EPG) data was collected providing tongue-palate contact patterns. Comparing the tongue shapes between the consonants and vowels reveals that only four classes of tongue shapes are needed to classify all the sounds measured. The classes are front raising, complete groove, back raising, and two-point displacement. The first three classes contain both vowels and consonants while the last consists of only consonants. The EPG patterns indicate three types of tongue-palate contact: bilateral, cross-sectional, and a combination of the two. Vowels use only the first pattern and consonants use all three. There was an observable distinction between the vowels and consonants in the EPG data, but not in the surface data. Stone concluded that the tongue actually has a limited repertoire of shapes and positions them against

the palate in different ways for consonants versus vowels to create narrow channels and divert airflow for the production of sounds.

Maeda [Mae90] creates an articulatory model of the vocal-tract that uses four parameters for the tongue by studying 1000 cineradiographic frames of spoken French. The four parameters are jaw position, tongue-dorsal position, tongue-dorsal shape and tongue-tip position. The focus of his research is gross vocal-tract shape and is not directly applicable to generating 3D shapes.

Research by Wilhelms-Tricarico [WT95] on a physiologically based model of speech production led to the creation of a finite element model of the tongue with 22 elements and 8 muscles. This initial model does not completely simulate the tongue but shows the feasibility of the methods. In later research, Wilhelms-Tricarico and Perkell [WTP97] create a way to control the model. This model is still under development, is quite computationally expensive, and is difficult to control, limiting its usefulness for facial animation.

Pelachaud et al. [PvOS94] develop a method of modeling and animating the tongue using soft objects, taking into account volume preservation and penetration issues. This is the first work on a highly deformable tongue model for the purpose of computer animation of speech. Based on the research of Stone [Sto91], Pelachaud creates a tongue skeleton composed of nine triangles. The deformation of the skeletal triangles is based on 18 parameters: nine lengths, six angles, and a starting point in 3-space.

During deformation, the approximate areas of the triangles are preserved and collision with the upper palate is detected and avoided. The skeletal triangles are considered charge centers of a potential field with an equi-potential surface created

by triangulating the field using local curvature. Their method allows for a tongue that can approximate tongue shapes during speech, as well as effects on the tongue due to contact with hard surfaces. Drawbacks of this method include the computationally expensive rendering and processing plus the large number of parameters.

4.3 Tongue Anatomy

The tongue consists of symmetric halves separated from each other in the mid line by a fibrous septum. Each half is composed of muscle fibers arranged in various directions with interposed fat and is supplied with vessels and nerves. The complex arrangement and direction of the muscular fibers gives the tongue the ability to assume the various shapes necessary for the enunciation of speech [Gra77].

The muscles of the tongue are broken into two types: intrinsic and extrinsic. The extrinsic muscles have external origins and place the tongue within the oral cavity. The intrinsic muscles are contained entirely within the tongue with the purpose of shaping the tongue itself. The extrinsic muscles are the styloglossus, hyoglossus, genioglossus, palatoglossus, and chondroglosus; the intrinsic muscles are the superior lingualis, inferior lingualis, transverse lingualis and vertical lingualis. The superior lingualis is the only unpaired muscle of the tongue [Gra77, BM84].

Figure 4.1a shows a mid-sagittal view of the tongue taken from the Visible Human Project [NLH99] data. The image shows the complexity and large size of the tongue. Typically, only the top and the tip of the tongue are visible, giving the false impression of a thin pancake-like organ. There are also vast differences in tongue shapes between individuals as can be seen from Figure 4.1c, which is a similar view of the Visible Human Female.

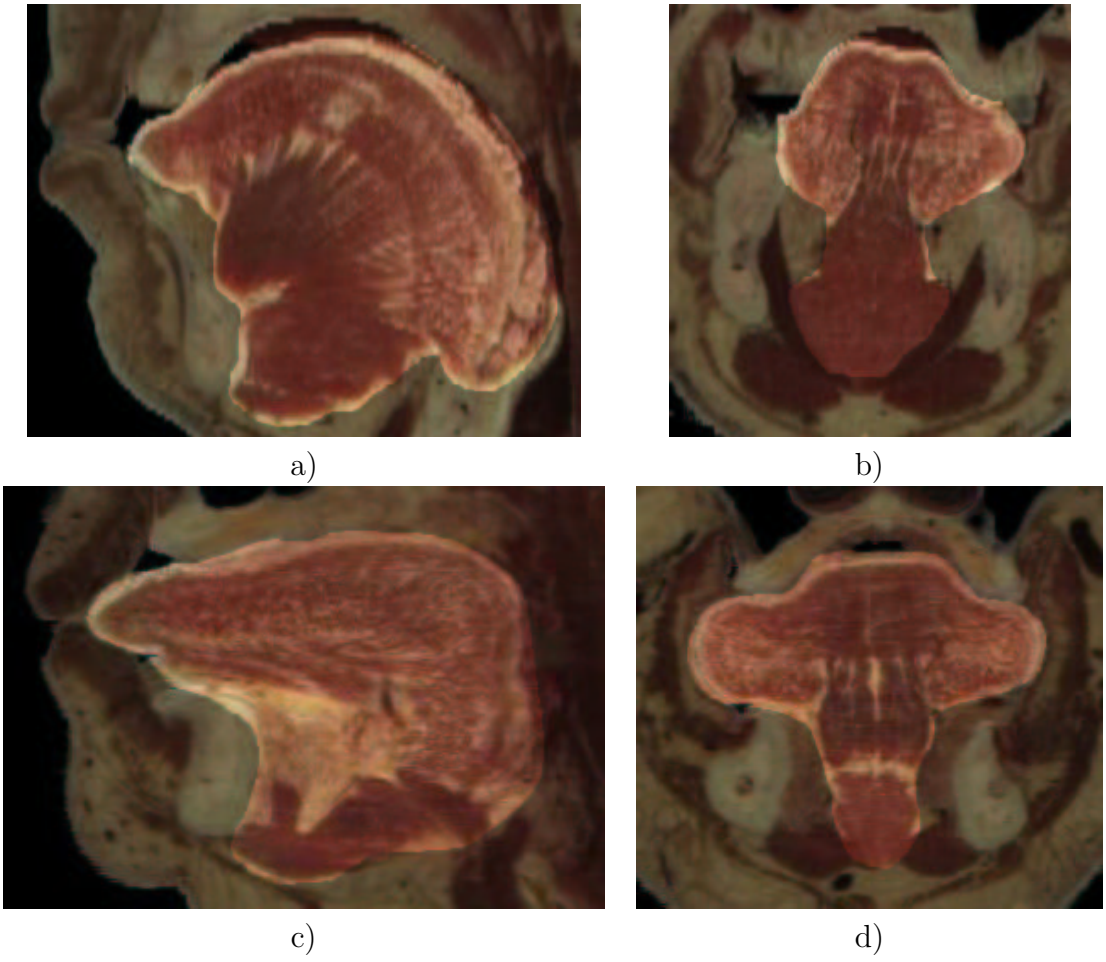


Figure 4.1: Tongue images from the Visible Human Project. a) A mid-sagittal view of the Visible Human Male. b) A coronal cross-section of the Visible Human Male. c) A mid-sagittal view of the Visible Human Female. d) A coronal cross-section of the Visible Human Female. The tongue has been highlighted in these images to better distinguish the tissue of the tongue.

Stone and Lundberg [SL96] observe that the tongue is a complex system with a fixed volume and made entirely of muscle, and its shape is systematically related to its position since the tongue volume can be redistributed but not changed.

The shape of the tongue is also influenced by contact with hard surfaces such as the hard palate and the teeth [SL96] as evidenced in Figure 4.1. Figure 4.1b depicts a coronal cross-section of the Visible Human Male while Figure 4.1d is a coronal view of the Visible Human Female. Notice that the male has a higher hard palette allowing the tongue to expand superiorly. The male's teeth on the left compress the tongue, while on the right side without teeth the tongue is allowed to expand. Also note in the sagittal views, Figures 4.1a and 4.1c, that without incisors the female tongue extends beyond the mandible.

4.4 Tongue Model

Our parameterized tongue model consists of geometry that defines the surface of the tongue, a parameterization that defines a space of possible tongue shapes, and a mapping that deforms the geometry based on the values of the parameters. Our tongue model is used to improve the intelligibility of the speech and to increase the realism of facial animation. Our application is a real time text-to-audiovisual system with requirements, in decreasing order of importance, of being:

- fast to render and deform,
- realistic in shape,
- capable of any tongue shape needed for speech,
- capable of a wide array of tongue shapes needed for general facial animation,

- able to support collision detection between the tongue and the rest of oral cavity,
- and able to preserve volume during movement and collision.

The first two requirements are addressed by defining a B-spline surface. To meet the next two requirements, we develop a tongue parameterization that we discuss in Section 4.4.2 below. We use this parameterization to abstract the specification of the tongue shape away from the actual geometry to make it easier to specify a particular tongue shape. Unfortunately, in order to meet real-time speeds, the final two requirements could not be met in our model due to their computational requirements.

A realistic tongue model should be capable of approximating the shape of any tongue in any natural position. For applications that need a faithful representation of the tongue, this is extremely important. For facial animation, that restriction can be relaxed somewhat. It is only important for those parts of the tongue that are clearly visible to be accurate. This includes the tip, the top from the front to the middle and the front portion of the underside. Many of the complex interactions between the tongue and hard surfaces are mostly invisible and can be ignored. Those that are visible are generally short-lived, allowing approximation methods to be used. The implementation we describe in Section 4.4.3 is only concerned with the first four requirements, and the final two are subjects of future research.

4.4.1 Geometry

B-spline patches result in a smooth surface that can be quickly deformed by displacing the control points. We created a B-spline surface capable of producing realistic tongue shapes. It is composed of an 8 x 13 grid of bi-cubic patches over 60 control

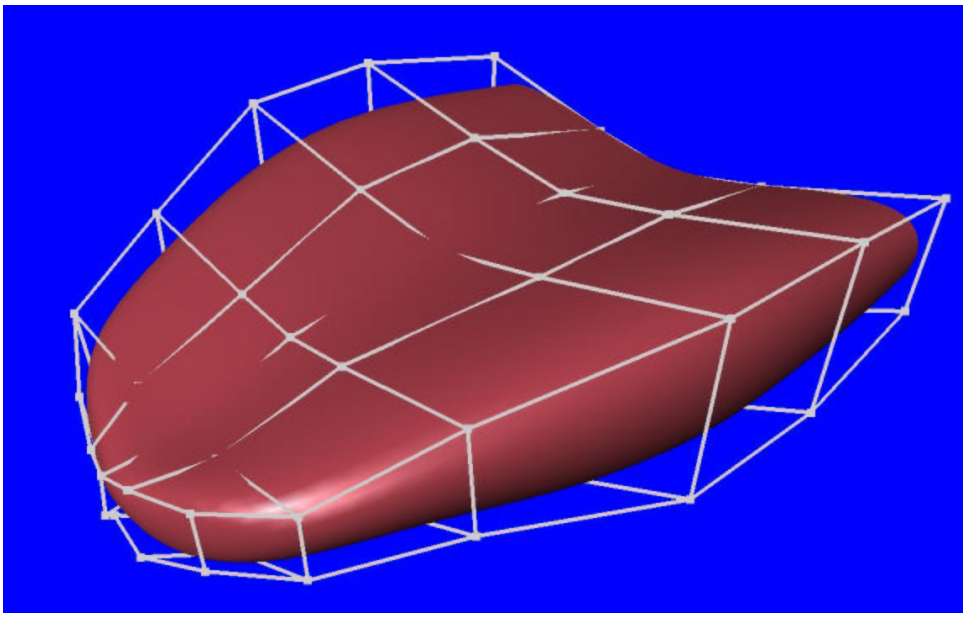


Figure 4.2: The tongue geometry and its control grid.

points. The control points, $P_{i,j}$, are arranged in a 6 x 10 grid, shown in Figure 4.2, with 6 control points along each column i representing the u parametric direction and located sagittally along the tongue from front to back. The rows j, represent the v parametric direction with 5 rows for the top of the tongue and 5 rows for the underside. Conceptually, the grid forms a rectangular parallelepiped with a cap at one end. Defining the bi-cubic patches over this grid creates two extraordinary points (with order three) at the tip, causing an acceptable loss of c^2 continuity.

This B-spline surface is capable of approximating any non-intersecting tongue shape except for a pronounced medial groove with steep walls. With the addition of another 20 control points and using an 8 x 17 grid however, the groove could also be approximated. A more complex grid, or a different choice of surface representation such as triangular or hierarchical [FB88] splines, can reduce the number of patches

and control points. For applications that need more realism and a larger space of possible shapes, post-processing the triangulation after collision detection is also a possibility.

4.4.2 Tongue Parameterization

The objective of parameterizing a model is to reduce the degrees of freedom to just the most significant while maintaining an adequate level of control. The 60 control points in our model represent 180 degrees of freedom, and specifying each one separately is an onerous task. Instead, we would like a parameterization with just a handful of parameters. This parameterization should also be natural to specify. For example, one such parameterization would be to specify the forces of each individual muscle, producing 10 to 17 parameters (depending on which muscles are used and which muscles are assumed symmetric) meeting the first criteria; however specifying forces or even relative muscle strengths is non-intuitive.

As noted by Stone and Lundberg [SL96], and from our own empirical research, the shape is systematically related to tongue position since the tongue volume can only be redistributed and not changed. This led us to believe just a small number of parameters would be needed to specify the tongue shape.

Using the four classes of tongue shapes identified by Stone and Lundberg [SL96] as a guide to creating the parameterization, we determined that tongue tip location (front raising), tongue dorsal height (back raising), and tongue lateral location (groove) are sufficient to deform the model into any shape necessary for the visual representation of speech. Our parameterization is not limited to our model and can

be used with other geometric models. We define the following parameters for the shape of the tongue model:

p1 - tongue tip X

p2 - tongue tip Y

p3 - tongue tip Z

p4 - tongue dorsal Y

p5 - tongue lateral X

p6 - tongue lateral Y

We assume symmetry between the two lateral sides of the tongue, which is adequate for most tongue positions, but does not allow the tongue to twist. Twisting the tongue is not a concern for our present system, but symmetric twisting can be added with another parameter for twist. Moreover, if non-symmetric twisting is desired, two more parameters are needed so that each lateral wing is defined separately.

4.4.3 Implementation

The above parameterization can be implemented in various ways. A sophisticated method would take into account volume preservation and deformation due to collision with the hard surfaces of the oral cavity. In a real-time facial animation system this sophistication is too costly to implement and is not strictly required. We use a simple and fast method that approximates volume preservation and gives good results for our real-time system.

The new control points $P'_{i,j}$ are calculated as

$$P'_{i,j} = P_{i,j} + \sum_{k=1}^{6} \omega_{k,j} p_k \alpha_{k,i}$$

Where $\omega_{k,j}$ is the weight for row j of parameter k and $\alpha_{k,i}$ is the weight for column i of parameter k. There are separate weights for x, y and z. For example, we use the following row weights in the x direction:

$$\omega_x = \begin{bmatrix} 1 & .9 & .6 & 3 & 0 & 0 \\ 1 & .5 & .1 & 0 & 0 & 0 \\ 1 & .9 & .9 & .2 & .2 & 0 \\ 0 & .1 & .3 & .9 & 1 & .2 \\ .5 & .9 & 1 & 1 & 1 & 1 \\ .1 & .7 & .9 & 1 & 1 & 1 \end{bmatrix}$$

Reducing the tongue width when the tip is extended, and expanding the tongue during retraction simulates volume preservation. This crude approximation gives reasonable results with no performance penalty.

The above implementation is highly sensitive to the shape of the tongue model in its rest position (all parameters 0). We model the Visible Human Female tongue in our images so they will not match up exactly with the tongue of the subject used in the Stone research. However, as seen in Figure 4.3, it still approximates the gross shape of the tongue and is adequate for our purposes. Figure 4.3 shows samples from the four categories of tongue shapes identified by Stone and Lundberg [SL96]. In Figure 4.3, for each phoneme, the left hand picture is the surface of the top of the tongue as measured with ultrasound [SL96], while the right hand side is our tongue model shaped for the same phoneme. The tongue tip is missing from Stone's reconstruction due to air between the ultrasound device and the tip. For /n/ the tip of our tongue model is not as curved, because it is not pressed up against the roof of the mouth as it was when captured with ultrasound on the left.

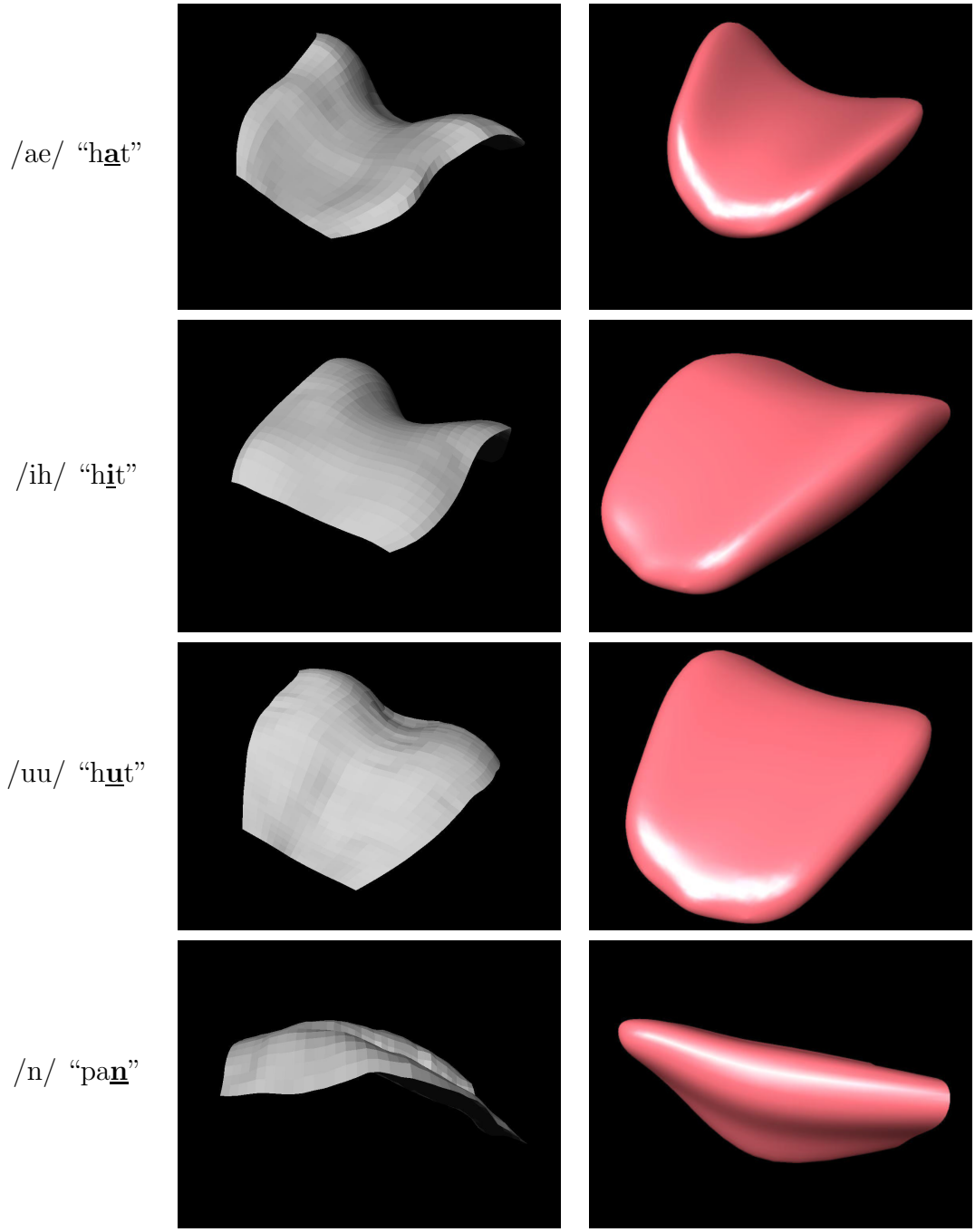


Figure 4.3: Samples from the four categories of tongue shapes from Stone and Lundberg[SL96]. The left image is a surface reconstruction from ultrasound of the tongue shape during sustained production of the phoneme, while the right hand side is our tongue model shaped to represent that phoneme. The tongue tip is missing from Stone's reconstruction because of air between the tip and the ultrasound device.

4.5 Rendering

In our real-time version, we render the tongue using Open Inventor [Wer94] and add a hand-crafted texture. This gives good results with acceptable computation. For off-line rendering, we use displacement and surface procedural shaders to increase realism. The tongue is usually under poor lighting conditions, far from the viewer and partially obscured reducing the priority of surface realism.

The displacement shader adds taste buds as bumps along the tongue's upper surface and adds veins to its underside. The displacement shader also allows for a grooved tongue. The surface shader has a user-defined color with the upper surface matte in appearance and the underside shiny. Tongue fuzz is displayed along the tops of the taste buds, as fuzz never forms along the actual surface of the tongue. The surface shader also gives the veins of the underside their blue color.

Our parameterized tongue model is capable of representing the tongue shapes produced during human speech as well as many of the non-speech tongue shapes needed during facial animation. The tongue is needed to increase intelligibility of the speech and to improve realism. Figures 4.4, 4.5

4.6 Results

and 4.6 show the mouth in phoneme positions and illustrate how the tongue aids in recognizing the sound. Figure 4.4 shows the tongue model included in our facial model in the position for the viseme /d/. Figure 4.5 shows the tongue in the shape of the phoneme /ay/ with the teeth, mandible and gums visible. Figure 4.6 shows a side-by-side comparison of the phoneme /n/ with and without the tongue model. Without a tongue, the mouth looks very unnatural, and by comparison, the image

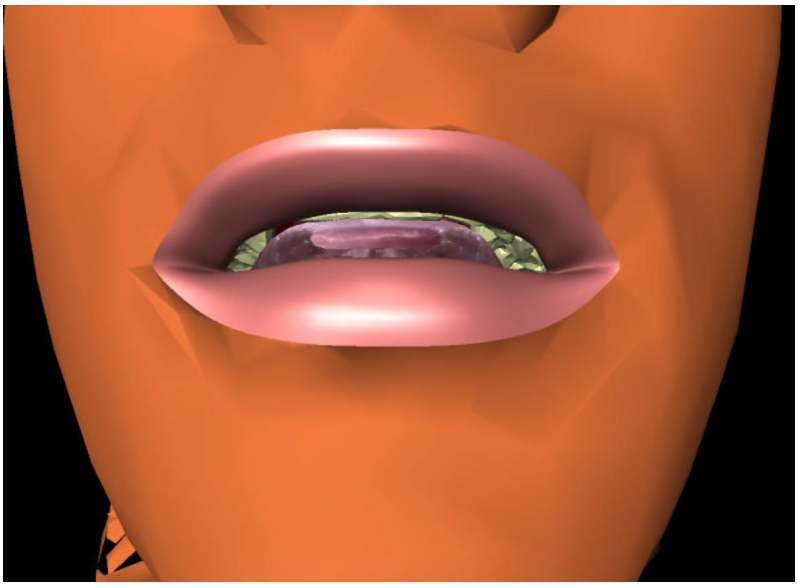


Figure 4.4: Our facial model in the viseme position for the phoneme /d/ including the tongue model we describe here.

Figure 4.5: The tongue model placed with a model of the oral cavity with the tongue in the position of the phoneme /ay/.

Figure 4.6: A side-by-side comparison of our facial model in the phoneme position /n/. On the left our tongue model has been added to the facial model, while on the right there is no tongue. The visible artifacts in the mouth on the right side are the back sides of the triangles from the rear of the head.

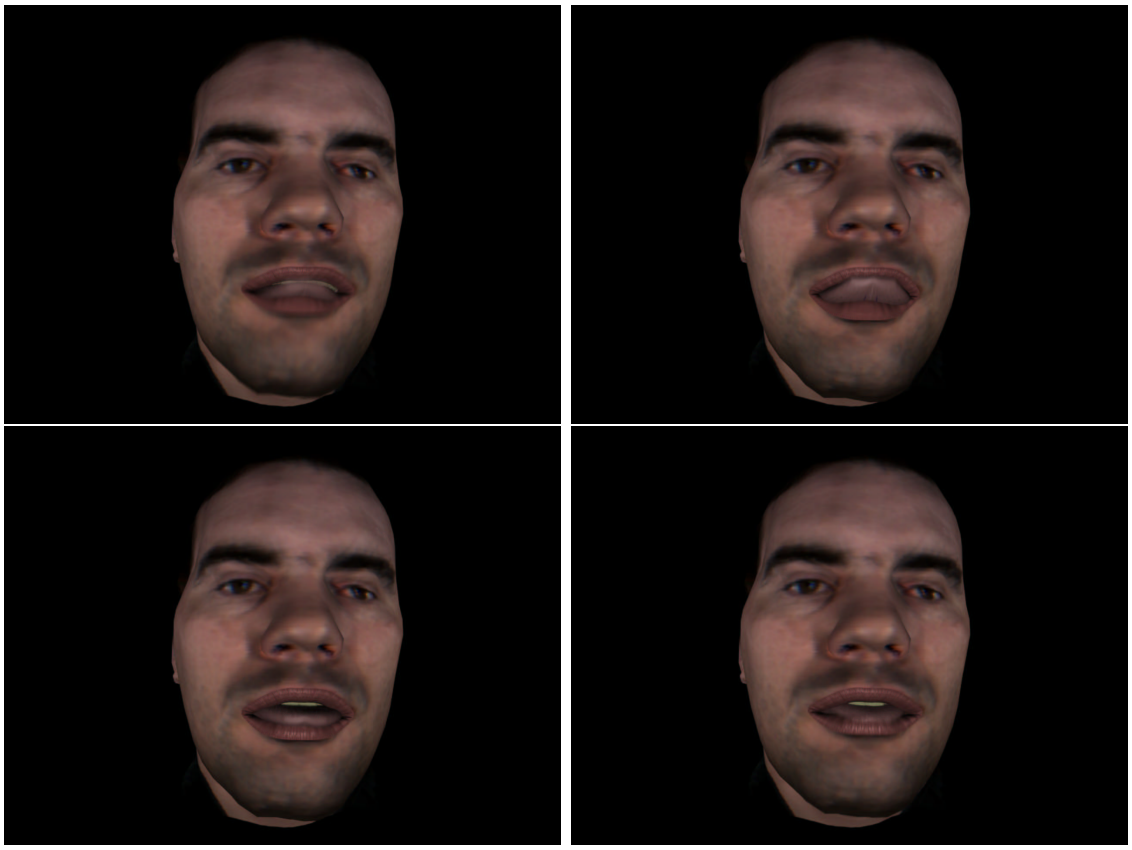


Figure 4.7: Frames from an animation of our facial model licking its lips.

with the tongue looks more realistic. This image shows how adding our tongue model will improve realism. The tongue model can also be used to increase the repertoire of facial expressions, such as licking the lips, shown in Figure 4.7, and sticking out the tongue, shown in Figure 4.8. Increasing the space of possible facial expressions helps bridge the believability gap.

4.7 Summary

Our parameterization of the tongue is able to approximate possible tongue shapes during speech for computer facial animation. Our parameterization is also sufficient

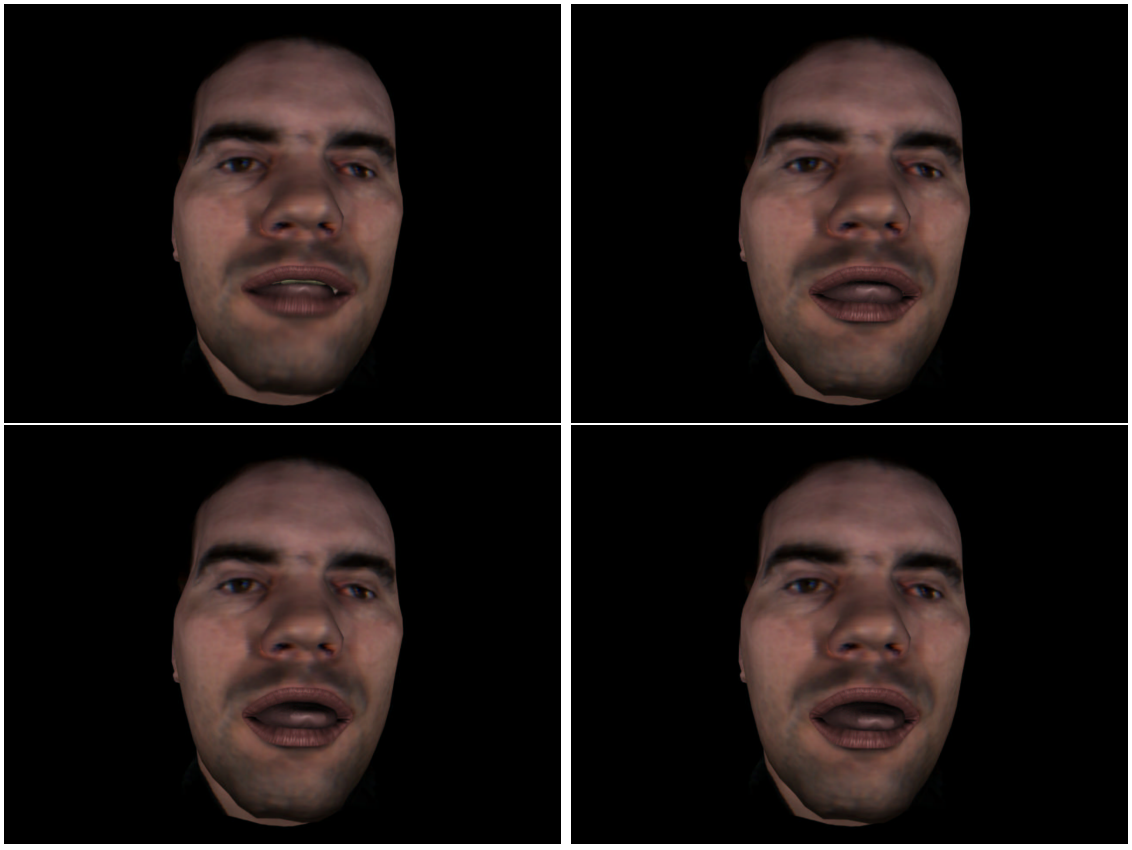


Figure 4.8: Frames from an animation of our facial model sticking out its tongue.

to approximate any tongue shape. However, our current interpretation of the parameters, which trades speed for accuracy, is not very accurate. While our implementation gives adequate shapes for speech and most facial animation requirements, it is not completely realistic. A much more sophisticated model is needed to accurately use the parameterization to full effect. Figure 4.3 shows how our shape for /n/ differs in the shape of the tip; however, the visible part of the tip (the inferior for this phoneme) is shaped correctly. Shapes that are more accurate can be achieved by adding a parameter for raising the middle section (immediately posterior to the tip) or with a more sophisticated implementation.

Realistic tongue shapes require collision detection and deformation due to the collisions. The collision detection method we use for the skin, described in Section 2.2.1, will disallow the tongue from penetrating the hard surface and give some deformation, but deformation is dependent on the number of triangles used to represent the tongue. A more sophisticated method that would adaptively change the surface mesh is needed for complete realism. Existing methods are computationally expensive and cannot be used in a real-time system, which is our goal.

A computer tongue model must preserve volume to be physically accurate. Our method of approximating volume preservation works well for shape changes due to muscle contractions, but not for deformation due to collisions. For the most accurate shapes, collisions and volume preservation must be handled together. The shape of the tongue is changed not only by the tissues that make up the tongue itself, but also by contact with other surfaces. In order to accurately model the tongue surface, the entire tongue body must be considered, which is a very complex problem.

Collisions between the tongue and oral cavity affect the shape of the soft parts of the oral cavity such as the cheeks and lips as well. For animation that is more realistic, these interactions need to be considered. The impact of the tongue on the oral cavity is external to the tongue model and must be processed independent of the model or the model must be modified to understand its environment.

CHAPTER 5

SPEECH-SYNCHRONIZED ANIMATION

To create convincing animated speech, the method of deforming the facial model over time is extremely important. A good animation method can overcome inadequacies in the facial model, but a good facial model will not look good if animated poorly. This chapter discusses how we animate our facial model and particularly how we animate the vocal tract articulators using a model of coarticulation.

5.1 Coarticulation

The speech production mechanism is a target-oriented system, with the vocal tract parts trying to reach these targets at the specified times [DJ77]. These targets are relatively close together around a rate of 10Hz. Because the parts of the vocal tract are made of tissue, they have masses, finite accelerations and decelerations, and velocity constraints. Because of this, speech is actually a continuous system, with no hard boundaries between sounds or words, contrary to that suggested by the written form. We may insert occasional pauses, but there is little to observe to mark as discrete units. Our auditory system is trained to register one of the expected phonemes when a specific target is approximated.

This blending between target values is called *coarticulation*. The movement of the vocal tract between these target locations provides important cues for identification. This motion from one target to the next must therefore be realistic to achieve intelligible animated speech.

During the act of speaking, the motions toward or away from a particular target can be as important as the actual approximation of the target. This is the case in glides and diphthongs. Glides are made by motions from a specified target to that of the following vowel. Diphthongs can be specified as two targets, with the transition between them dominating the targets themselves. Transitions also play a crucial role in the production of stop and nasal consonants. The absence of sound for these consonants carries the information, while the transition to it cues the place of the stop production [DJ77].

The listener tries to discern meaning from utterance, and she segments the utterance into meaningful word and phrase units; this is referred to as *segmentation*. Concurrently, the speaker attempts to convey the meaning, often with minimal effort. A speaker may adjust his speech to make distinguishing the targets easier. For instance, slowing the speech might be employed. Such an adjustment may be done in response to a perceived lack of understanding the meaning by the listener. The goal of speech is to convey information with minimal articulatory effort, not to produce speech that sounds good.

5.1.1 Previous Work

Coarticulation is an important issue to consider in order to produce realistic, intelligible, computer-animated speech. If the visual manifestation of coarticulation

is missing, important information for the listener will be absent making the speech harder to understand as well as visually incorrect. For these reasons, many researchers have attempted to model coarticulation effects.

Pelachaud et al. [Pel91, PBS91, PVY93] handle coarticulation using a look-ahead model [KM77] which considers articulatory adjustment on a sequence of consonants followed or preceded by a vowel. They integrate geometric and temporal constraints into the rules to make up for the incompleteness of the model. The algorithm works in 3 steps: 1) apply a set of coarticulation rules to context dependent clusters, 2) consider relaxation and contraction time of muscles in an attempt to find influences on neighboring phonemes, and 3) check the geometric relationship between successive actions.

Cohen and Massaro [CM93] extend Parke's model for better speech synchrony by adding more parameters for the lips and also a simple tongue. They include a good discussion of coarticulation research by the speech and hearing community. They use the articulatory gesture model of Löfqvist [Löf90], which uses the idea of dominance functions. Each segment of speech has a dominance function for each articulator that increases then decreases over time during articulation. Adjacent segments have overlapping dominance functions, which are blended. They use a variant of the negative exponential function

$$D_{sp} = \begin{cases} \alpha_{sp} e^{-\theta_{\leftarrow sp}|\tau|^c} & \text{if } \tau \geq 0 \\ \alpha_{sp} e^{-\theta_{\rightarrow sp}|\tau|^c} & \text{if } \tau < 0 \end{cases}$$

where D_{sp} is the dominance of parameter p for speech segment s, α gives the magnitude of the dominance function, τ is the time distance from the segment center, c controls the width of the dominance function, $\theta_{\leftarrow sp}$ is the rate on the anticipatory side, and $\theta_{\rightarrow sp}$ is the rate after the articulation.

Waters and Levergood [WL93] describe DECface, which is a system that generates lip-synchronized facial animation from text input. Using DECTalk to generate phonemes and a waveform, they associate a keyframe with each phoneme and interpolate between them. To handle coarticulation effects between the phonemes they assign a mass to the nodes and then use Hooke's law for spring calculations, with Euler integration, to approximate the elastic nature of the tissue. A method to deal with coarticulation, popular in 2D image techniques, is to use triphones instead of phonemes. A triphone is a phoneme within the context of every possible preceding and succeeding phoneme. More information must be stored but the resulting transitions are more realistic. This is popular in vision techniques where triphones can be input as training data then reused for animation.

Brooke and Scott [BS94] describe a system that uses principal component analysis (PCA) and a hidden Markov model (HMM) to create animated speech of 2D images. PCA is performed on a standard test set of 100 3-digit numbers spoken by a native British-English speaker. Fifteen principal components are identified that account for over 80% of the variance. The image data is hand-segmented with the encoded frame sequences phonetically labeled and the 15 principal components describe the image of any phoneme. The data is used to train the HMM model and for each triphone HMMs are constructed with between 3 and 9 states. Animation is achieved by fitting a quadratic through the states, which in a continuous stream of PCA coefficients. The coefficients do not exactly represent the statistics of the HMMs but they only deviate slightly and produce smoother transitions.

5.1.2 Our Coarticulation Model

Our animation system is track-based, with one parameter per track and possibly many tracks per parameter. A track consists of a curve that describes the value of the parameter over time. Our system allows for multiple types of curves including spline approximation, spline interpolation, physics-based, and our coarticulation model. We use NURBS, which include all of the popular splines as a subset, and we allow any order of curve. We also allow for defining the targets as point masses with spring constraints. Curves can also be defined with our coarticulation model. We use the coarticulation model for the parameters that control speech and the other curve types for animating facial expressions.

Our coarticulation model is a modification of the Cohen and Massaro [CM93] coarticulation model. Our model has four differences:

1. We define a phoneme by a curve instead of a single keyframe.

2. We allow the influence of a phoneme to be limited spatially.

3. We use a different value for c, giving the peak of the dominance function a convex instead of a concave shape.

4. We allow combining of the target-based approximation of the base coarticulation method with other methods of approximation, such as splines, as well as with interpolation methods.

A phoneme is a segment of speech uttered over a finite length of time. That phoneme is generated by a dynamic shaping of the vocal tract by movement of the articulators of the vocal tract. A common practice is to use a single keyframe position

for each phoneme and interpolate those keyframes to achieve animation. However, interpolating to a single keyframe position will not be able to accurately recreate the motion of the articulators. This is particularly apparent for the diphthongs, which have two distinct shapes, and the stops, which not only stop but also release air.

Each parameter of the facial model has a curve that animates that parameter over time. The parameter curve is a combination of the curves for each phoneme. A phoneme curve is defined by specifying the control points of that curve. A typical control point is $(u, \rho, \alpha, \theta, \phi, \xi, \chi, c)$ where u specifies how close to the beginning of the phoneme that control point is, ρ is the value of the parameter for that control point, and the remainder are optional and control the shape of the curve as explained below.

The control points define the shape of the phoneme curve and together, all of the phoneme curves define the parameter curve. To combine the phoneme curves, each phoneme control point gets mapped to the parameter curve as the tuple $T = (t, \rho, \alpha, \theta, \phi, \xi, \chi, c, \sigma)$. The position, t, of the control point on the curve is calculated with $t = \sigma + ud$ where σ is the starting time of the phoneme, and d is the duration of the phoneme. Both σ and d are generated by the text-to-speech process.

The control points are in effect target values for the parameter being animated. These targets are then approximated by a curve using our coarticulation model. The curve is a weighted average of dominance functions and calculated with

$$C = \frac{\sum_{i=1}^{n}(D_i T_i)}{\sum_{i=1}^{n} D_i}$$

The dominance function for target T_i is denoted as D_i and calculated as

$$D_i = \begin{cases} 0 & \text{if } \tau < \chi_i \text{ or } \tau > \xi_i \\ \alpha_i e^{-\theta_i |\tau|^{c_i}} & \text{if } \tau < 0 \text{ and} \tau \geq \xi_i \\ \alpha_i e^{-\phi_i |\tau|^{c_i}} & \text{if } \tau \geq 0 \text{ and} \tau \leq \chi_i \end{cases}$$

where ξ_i is the distance in t before the target that the target will influence, χ_i is how far past the target there is still influence, α_i is the magnitude, θ_i is the growth rate of the influence on the anticipatory side, ϕ_i is the rate of decay of the influence on following targets, $\tau = t - \sigma_i$ is the distance from the target in time, and c_i controls the shape of the peak of the dominance function. The introduction of χ and ξ to the calculations give more control over the shape of the final curve by allowing for a dominance function to decay slowly but still have limited influence.

Cohen and Massaro [CM93] suggest using a value of 1 for c, however, we believe a value of 2 gives better curves. Figure 5.1 depicts a curve using $c = 1$ along with its first and second derivatives for that curve. The dashed lines are the dominance functions, the solid line is the parameter curve, the circles depict the target values, the vertical lines depict the borders of the phonemes, and the phonemes are listed above their respective curve segments. Since the dominance function is concave on each side of the apex, there is a discontinuity in the first derivative. This causes discontinuities in the first and second derivatives for the resulting coarticulation curve. Discontinuities are unrealistic and cause visible artifacts in the animation such as jerkiness. Using $c = 2$, we get a convex curve at the apex of the dominance function which does not have a discontinuity. An example curve along with its first and second derivatives are shown in Figure 5.2. The first and second derivatives are continuous and are generally much smoother. Figure 5.3 shows the resulting curves, first, and second derivatives

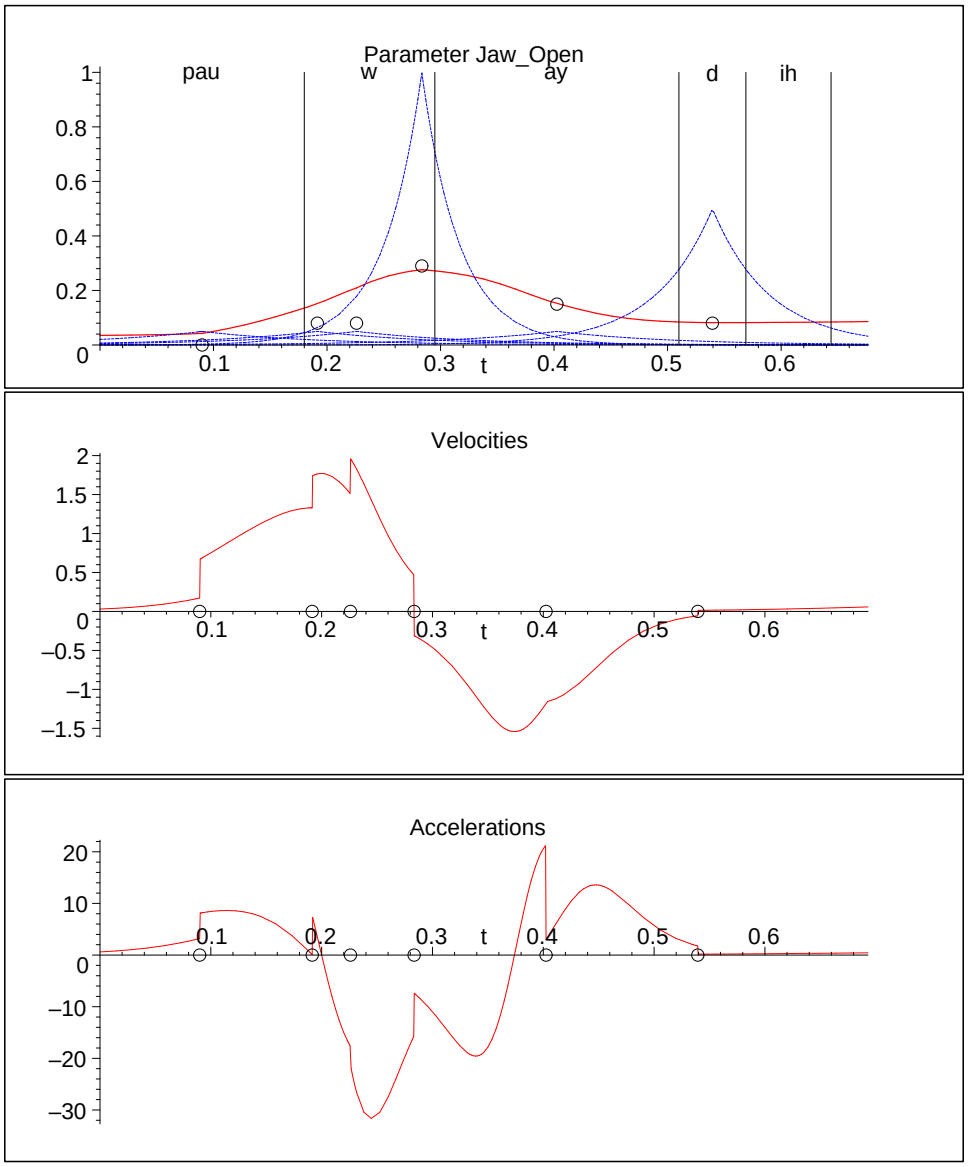


Figure 5.1: A curve using the coarticulation method of Cohen and Massaro [CM93] along with the first and second derivative curves.

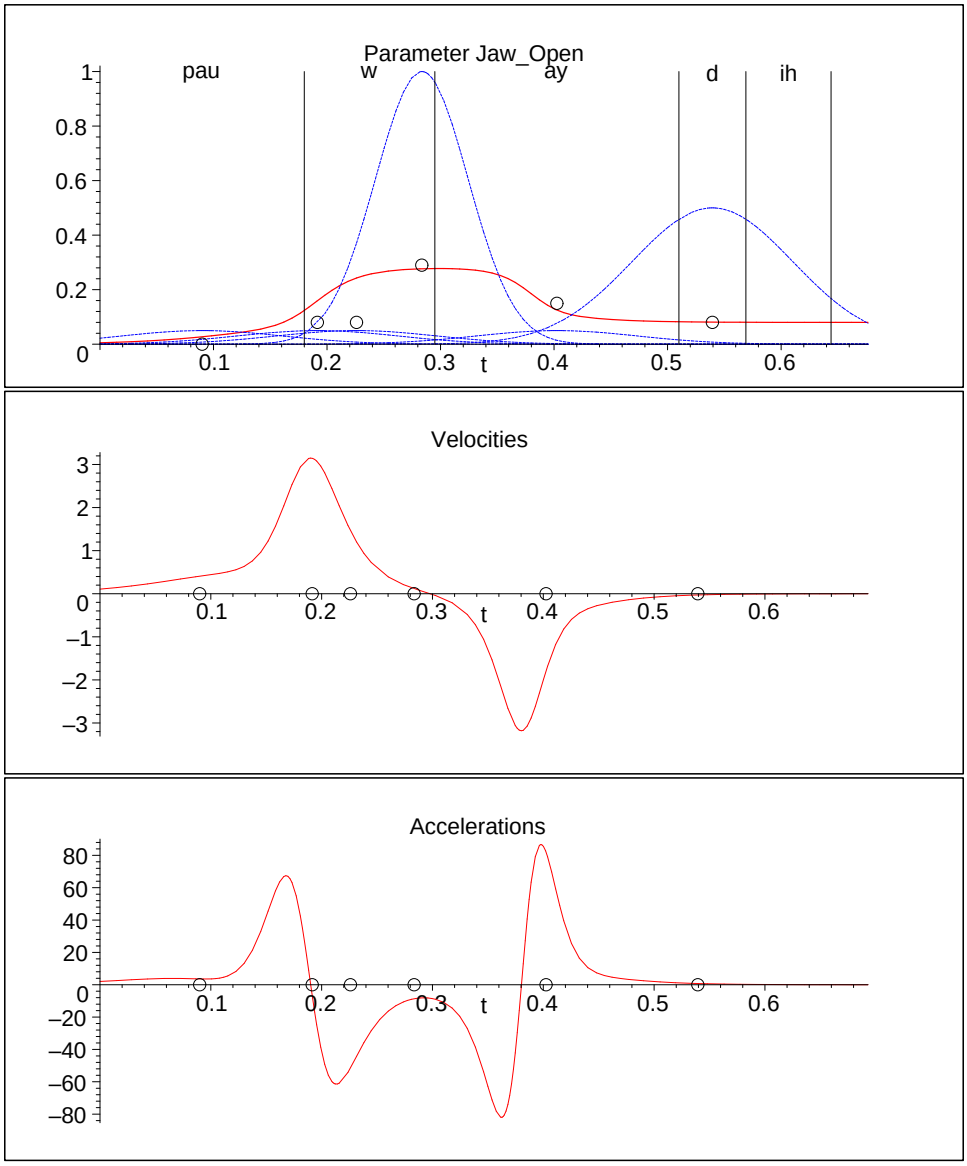


Figure 5.2: Our version of coarticulation along with the first and second derivative curves.

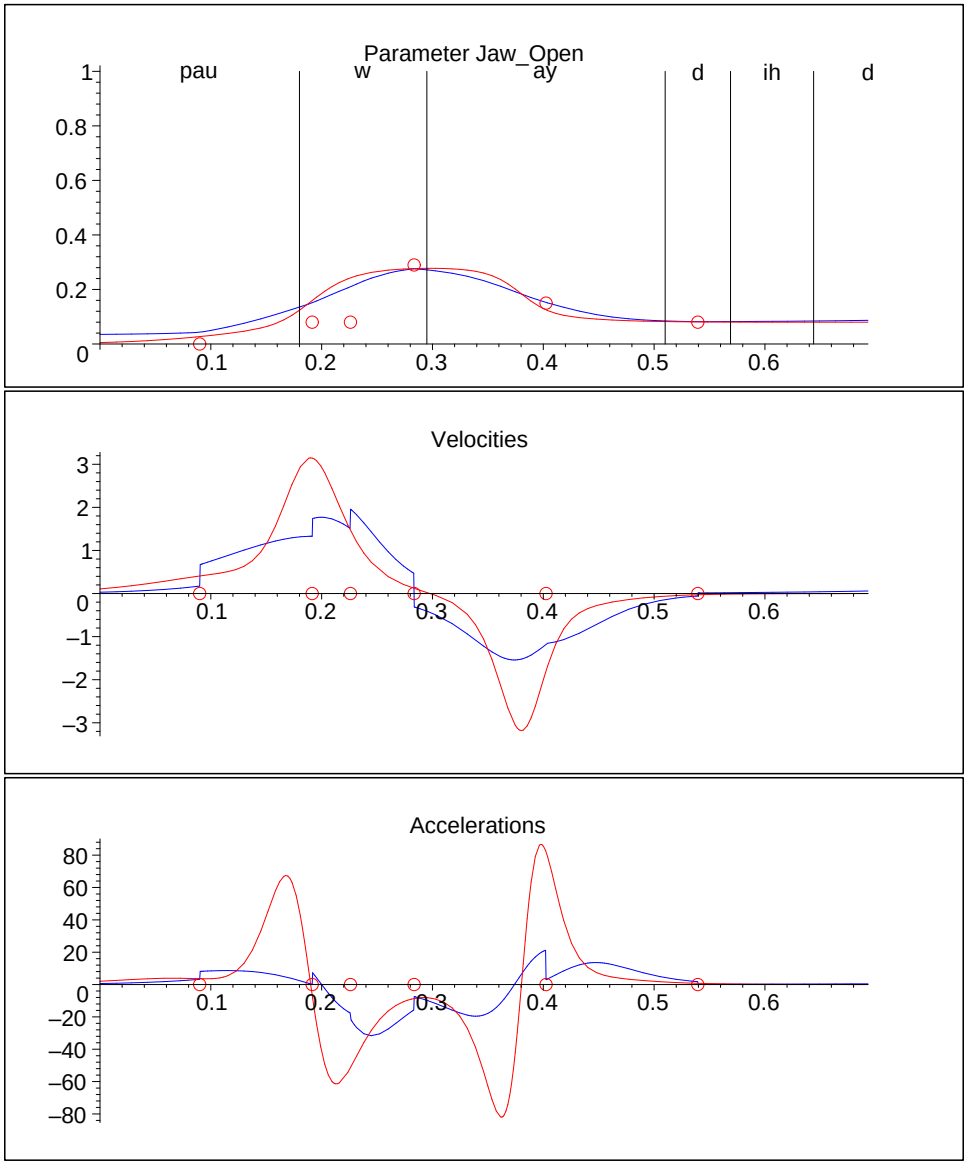


Figure 5.3: Comparison between our coarticulation and Cohen and Massaro coarticulation. Even though the curves look somewhat similar, the first and second derivatives from our method are smoother and without discontinuities.

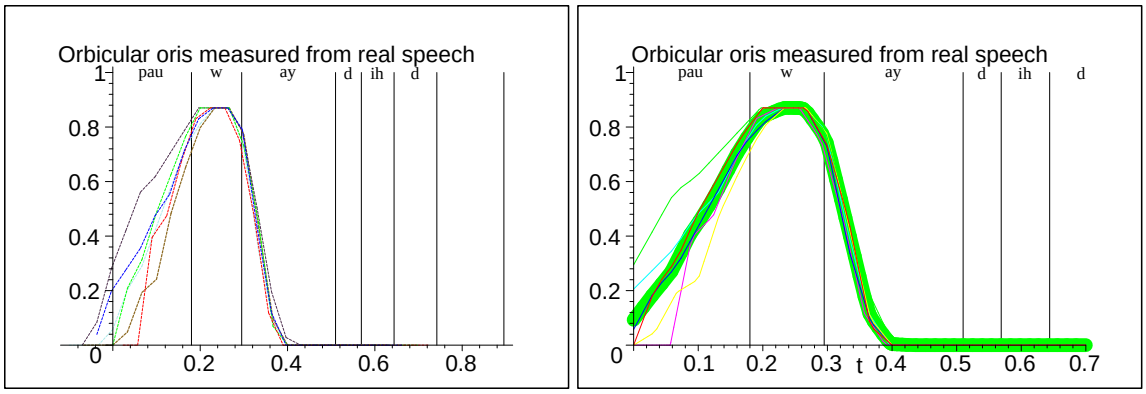


Figure 5.4: Curves for the orbicularis oris measured by hand from six different video recordings. The graph on the right shows the six curves as well as their average.

using both coarticulation models in the same plot. The parameter curves for the two coarticulation models look similar but have very different properties.

To compare the visual speech generated by our system with real speech, a subject was recorded repeating the same phrase. The phrase used was "Why did Ken set the soggy net on top of his deck?". This phrase comes from early speech research [Per69] and is included in the cineradiographic database [MVBT95]. Using this phrase allows us compare the movement of the vocal tract of our synthetic speech to the vocal tract movements of real speech that are not visible to a video recorder. The subject listened to the audio while trying to speak in-sync with that audio. These videos were then digitized and roughly synchronized with the animation produced by our system. For six of the videos, the parameter for the orbicularis oris muscle was estimated by hand for each frame. The estimate was based on the width of the lips due to the contraction of the orbicularis oris muscle. Figure 5.4 shows the piecewise linear curves that interpolate the data measured by hand, as well as the average of those six curves.

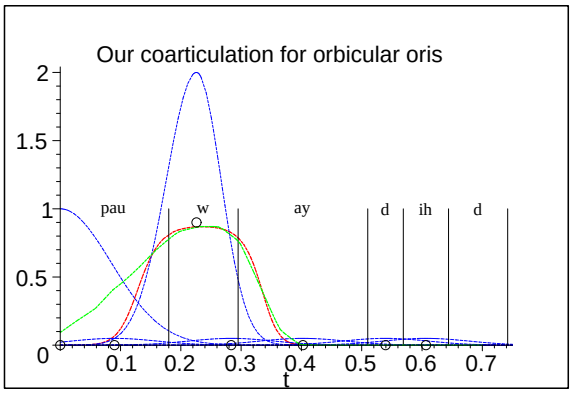


Figure 5.5: The curve generated by our coarticulation model for the orbicularis oris parameter.

Figure 5.5 shows the curve generated automatically by our system for the same phrase with the measured curve for comparison. The generated curve matches the measured curve closely, especially from just before the onset of the /w/ to the end of the segment. The beginning of the curves do not match, which could be due to many factors. We believe the most likely reason the beginning of the curve does not match is that the subject was filmed repeating the phrase several times while trying to match the audio. It is quite likely the subject was anticipating the lip rounding in order to be synchronized with the audio playing in his ears. The high variability in the measured curves seems to support this hypothesis. During more natural speech this earlier anticipatory rounding may not occur.

In previous animated speech research, the most common method is to treat each phoneme as a keyframe and use some form of interpolation. The easiest method is to use linear interpolation. Figure 5.6 shows linear interpolation of the targets versus our coarticulation model with the measured curved overlayed. The two plots show standard linear interpolation, which is the common method of one keyframe

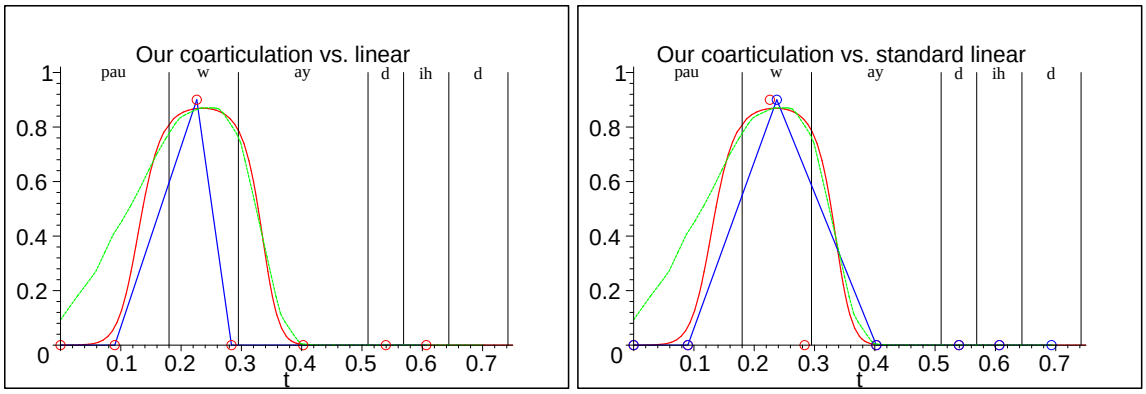


Figure 5.6: A comparison between the curves generated by our coarticulation model and linear interpolation as well as a measured curve.

per phoneme, and linear interpolation of the multiple keys per phoneme. The plots show that the linear interpolation will not hold the lips in the rounded position for very long. In fact, when animating at 24 or 30Hz, the frames could fall on either side of the peak resulting in a much lower peak value. With linear interpolation the anticipatory rounding plus the lag after the rounding does not occur. This is a byproduct of interpolation.

To solve some of the problems of linear interpolation, an acceleration and deceleration near the keys can be given and a cubic interpolation used. Figure 5.7 shows the natural spline, or a blending of the interpolating cubic polynomials, through the targets. Again, plots using the standard technique of one key per phoneme and using multiple keys are shown. Here we do get the acceleration and deceleration effect so the lips will stay rounded longer. However, the resulting curve will overshoot the targets resulting in higher and lower values of the parameter, which could cause disastrous results.

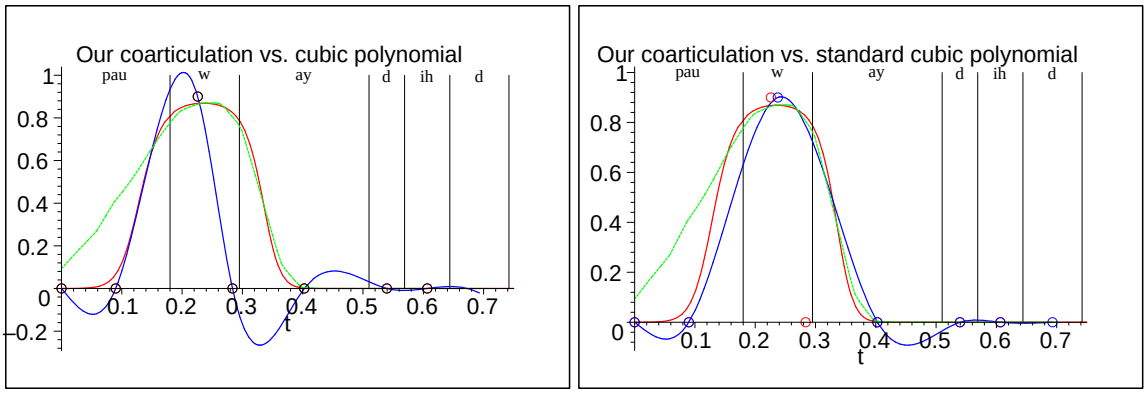


Figure 5.7: A comparison between the curves generated by our coarticulation model and cubic polynomial interpolation as well as a measured curve.

Instead of interpolation we could use approximation of the keys. Bezier approximations are a popular choice and are depicted in Figure 5.8. Here we see that the curve will not get close enough to the peak values as we would like. Adding a weight to a key will allow the curve to get closer resulting in a rational Bezier curve, which is depicted in Figure 5.9. Rational Bezier curvess allow anticipating and lagging the lip rounding, but they lose the acceleration and deceleration. We could use an interpolating Bezier curve and just specify the tangents at each key location. This would allow us to reach our target and have an acceleration and deceleration. However, if we interpolate the keys we still have the problem of not allowing enough anticipatory rounding.

B-splines are another popular curve choice for keyframe animation. As with Bezier curves, the B-spline curve may not approach important extreme keys, as seen in Figure 5.10. Rational B-splines, shown in Figure 5.11 can help, but again with the loss of the acceleration and deceleration properties. The order of the curve can also become a problem as higher-order curves can be very wild. Another method to get

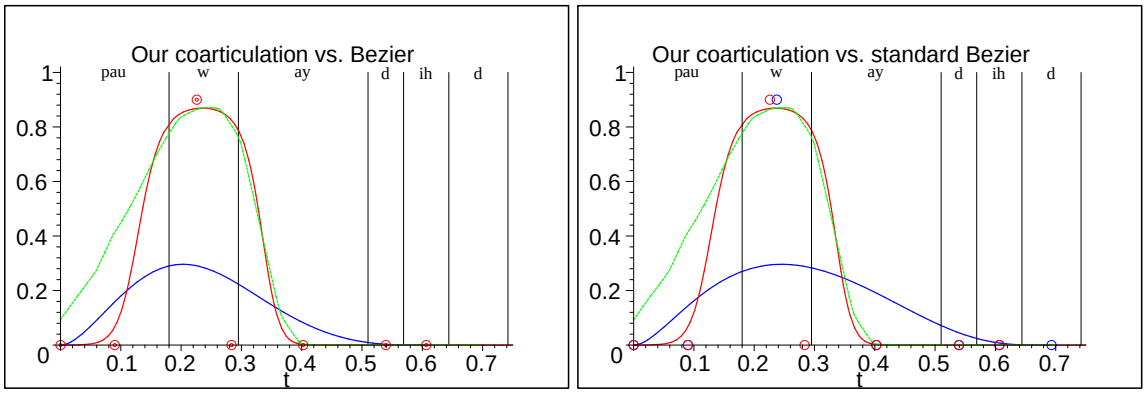


Figure 5.8: A comparison between the curves generated by our coarticulation model and Bezier approximation as well as a measured curve.

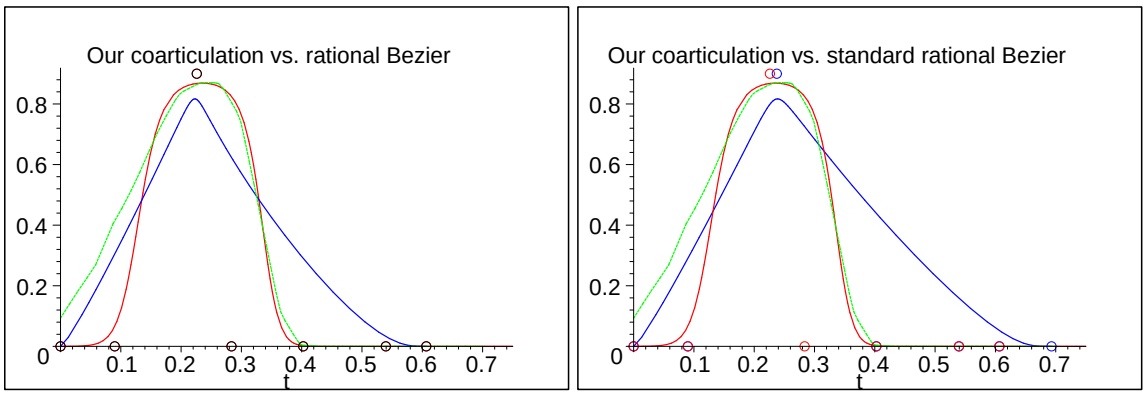


Figure 5.9: A comparison between the curves generated by our coarticulation model and rational Bezier approximation as well as a measured curve.

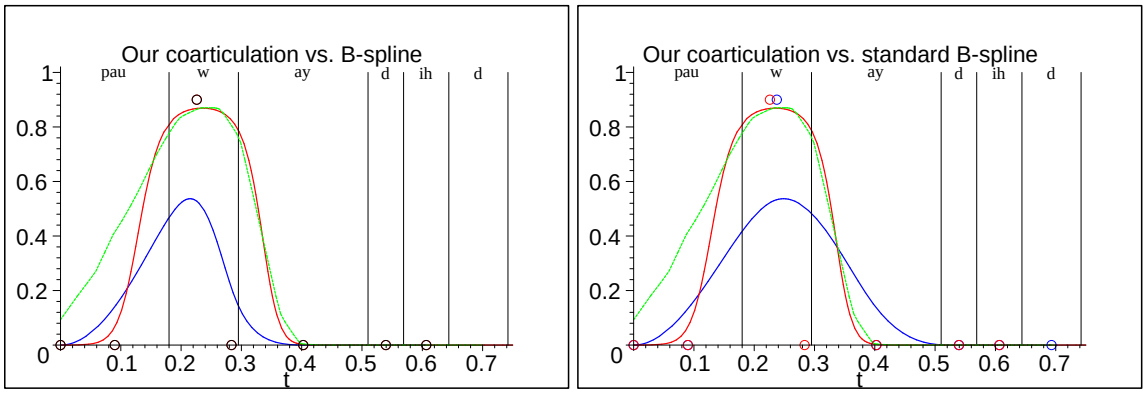


Figure 5.10: A comparison between the curves generated by our coarticulation model and B-spline approximation as well as a measured curve.

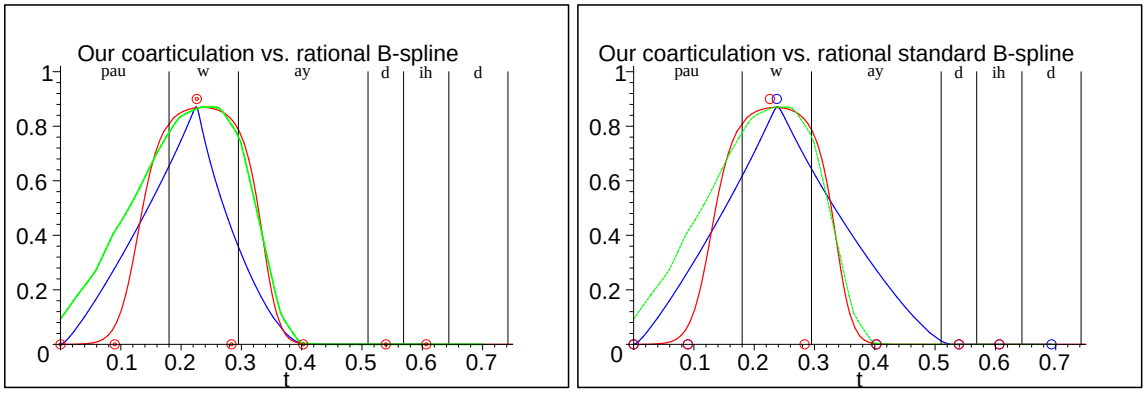


Figure 5.11: A comparison between the curves generated by our coarticulation model and rational B-spline approximation as well as a measured curve.

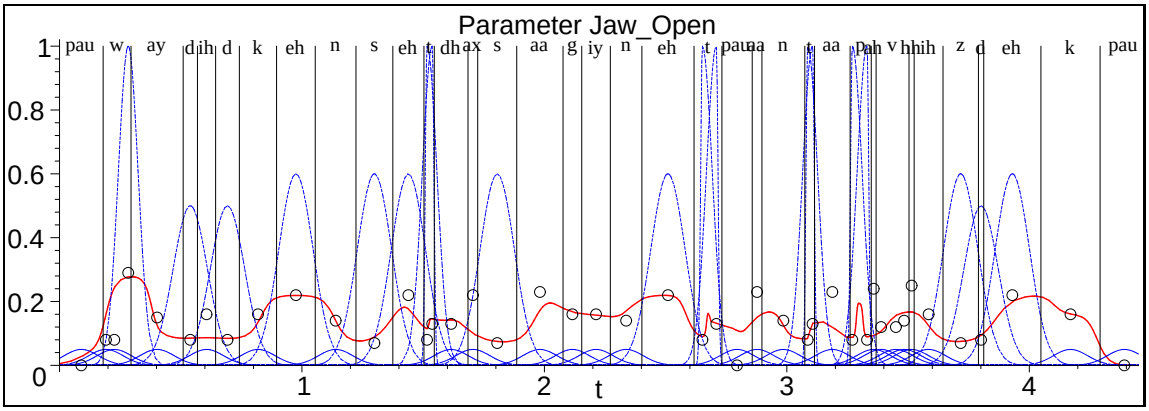


Figure 5.12: A curve for the Jaw_Open parameter generated by our coarticulation model for the phrase "Why did Ken set the soggy net on top of his deck?".

the curve to approach a key is to have multiple knots for the key. The curve will indeed approach that location, but a discontinuity in the derivatives will result.

5.2 Results

An example curve generated by our system for the phrase: "Why did Ken set the soggy net on top of his deck?" for the parameter Jaw_Open is show in Figure 5.12. The Jaw_Open parameter controls rotation of the jaw downward, causing the mouth to open or close. The dashed lines are the dominance functions for each target, which controls how close the parameter curve gets to the target value. The targets are displayed as circles. The boundaries of the phonemes are denoted by vertical lines with the phonemes displayed above the portion of the curve for that phoneme.

5.3 Summary

Our coarticulation model is in the early stages and more development is required. Particularly we need more analysis of actual speech to verify that our model gives

good prediction of articulatory movement, and if not, to adjust the parameters to get better prediction. It may also be necessary to modify our coarticulation model to give curve shapes we have yet to encounter.

One modification that is probably necessary is to have velocity and acceleration constraints on the curves. The tissues involved have such constraints so the model should not produce animation that violates these constraints. Unfortunately it is difficult to measure velocities and accelerations of the articulators themselves, and even more difficult to measure them for the parameters of our model. Video is an available source of information, but sampling the curve at 30Hz is inadequate to get valid velocity and acceleration information. Access to high-speed motion capture systems will help, but the measurements are mostly indirect and therefore not as accurate. Direct measurements are difficult because attaching a sensor to the articulator will cause the speaker to change her speech patterns.

CHAPTER 6

TALKINGHEAD: A TEXT-TO-AUDIOVISUAL-SPEECH SYSTEM

6.1 Introduction

We combine our facial model and animation techniques into a text-to-audiovisual-speech (TTAVS) system. Plain text is fed to the system and an animation of a talking head is produced. Both the audio and video are automatically produced by the system from the input text alone. Such a system is capable of rapidly generating animations of speech. Example uses are a user interface to a catalog, a talking kiosk for visitors, a web site greeter, a book reader, rapid creation of plays or films from scripts, etc.

The advantages of a text-to-audiovisual-speech system over using prerecorded video is storage and cost. Storing the hundreds or thousands of possible hours of video is tremendous, and some may possibly never be viewed. As well, creating the video is an expensive endeavor in time and resources. An actor must be paid along with camera operators to shoot the raw footage. The equipment and location costs must also be considered. The raw footage must then be edited for use. Simple changes in the subject matter could cause a significant fraction of the existing footage to be rendered useless. Instead of recording an entire soundtrack, recording very short

clips, say of words or phrases, and editing them together is a much cheaper solution. However, changing the character, lighting conditions, or even the language requires a whole new set of recorded clips. Plus the clips will not merge together seamlessly. These costs make applications, such as a talking kiosk, prohibitively expensive using video footage of real humans.

On the other hand, a system such as ours has many advantages including:

- Cheap production and storage costs.
- Only video that is viewed need be created.
- Updates are immediately effective.
- Multiple characters can be used.
- Multiple languages can be used.
- Delivery of the information can be easily modified, such as using faster or slower speech.

6.2 System Overview

Figure 6.1 shows a schematic overview of our text-to-audiovisual-speech system. The first stage of the process, detailed in Section 6.3, converts the input text into phonemes. The phonemes are then used to create a speech waveform. The first stage is done with existing software. The next stage of the process, discussed in Section 6.4, is the viseme and expression generator, which converts the phonemes into visemes and adds facial expressions to the animation. The visemes are a set of target values used by the coarticulation model. Facial expressions are generated

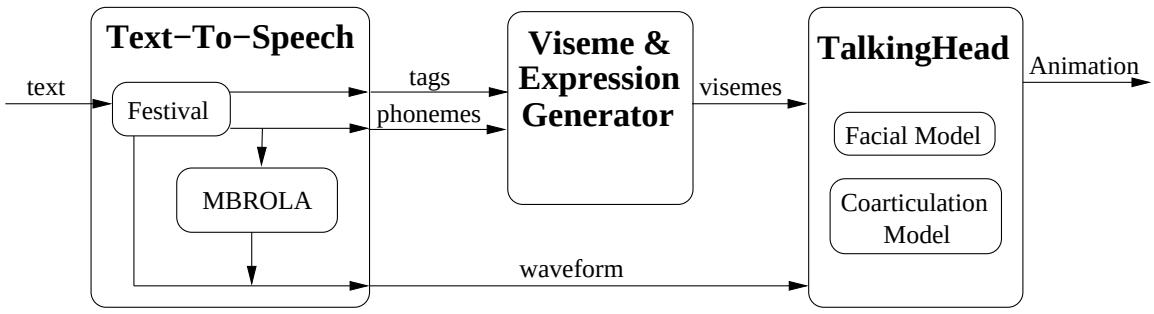


Figure 6.1: TalkingHead System Overview

from tags in the input text. Finally, the TalkingHead software uses the visemes to animate the facial model, which is combined with the speech waveform to produce speech-synchronized animation. The TalkingHead animation system is presented in Section 6.5. The last two stages use the research described in this dissertation.

6.3 Text-To-Speech Synthesis

Text-to-speech processing is achieved using Festival [BTCC00]. Festival is a freely available, general, multi-lingual speech synthesis system offering a full text-to-speech system. Festival parses the text into a stream of phonemes, which include the necessary information to produce the speech waveform such as frequency and timing.

The phonemes are then used to create a speech waveform by a speech synthesizer such as a Festival voice or MBROLA [MBR99]. Festival voices consist of a set of diphones and some scheme code to provide a front end, including text analysis, pronunciation, and prosody prediction. The voices are separate from Festival itself, and are distributed as packages. They come as a language-sex pair, such as a British English female or a German male. The Festival distribution comes with many

voices, and the creation of new voices has been made rather straightforward leading to numerous voices created and distributed by other researchers.

The system is designed to create a new speech waveform for each animation produced. Although great strides have been made in the quality of computer synthesized speech, it still does not quite sound natural. It may be desired to use a recorded waveform along with the animated visual speech. This can be done currently in our system, but the timing of the phonemes must be calculated by hand to match the recorded dialog. While there are some promising techniques that could be used to automate this process, it is not the focus of our research.

6.4 The Viseme And Expression Generator

The job of the viseme and expression generator is to convert the phonemes into a set of visemes and generate facial expressions. Most researchers [BL85, Ber87, NHS88, CM93, NHS88, Par72, Par75, Par74, Wat87] map several phonemes into a single viseme based on speech-reading [Eri89, JB71, KBG85, NNT30, Wal82] techniques. This is also the method used in traditional, hand-drawn animations [Mad69]. In these methods, a single keyframe is used to represent a phoneme. However, a phoneme is not a static configuration, but rather a dynamic shaping of the vocal tract.

In our method, a viseme is actually a curve that can be of any order. We do this by creating targets, which are control points for a curve. A viseme is defined by as many target positions as are necessary to create the desired shape. The targets are generally control points for our coarticulation model. However, the targets may also be approximated or interpolated with other curves, or as combinations of different curves. Each phoneme has a separate definition. However, it may be very similar to

```
(p ((.1 1 (jaw_open .08 1 300 10000)(jaw_in 0)(jaw_side 0)(orb_oris 0)
    (l_ris 0)(r_ris 0)(l_plat .24)(r_plat .24)(l_zyg 0.21)(r_zyg 0.21)
    (l_lev 0)(r_lev 0)(dep_inf 0)(dep_oris 0)(mentalis .34 2 300 9000)
    (l_buc 0)(r_buc 0)(r_wing_up 0)(inc_sup .17 1 300 9000)(inc_inf 0)
    (tip_side 0)(tip_up 0)(out 0)(mid_up 0)(l_wing_out 0)(l_wing_up 0)
    (r_wing_out 0))
   (.8 1 (jaw_open 0.08 1 10000 300)(jaw_in 0)(jaw_side 0)(orb_oris 0)
    (l_ris 0)(r_ris 0)(l_plat .24)(r_plat .24)(l_zyg 0.14)(r_zyg 0.14)
    (l_lev 0)(r_lev 0)(dep_inf 0)(dep_oris 0)(mentalis .34 1 9000 300)
    (l_buc 0)(r_buc 0)(r_wing_up 0)(inc_sup .17 1 9000 300)(inc_inf 0)
    (tip_side 0)(tip_up 0)(out 0)(mid_up 0)(l_wing_out 0)(l_wing_up 0)
    (r_wing_out 0))))
```

Figure 6.2: An example viseme definition for the phoneme /p/. This definition is used by the viseme generator to create list of target values that the animation module uses to create the animation.

or exactly the same as the definition of other phonemes. The set of viseme definitions are associated with particular characters and voices.

Each phoneme has an associated viseme, which is a list of control points that define the shape of the viseme curve. Each phoneme has a length and a control point has a value that places that control point at the appropriate time during the time span of the phoneme. The control point has target values for all necessary facial model parameters along with appropriate parameters for the coarticulation model. The parameter values for the coarticulation model allow the facial model parameter curve to be shaped as needed. The coarticulation model is discussed in detail in Section 5.1.

Figure 6.2 shows an example definition of a viseme for /p/. In this case, there are two control points to shape the viseme curves. One of the control points is located 10% of the way into the phoneme and the other at 80% along the way. The first control

point is at $u = .1$. For the jaw_open parameter the value is .08 with a dominance of 1 and the function decays on the leading edge at a rate of 300 and on the trailing edge at a rate of 10000. For the jaw_in parameter the values is 0 and default values are used for the coarticulation model. The default values are typically .05 for dominance and a decay rate of 85. The default values are defined separately for each facial model parameter as part of the viseme set.

Currently, visemes are created by ad hoc methods, using a mirror or film footage as a reference. A target viseme position is created using a user interface. Appropriate coarticulation model parameters are then determined to get the parameter curves to match the observed motion. We are researching methods to automatically obtain motion curves for the facial model parameters. Unfortunately, measurements can only be obtained indirectly, as direct measurements often change the method of speaking. In addition, only the resulting motion can be determined, which must then be mapped back to parameter values. It is possible for two different sets of parameter values to describe the same facial shape.

At this stage, keyframes for expressions and head movement are also added to the animation. Our system is capable of reading tags in the speech for expressions such as smile, blink, frown, wink, etc. Conversational cues could also be inserted automatically as is done in systems such as that of Cassell et al. [CPB+94]. We are at an early stage of development of such improvements. These extra facial expressions are necessary to produce life-like realistic speech. However, our focus is currently on synchronizing movement of the visible parts of the vocal tract with the speech.

6.5 The TalkingHead Animation Subsystem

The final stage is TalkingHead, which combines a visual signal with the speech waveform to create animated speech. TalkingHead takes an animation definition, a facial model, and the speech waveform and produces animated speech. An animation definition is a collection of animation tracks with each track controlling a facial model parameter. There may possibly be more than one track controlling the same parameter and they are combined in one of many possibly ways based on properties of the animation tracks.

Figure 6.3 shows a screen shot of TalkingHead, which is the part of our TTAVS system that animates our facial model. TalkingHead is also used to create and test the parameter values for the visemes. There is a slider for all of the facial model parameters as well as sliders to control visibility and the animation.

For near real-time display, we use Open Inventor^{TM} [Wer94] which, unoptimized, results in a 10-15Hz display rate on a mid-line SGI graphics workstation. For a more photo-realistic rendering, we use our custom shaders for the lips and tongue producing animations at less than ten seconds per frame.

6.5.1 Different Characters

The system is capable of animating different human or human-like characters. To be human-like, the lips must look and behave similar to that of a human's and the jaw motion must be analogous to human jaw motion. A character can have different geometry, texture maps, voice, language, sex, or way of speaking. To change the geometry requires fitting the lip model and skull to the new geometry and picking the characteristic points. Along with fitting the lip model, the displacements of

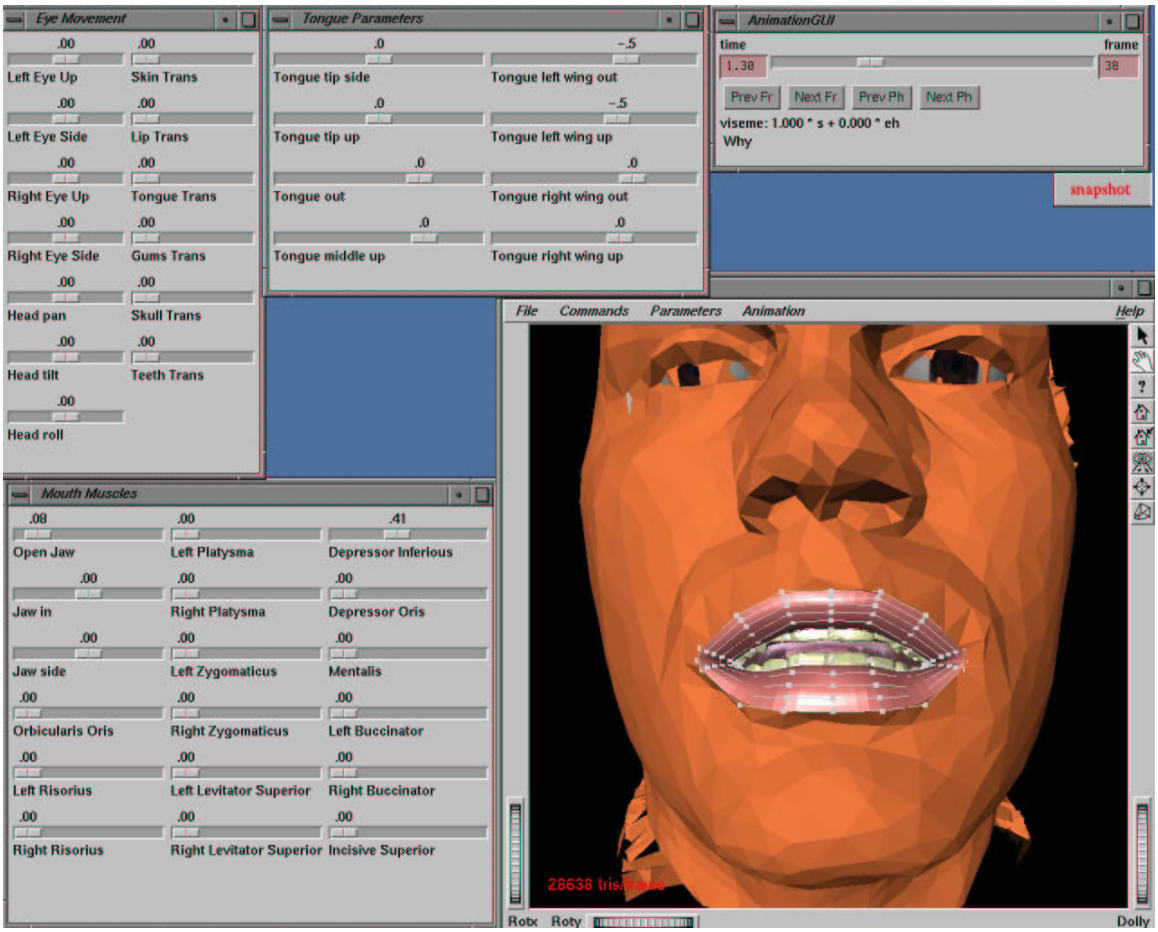


Figure 6.3: Screen shot of TalkingHead, which is the part of our system that animates our facial model to create animated speech.

the lip control points for each of the parameters need to be defined. Texture maps are changed by using new texture images and possibly changing the texture coordinates. The voice, language and sex are features of the text-to-speech software and are changed by selecting a new voice that has the characteristics desired. Changing sex may also involve changing the motion definition files of the lip model. Speaking patterns can be changed by using a different set of text-to-phoneme rules or by changing the shape of the viseme curves.

6.6 Summary

TalkingHead, our text-to-audiovisual-speech system, is capable of quickly producing animated speech. Multiple languages and multiple characters are easily used. Because we use a 3D method, lighting conditions and camera positions are easily changed. Video does not need to be prerecorded reducing the storage requirements and only those animations that are viewed need to be generated. It also is a simple process to change speech properties, such as making the rate of speech either faster or slower, depending on the listener's preferences. Our system design is perfect for applications such as a talking kiosk or web greeter and the individual parts can be used in many other applications.

CHAPTER 7

RESULTS

We have thus far described our facial model designed for animated speech, as well as our techniques for animating our facial model. We also described our text-to-audiovisual-speech system which combines our facial model with our animation methods. Our facial model, or parts of our facial model, could certainly be incorporated into other facial models. It could also be animated with other techniques. In addition, our animation techniques are not restricted to our facial model.

The goal of our research is to move towards individual realism. Specifically, we are interested in advancing the art of animated speech. We have designed our facial model with static individual realism in mind, but the model is still incomplete. It only deals with shapes needed for speech. Our animation techniques are designed for dynamic individual realism but we are currently only interested in dynamic realism of speech.

In this chapter, we present results of our facial model and our animation techniques. Some of these results have been previously presented in early chapters, and are included to make this chapter complete. We first give modeling results describing the capabilities of our tongue and lip models. Finally, we present results of our animation techniques.

7.1 Results From Our Facial Model

We have created a facial model specifically designed for animating speech. The animation of speech requires highly deformable lips and a tongue because the mouth conveys important information about the speech signal. We desired a facial model that could represent the geometry of any human. To achieve this we accept facial geometry as input to the facial model. The anticipated source of the geometry is laser scanners or human modelers. To guarantee enough geometric complexity in the lips and tongue we create highly deformable models of both the tongue and lips. This section presents results of our facial model emphasizing the tongue and lip models.

7.1.1 The Tongue Model

Figure 7.1 shows samples from the four categories of tongue shapes identified by Stone and Lundberg [SL96]. Stone and Lundberg used an ultrasound device to reconstruct the surface of the tongue during sustained production of phonemes. The tongue tip is missing from Stone's reconstruction because of air between the tip and the ultrasound device. On the left are the surface reconstructions from their research while the right hand side is our tongue model shaped to represent the same phoneme. Our tongue model is capable of representing the general shape of the surface of the tongue of the four categories identified by Stone and Lundberg as needed for American English. For /n/ the tip of our tongue model is not as curved, because it is not pressed up against the roof of the mouth as it was when captured with ultrasound on the left. Our tongue model is shaped to represent the Visible Human Female, while the tongue reconstructions by Stone and Lundberg are of their subject.

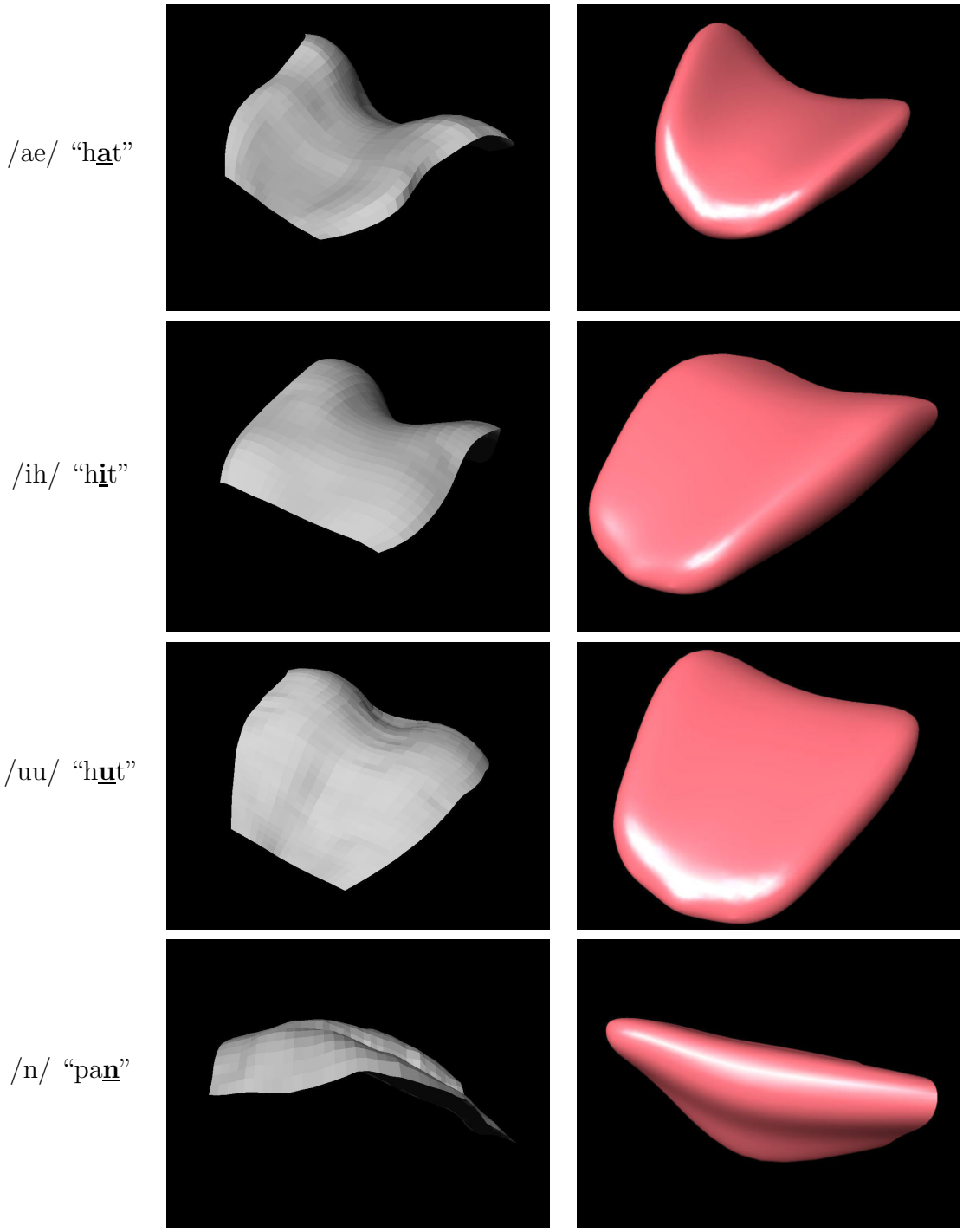


Figure 7.1: Samples from the four categories of tongue shapes from Stone and Lundberg [SL96]. The left image is a surface reconstruction from ultrasound of the tongue shape during sustained production of the phoneme, while the right hand side is our tongue model shaped to represent that phoneme.

Figure 7.2: A side-by-side comparison of our facial model in the phoneme position /n/. On the left our tongue model has been added to the facial model, while on the right there is no tongue. The visible artifacts in the mouth on the right side are the back sides of the triangles from the rear of the head.

The tongue has not been given much consideration in facial animation and modeling in the past. The most common reasons given are the tongue does not play an important role in discerning between sounds and it is mostly obscured. In our opinion, the tongue gives important information and is critical to distinguish certain sounds such as /th/ and /l/. Furthermore a complete lack of a tongue, or a tongue that does not behave correctly drastically reduces the believability of a facial model. Figure 7.2 shows our facial model in the position of the phoneme /l/, both with and without the

Figure 7.3: Frames from an animation of our facial model licking its lips.

tongue. The image on the left has the tongue while the image on the right is missing the tongue. In the image without the tongue, it is impossible to discern what sound is being made and the lack of tongue is extremely noticeable.

In addition to representing the tongue shapes needed during speech, our model is capable of other tongue shapes needed for general facial expressions. Figure 7.3 is made up of a selection of frames from an animation of our facial model licking its lips. An action such as licking the lips is a natural action that can occur during a conversation and adds to the believability of the facial model.

A character could also be rude, taunting, or teasing and stick their tongue out at the listener. Figure 7.4 is composed of frames from an animation of our facial model showing some ill-mannered behavior. Figure 7.5 shows both a side and front view of our facial model with its tongue hanging out. These images show that a highly

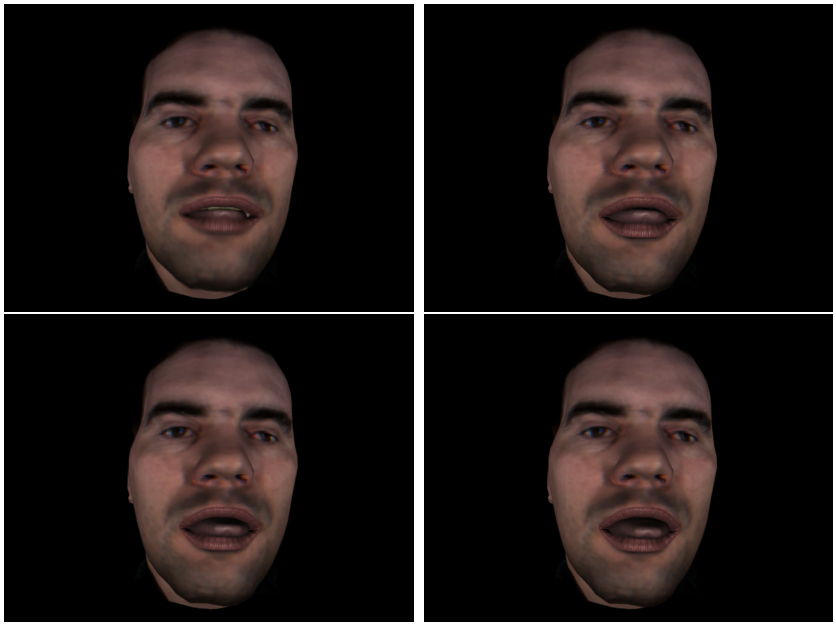


Figure 7.4: Frames from an animation of our facial model sticking out its tongue.

Figure 7.5: Our facial model sticking out its tongue.

Figure 7.6: Our facial model with its lips puckered as if to kiss.

deformable tongue is not only beneficial for producing animated speech, but also for showing emotion and life functions that make an animation more realistic.

7.1.2 The Lip Model

The lip model is designed to represent both speech and facial expressions. In Section 7.2 the results from our animation methods are presented and show the ability of the lip model to represent the shapes needed for speech. In this section, we exhibit the lip model in various facial expressions.

Figure 7.6 depicts a front and side view of our facial model as it puckers its lips in anticipation of a kiss. Notice in the side view how the lips actually protrude outward.

Figure 7.7: Our facial model is happy.

The protrusion of the lips during the contraction of the orbicularis oris muscle is important for realism. Also, notice that the lips are rotated outward.

Showing emotion is very important for a facial model. Facial expressions can convey the emotional state of the speaker and is important in determining the meaning of speech. It is also important for creating animations of characters that need to show life. A talking head that only speaks without emotion seems robotic and boring. In addition, emotional emblems can be used to express an emotion to a listener that the speaker is not currently experiencing. Figure 7.7 shows our lip model in a good mood. While in Figure 7.8 our facial model appears unhappy. In Figure 7.9 the facial model is depicted with a very big, open-mouthed smile. Finally we show the facial model in a half smile, possibly a smirk, in Figure 7.10. Note that the images do not quite

Figure 7.8: Our facial model expressing sadness.

Figure 7.9: Our facial model in a very happy mood, with a very big grin.

Figure 7.10: Our facial model in the position of a wry smile.

display the intended emotion. The emotions may even seem fake. This is because the emotion is only being displayed by the mouth and not by the eyebrows and eyes. Competing emotions in the different regions of the face (eyes, eyebrows, mouth) is a sign of deceit [EF84, Ekm85].

7.2 Animation Results

One method of evaluating our resulting speech is to compare it against real speech. To accomplish this, we recorded our subject speaking the phrase: "Why did Ken set the soggy net on top of his deck?". We created synthetic speech of the same phrase using the techniques described in this dissertation. In order to compare the two utterances, we specified the timing of the phonemes for the synthetic speech by hand so that it would closely match the timing of the real speech. Figure 7.11. shows the

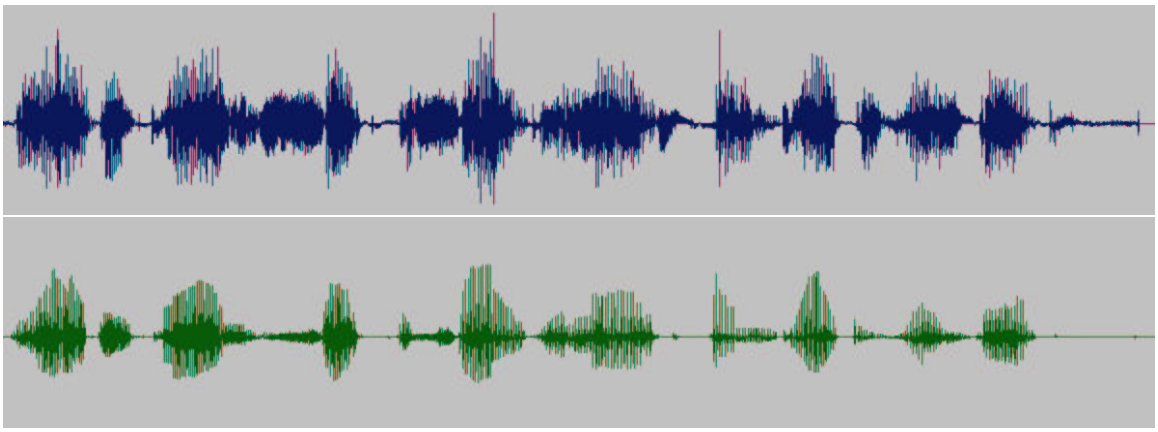


Figure 7.11: A comparison of the waveforms for the synthetic and natural speech. The top waveform is from the real speech and the bottom waveform is synthetic.

two speech waveforms. Note that the timing is indeed quite close, but not exact. Boundaries of words can often be matched very closely, however, there is not always a clearly distinguishable separation between phonemes. Without a clear start and stop of each phoneme, it is impossible to match the timing exactly.

To make a better comparison, we recorded the real speech in both a head-on and profile view by using a mirror at roughly a 45 degree angle. A mirror allows the acquisition of synchronized, orthogonal (or near orthogonal) views without expensive hardware and multiple cameras. The real speech was digitized and the mouth area was cropped from the frames. The synthetic speech was also rendered using the same two views and also cropped. The two cropped images were then combined into a single frame resulting in an animation that showed both synthetic and real speech in rough synchronization from two separate views.

There are a few things to consider before looking at the frames. First, the synthetic speech does not use prosodic information so the difference in emphasis and pitch will

cause different target positions. Also, consider that the real speech was filmed with the subject repeating the phrase while listening to the synthetic speech. This was done to get similar timing and similar prosodic features to the speech. As a result, the beginning mouth position of the real speech is not in a neutral position (closed mouth). Remember that the timing of the phrases are very similar but they are not exactly the same, a difference of only .03 seconds will cause movement an entire frame later. Note that the model may indeed look similar, but it is not meant to be an exact match. For example, the lips are not exactly shaped the same way and the tongue is shaped like that of the Visible Human Female and not like that of the subject. Thus, the tongue is much more flat in the mouth for the synthetic speech than it is for the real speech. Finally, the synthetic speech is rendered with different lighting conditions. Specifically, there is a light shining more directly into the synthetic mouth and there are no shadows cast in the synthetic mouth.

Figures 7.12-7.20 show the frames of the comparison. You should notice that the real speech starts with the mouth in a more open position and starts rounding for the /w/ a few frames earlier. The open mouth is just a different starting position than our neutral position for the synthetic speech. The earlier rounding could either be due to more emphasis on "why" for the real speech, or because the viseme for /w/ needs to have a little more initial dominance.

The synthetic speech matches very closely from the /w/ in frame 9 up until around frame 44, which is the start of the /eh/ in "set." This is because the viseme for /eh/ allows for a more open mouth. This could be due to the viseme definition being slightly incorrect for this subject or due to a lower emphasis on the word "set" in this utterance.

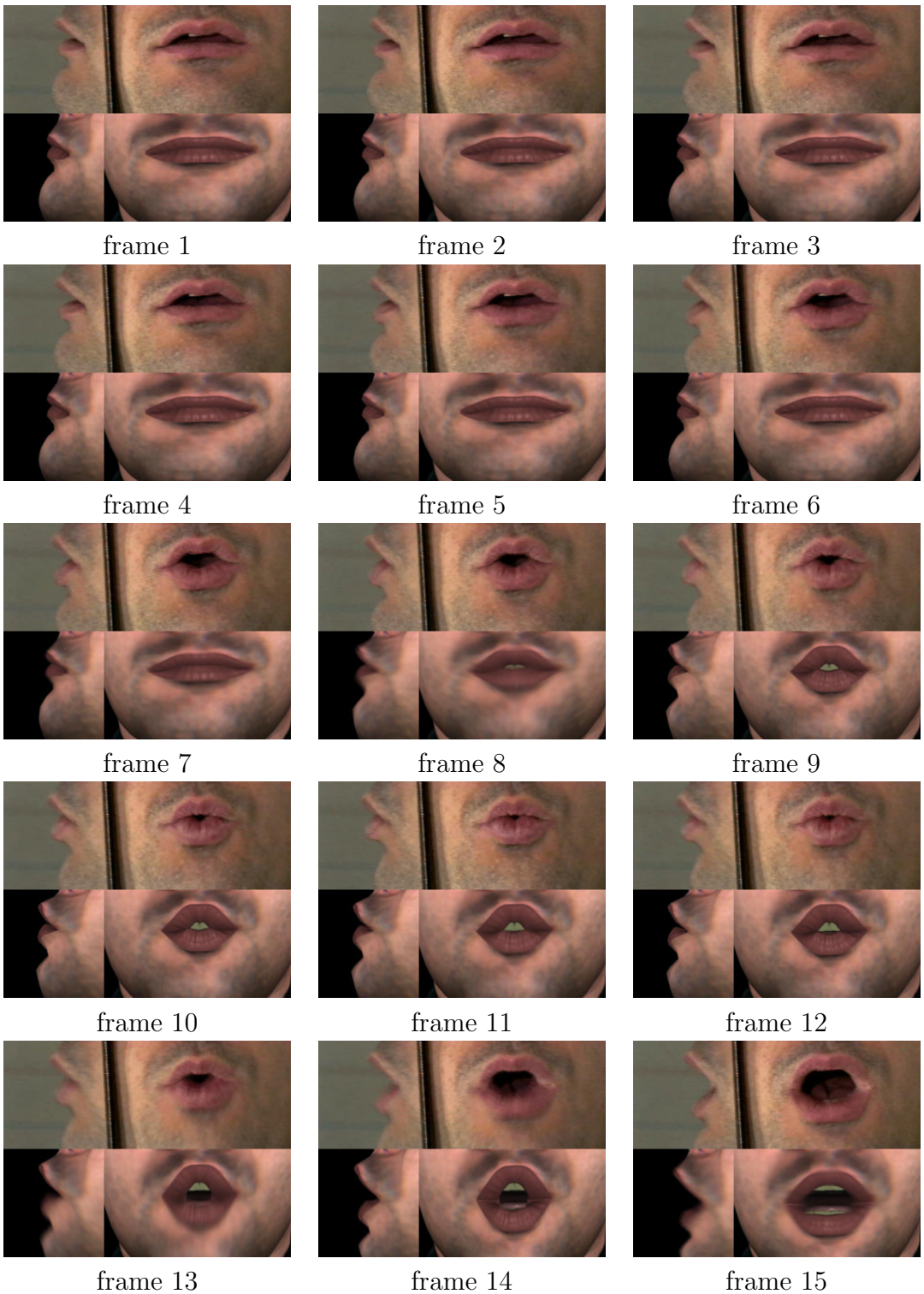


Figure 7.12: Frames 1-15 of a comparison of computer generated speech with recorded speech for the phrase: "Why did Ken set the soggy net on top of his head?".

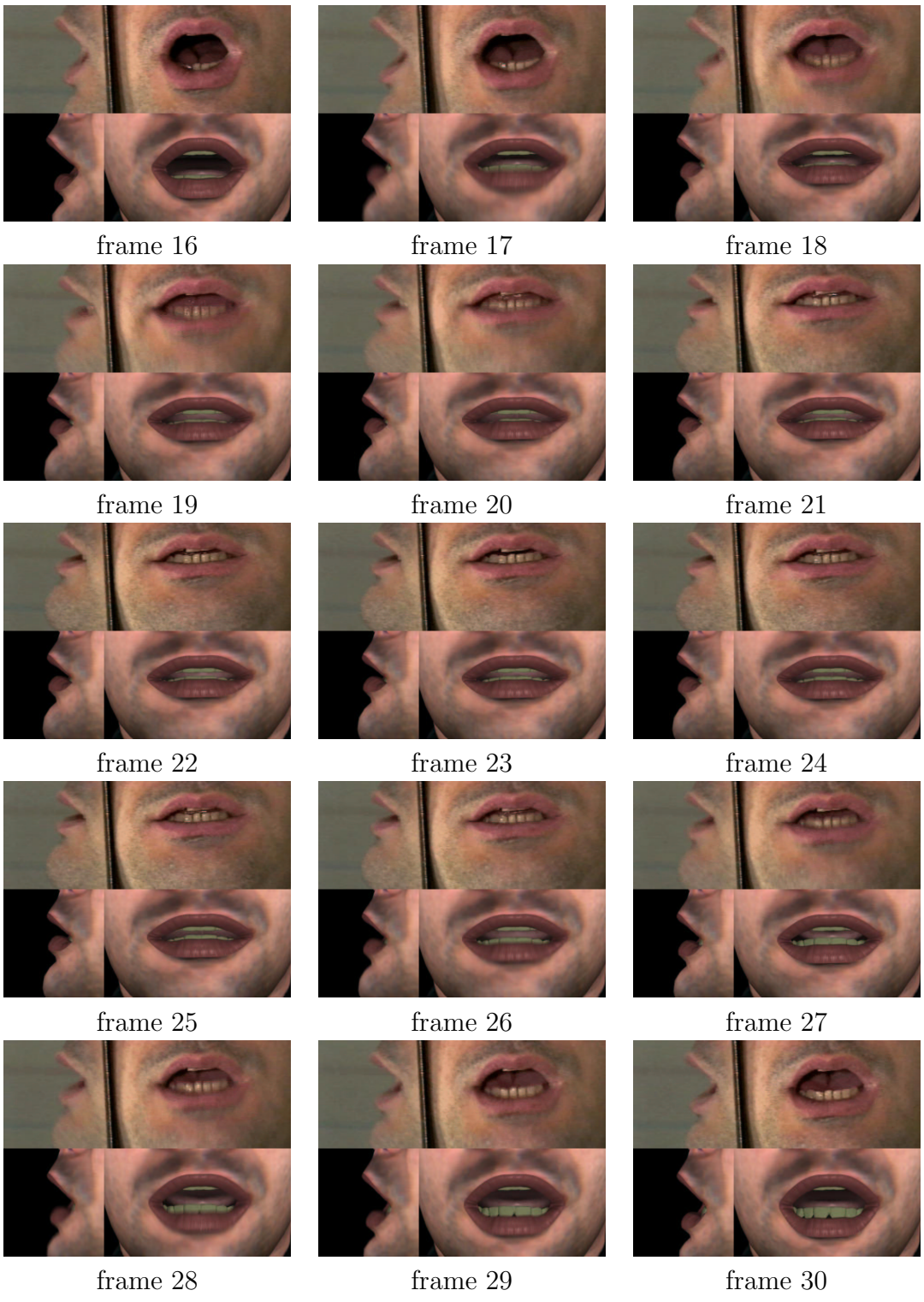


Figure 7.13: Frames 16-30 of a comparison of computer generated speech with recorded speech for the phrase: “Why did Ken set the soggy net on top of his head?”.

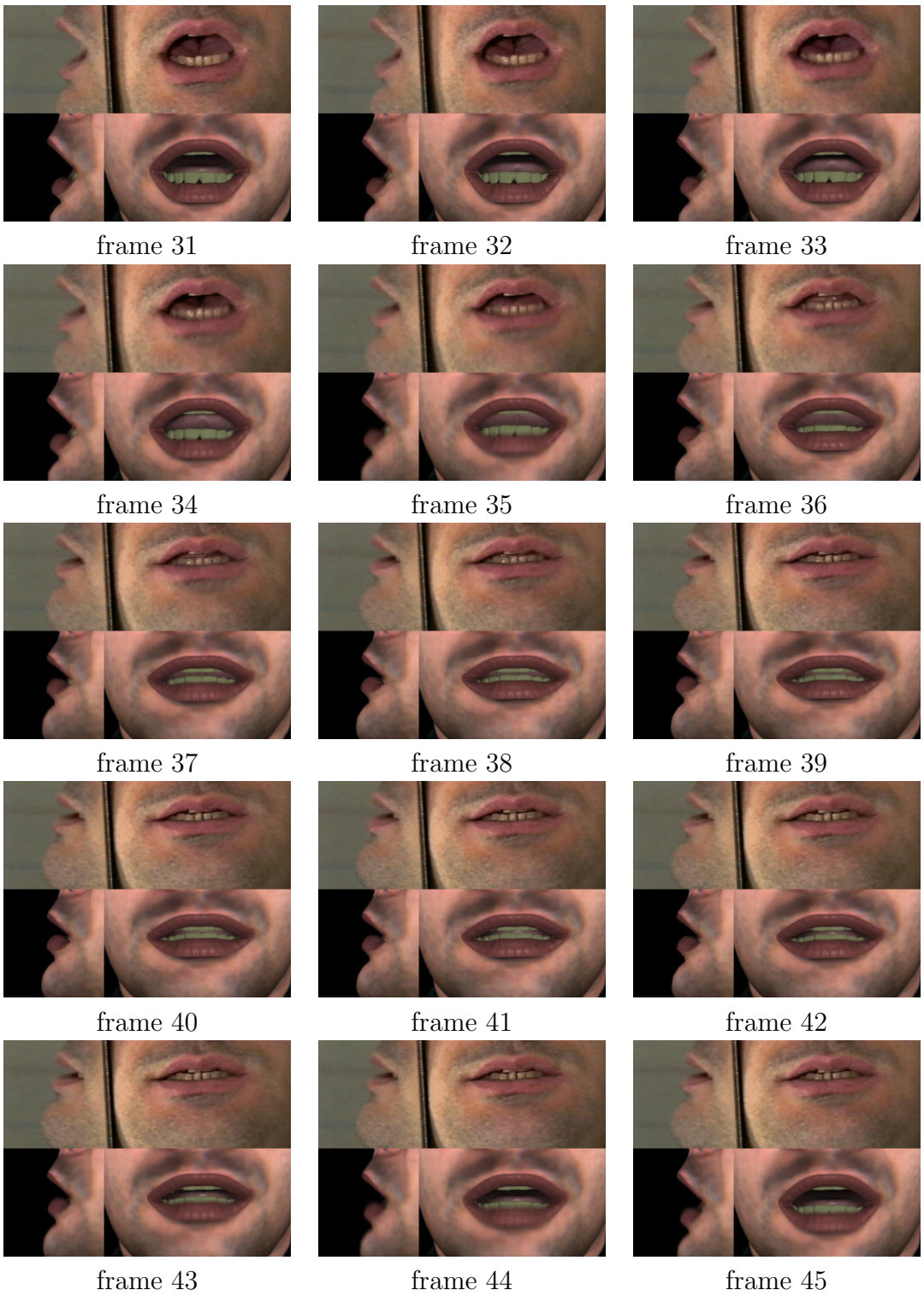


Figure 7.14: Frames 31-45 of a comparison of computer generated speech with recorded speech for the phrase: “Why did Ken set the soggy net on top of his head?”.

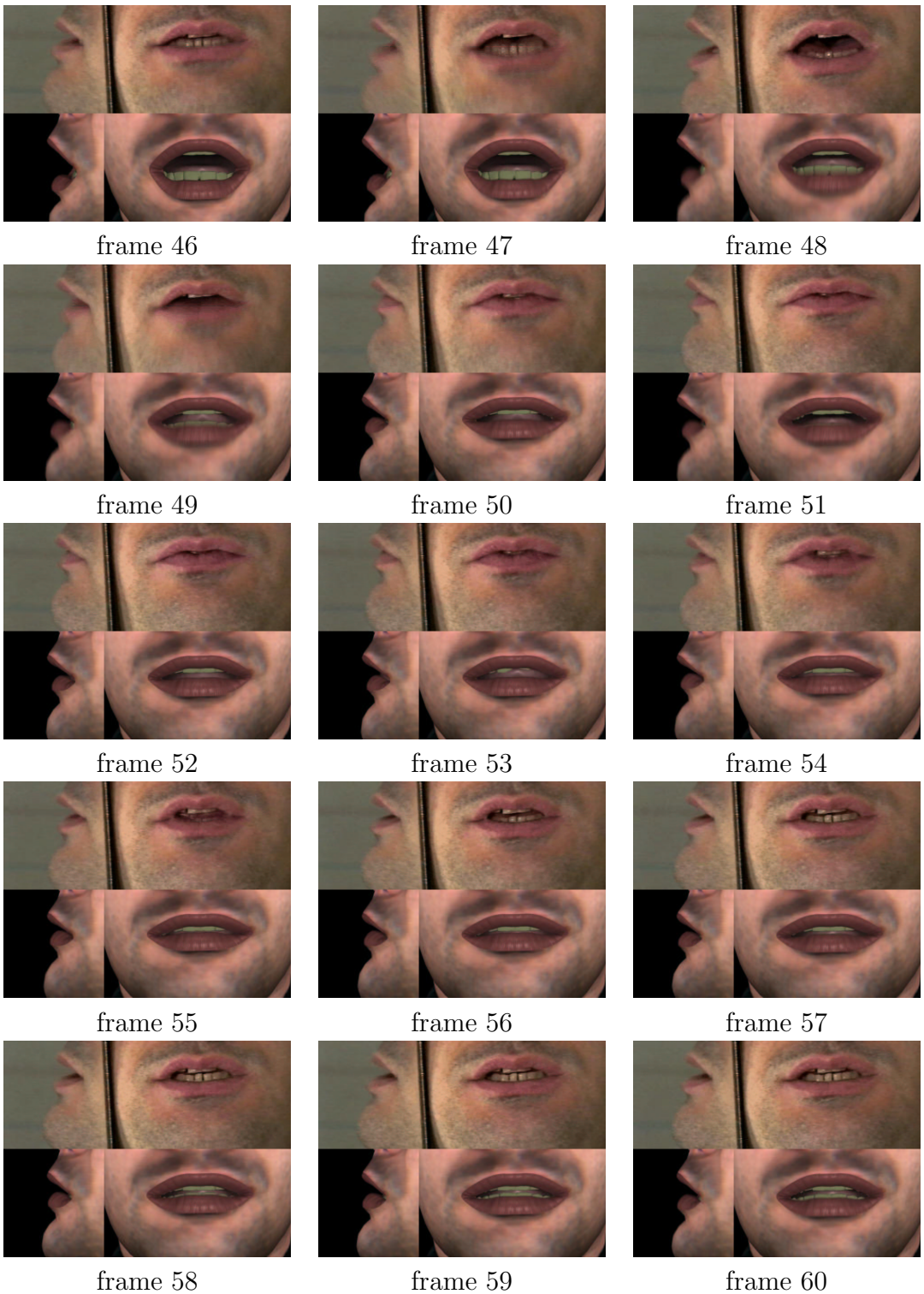


Figure 7.15: Frames 46-60 of a comparison of computer generated speech with recorded speech for the phrase: "Why did Ken set the soggy net on top of his head?".

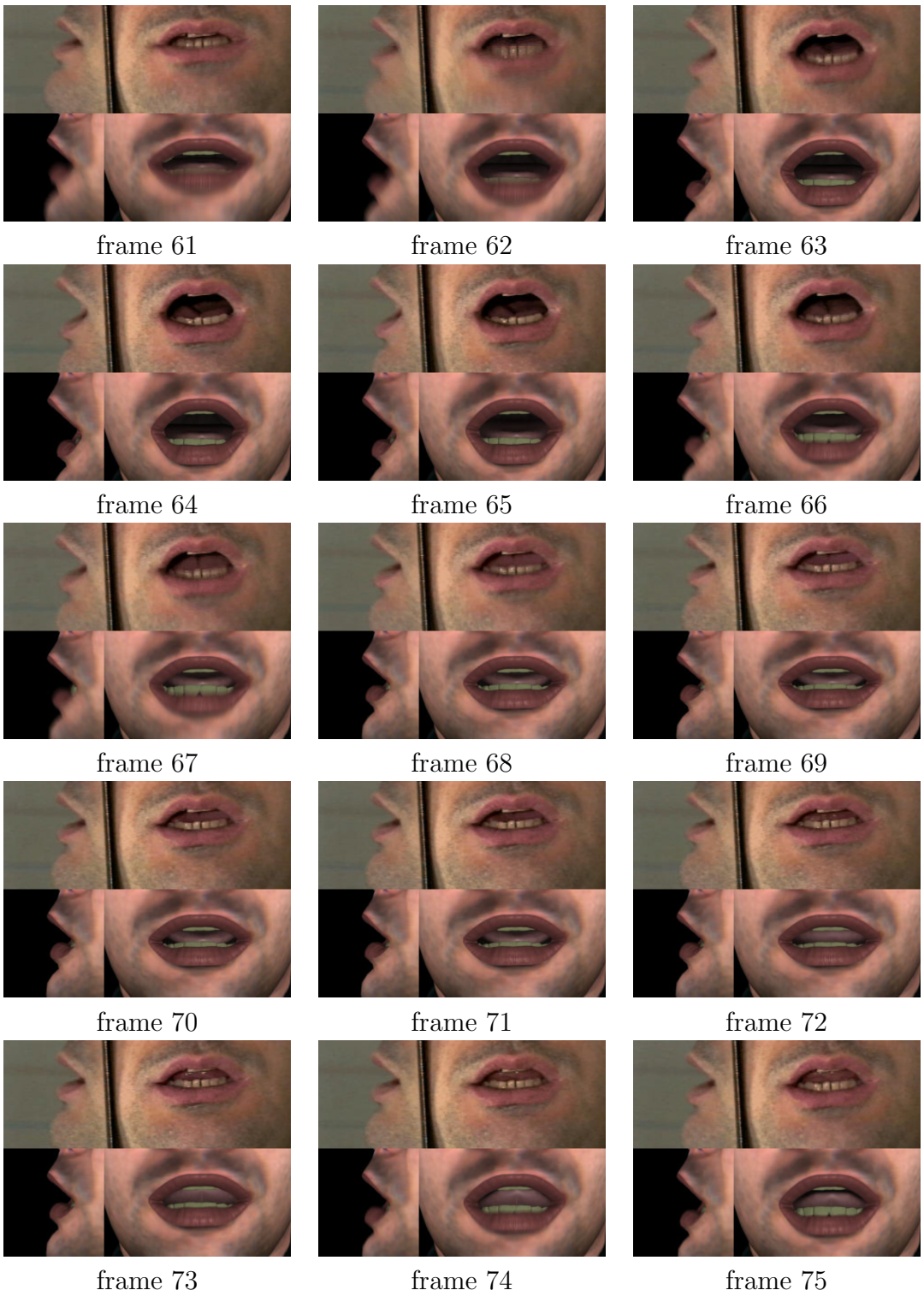


Figure 7.16: Frames 61-75 of a comparison of computer generated speech with recorded speech for the phrase: “Why did Ken set the soggy net on top of his head?”.

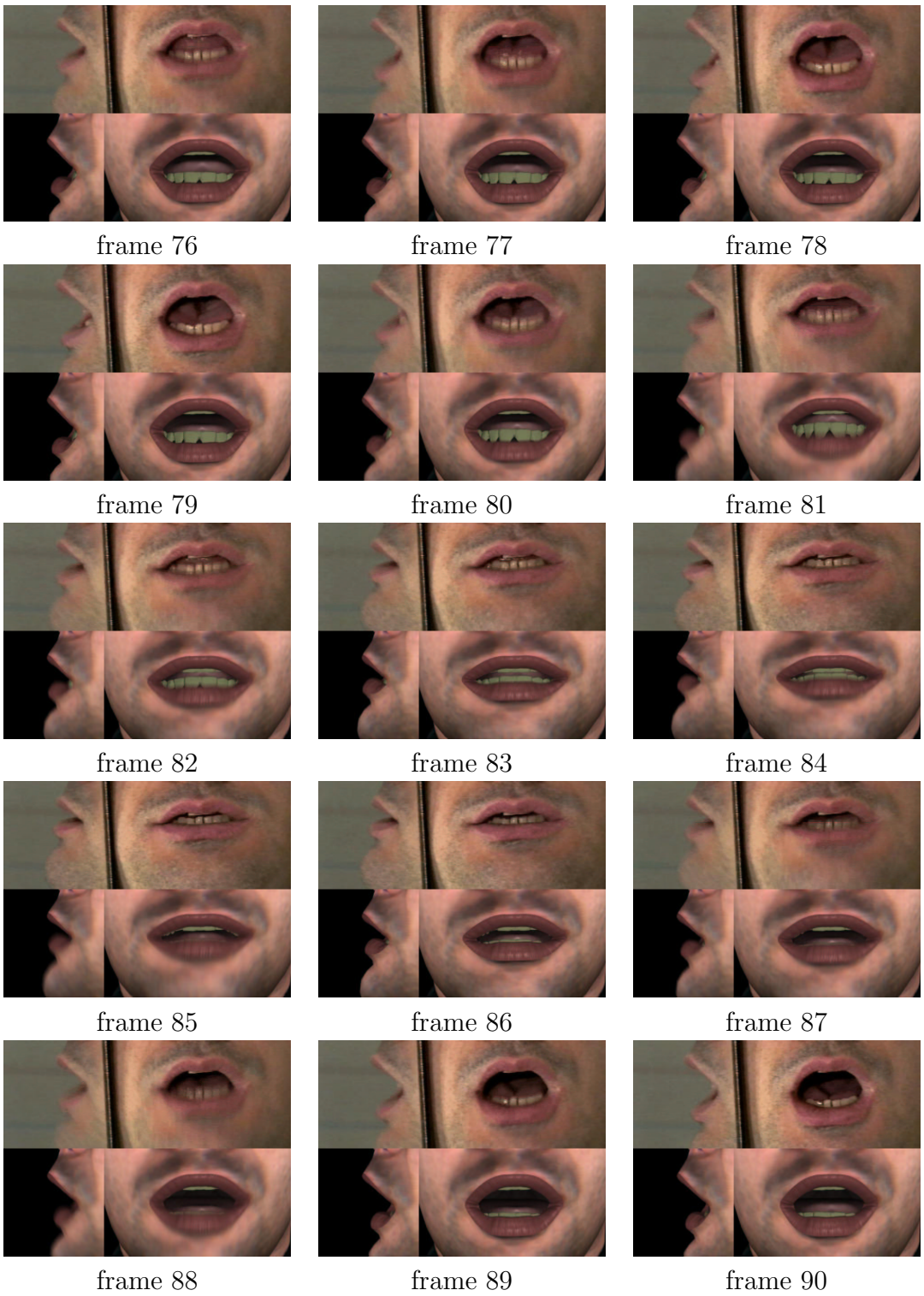


Figure 7.17: Frames 76-90 of a comparison of computer generated speech with recorded speech for the phrase: "Why did Ken set the soggy net on top of his head?".

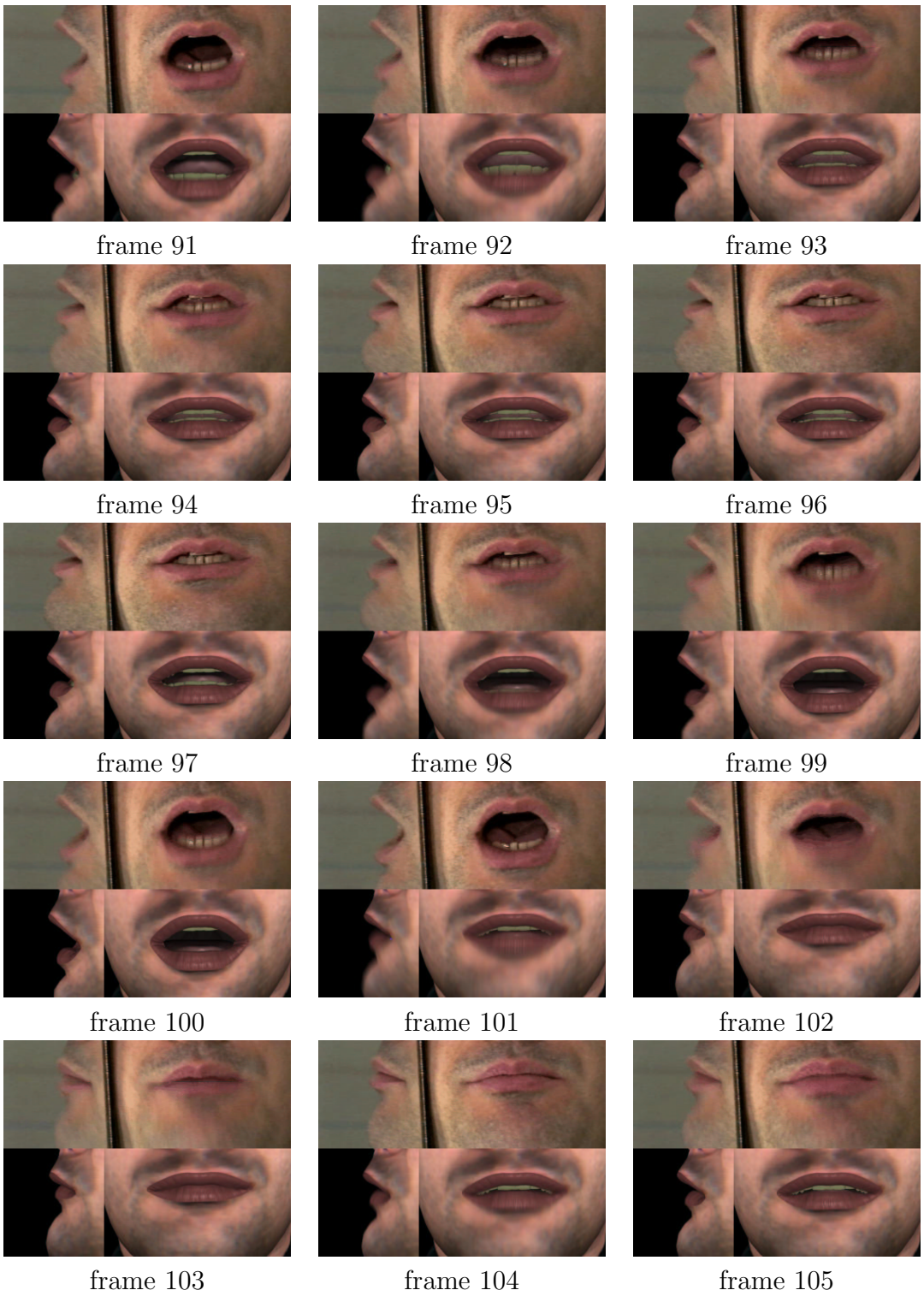


Figure 7.18: Frames 91-105 of a comparison of computer generated speech with recorded speech for the phrase: "Why did Ken set the soggy net on top of his head?".

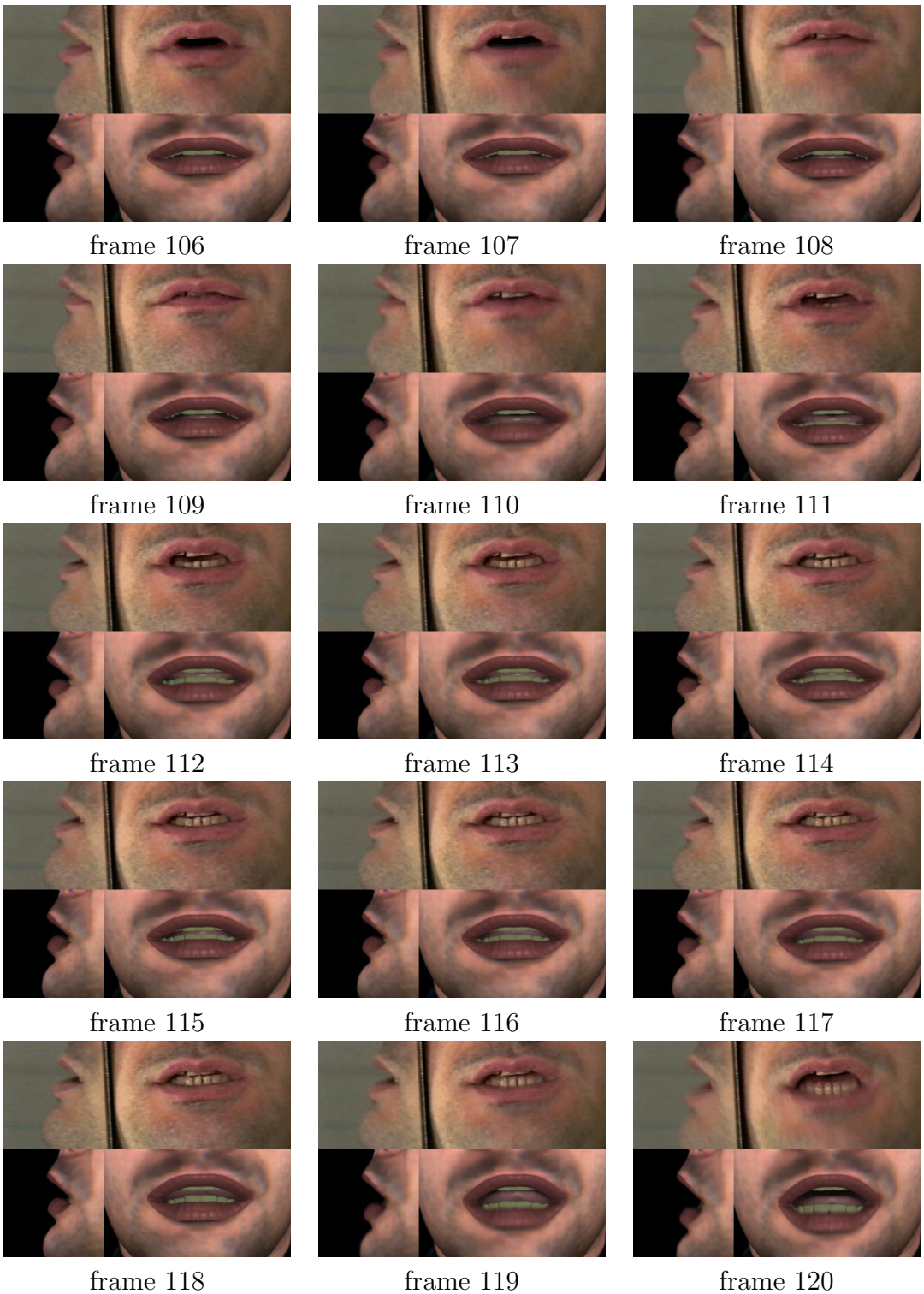


Figure 7.19: Frames 106-120 of a comparison of computer generated speech with recorded speech for the phrase: “Why did Ken set the soggy net on top of his head?”.

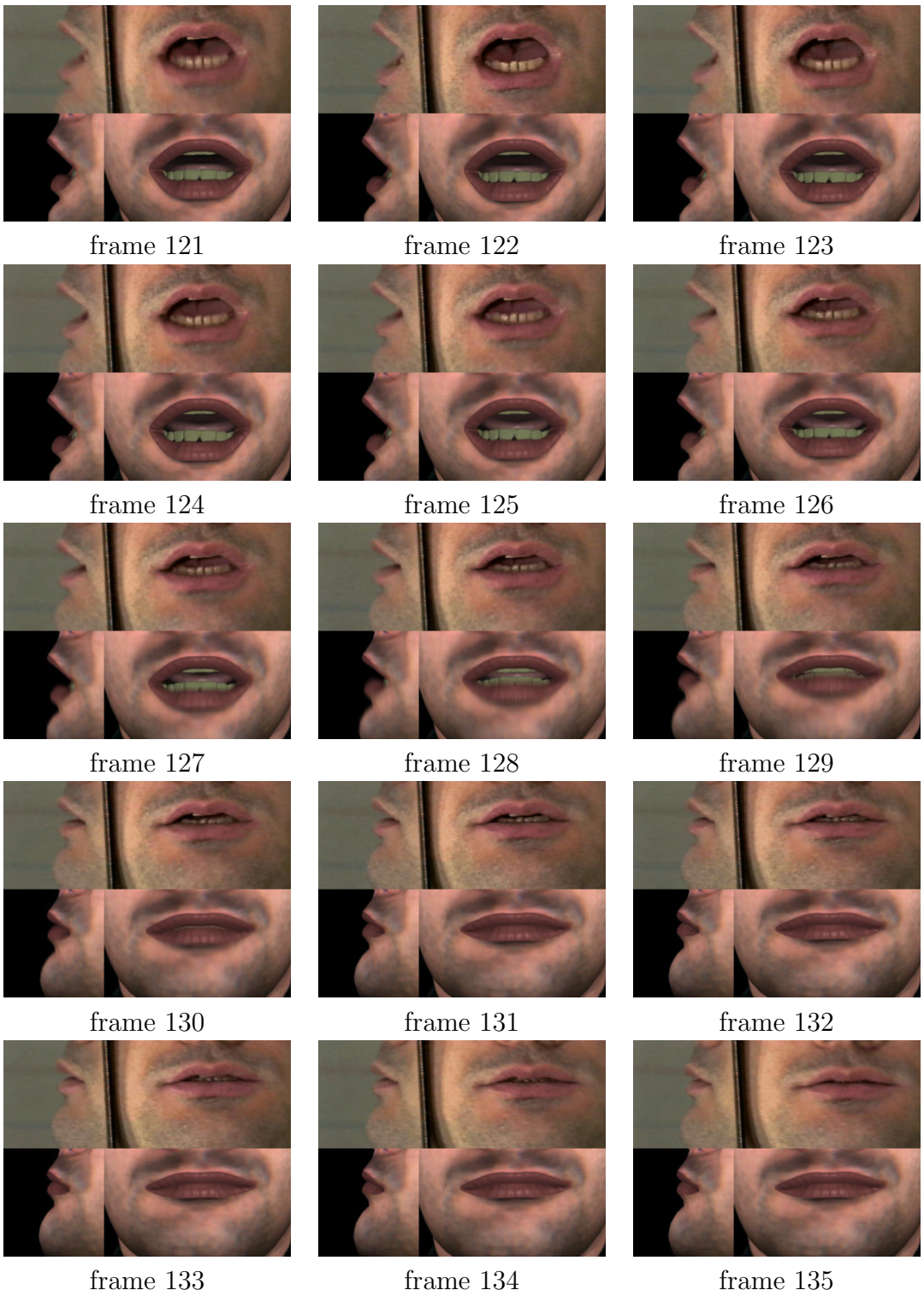


Figure 7.20: Frames 121-135 of a comparison of computer generated speech with recorded speech for the phrase: "Why did Ken set the soggy net on top of his head?".

Frames 50-54, which is the end of the /t/ and the beginning of /dh/ in "set the", also show a deviation. This is mostly due to an inadequacy in the lip model. The lip model does not properly allow the lips corners to get compressed together. The best examples of this effect are in frames 51-55 for /dh/ in "the" and in frame 105, which is the release of air for /p/ in "top."

In Frame 101, the synthetic mouth is starting to close about a frame earlier than the real speech for the /p/ in "top." This is most likely due to a slight difference in timing, but could also be due to an incorrect dominance for the /p/. The air is also released for the /p/ a little earlier for the synthetic speech.

Frame 106 is the start of the /v/ from "of" and shows a little difference from the synthetic speech. However, in the synthetic speech the lower lip shows some motion blurring which during animation will give the impression of quick movement as the lower lip tucks under the upper teeth to produce the /f/ and /v/ sound.

Frames 112-118 show a difference in the rounding of the lips for "his". This is due to the corners of the synthetic lips not properly compressing together. At the end of the phrase the synthetic mouth moves to a closed position a little quicker than the real lips, due simply to the neutral position of the synthetic face.

Besides the difference in the lip positions, there are some differences in tongue positions. The synthetic tongue has a tendency to move a little more quickly. This is because the coarticulation model only considers one parameter at a time. It does not understand that if the jaw is closing, then the tongue is also indirectly moving upward and therefore does not need to necessarily be moved directly upward. This is best seen in frames 32-35, which is the transition between the phonemes /eh/ and /n/ in "Ken", where the tongue moves towards the roof of the mouth too early.

Figure 7.21 and 7.22 show a typical animation produced by this system. This is the standard rendering for an off-line animation. The synthetic head is saying the phrase: "Hello world."

The goal of this work is to produce realistic animations. Unfortunately, the current printout of this dissertation cannot display an animation. However, the electronic version of this document can display animations. Figure 7.23 shows the first frame of two animations produced by this system. If you are viewing the electronic version, click on an image to view the animation. The "Abstract" animation is our facial model reciting the abstract of this dissertation. The "Comparison" animation shows a comparison between synthetic speech and real speech, and is the animation shown in Figures 7.12-7.20. All of these animations were produced using the research described in this dissertation.

Figure 7.24 contains images of the first frame of animations that shows further capabilities (other than speech) of our facial model. Note, these animations are only viewable in the electronic version of the thesis. Click on an image to view the animation. The "Tongue Tricks" animation is an animation showing the facial model sticking out its tongue and licking its lips. The "Happy Model" shows the facial model speeking with a smile. Both the "Tongue Tricks" and "Happy Model" animations were creating without using our coarticulation model. They were created using a single keyframe per viseme that are linearly interpolated.

7.3 Summary

Note that our goal here is only dynamic realism, and not individual realism. We do achieve a degree of individual realism since we designed the visemes based on the

Figure 7.21: Frames 1-15 of an animation of the phrase: "Hello world."

Figure 7.22: Frames 16-30 of an animation of the phrase: “Hello world.”

Abstract

Comparison

Figure 7.23: Animations produced by the techniques described in this dissertation. These animations are only available in the electronic version of the thesis. Click on an image to view the animation. The machine you are using to view the animations must be capable of playing the animaitons at 30 frames per second. The "Abstract" animation is our facial model reciting the abstract of this dissertation. The "Comparison" animation shows a comparison between synthetic speech and real speech, and is the animation shown in Figures 7.12-7.20.

Tongue Tricks

Happy Model

Figure 7.24: Animations produced by the techniques described in this dissertation. Note, these animations are only viewable in the electronic version of the thesis. Click on an image to view the animation. The "Tongue Tricks" animation is an animation showing the facial model sticking out its tongue and licking its lips. The "Happy Model" shows the facial model speeking with a smile. Both the "Tongue Tricks" and "Happy Model" animations were creating without using our coarticulation model.

subject, and the geometry of the model is that of the subject. However, our model shown is low-resolution and the lips were fitted to the low-resolution version. The lips therefore, do not quite match that of the real speech. Without prosodic features, the motion will not exactly match real speech that includes prosodic features. Our results show both static and dynamic realism, however we do not quite achieve static nor dynamic individual realism.

CHAPTER 8

FUTURE RESEARCH

The work presented here is in the early stages with more research yet to be performed. In this chapter we discuss research yet to be completed that is currently in various stages of development. This new research involves modeling, animating and improving our text-to-audiovisual-speech system. We also have a plan for a validation study.

8.1 Future Modeling Research

Our facial model can be used to create intelligible animated speech with realistic motion but more work remains. Our current focus is on the mouth area. General facial expressions are still required to complete the facial model. Also, as the muscles move the mouth area, the displacement of the muscle tissue causes changes in the facial surface in other areas of the face. These changes, due to the muscle movement need to be incorporated into the facial model.

Collision detection, however, is a difficult problem and solutions tend to be very slow. For a human computer interface using speech, real-time generation of the speech is required. Real-time animation leaves precious little computing time to perform accurate collision detection.

As we have gained experience creating animations and comparing the synthetic speech with real speech we have learned some of the inadequacies of our facial model. Collision detection, and the appropriate deformations due to collisions would increase realism. The lack of the collision is apparent for the tongue in phonemes such as /dh/, ("the"), where the tongue should flow around the upper teeth. For sounds that bring the lips together then apart, such as /p/, the corners of the mouth compress together and may actually stick together as they part.

8.1.1 Hair

Hair is an important visual characteristic of human heads. For a believable talking head hair is an important attribute to include. Not only the hair on top of the head, but also mustaches, beards and eyebrows must be modeled. The hair must not only be rendered realistically, it must also move realistically. See Section A.3.6 for a discussion of the issues involved in modeling and rendering hair and some current solutions.

8.1.2 Rendering Skin

We are currently researching a skin rendering algorithm that is a combination of the reflection model of Hanrahan and Kruger [HK93] and the rendering and modeling method of Wu et al. [WKMT96]. Hanrahan and Kruger [HK93] create a model for the subsurface scattering of light in layered surfaces using one-dimensional transport theory. This method replaces Lambertian diffuse reflection with a more accurate model of diffuse reflection that considers how the light interacts with different layers. Wu et al. [WKMT96] model the skin with both a micro and macro structure. At the micro level, they create a tile texture pattern using a planar Delaunay triangulation with a hierarchical structure and render the micro furrows using bump mapping. At

the macro level, they model wrinkles. Combining the techniques of Wu et al. and Hanrahan and Kruger may result in a skin model that has correct surface detail and accurate lighting properties.

8.2 Future Animation Research

Phonemes are commonly used as the smallest concatenative units of speech. And that is what we use for our research. However, the same phoneme does not always look the same. The two effects that model these differences are coarticulation and prosody. Our current research has a coarticulation model but does not handle prosody. Adding prosody to our system will allow us to create more realistic animated speech.

Our coarticulation model allows us to specify a viseme as a curve for each facial model parameter. The coarticulation model also allows us to define how visemes will influence other visemes in the utterance. The coarticulation model gives us greater control over the shape of the curve for the parameter for the entire utterance than in other animation techniques that use a single key position for each viseme. However, we have not yet had sufficient experience with the model for different individuals and languages to judge its generality. One source of information is using x-ray films of the vocal tract during speech.

Our coarticulation model does not consider the entire vocal tract, but only a single parameter, at the same time. In reality, there can be more than one way to produce the same sound. For instance, to change the volume of the oral cavity, the tongue can be raised or lowered directly, or indirectly by movement of the jaw. Another example is when the tip of the tongue needs to reach the roof of the mouth. When started with the mouth opened, the tongue is quite far away. If the jaw also is going

to be closed, the tip of the tongue can move most of the way without effort using the motion of the jaw to get closer to its target location. Then, the tip can finish the movement to the target once the jaw has slowed or halted its movement. Another example is movement of the lower lip which can be moved outward via direct muscle action or indirectly with movement of the jaw. A possible method to handle these multiple path situations is to use a rule based system that recognizes such situations and changes the normal control points for the affected tracks, to achieve the desired motion.

8.2.1 Syllables as Concatenative Units

A common approach, and the one we currently use, is to define speech as made up of phonemes. That is, phonemes are the smallest concatenative units of speech. Instead of using phonemes, syllables may be a better choice [FL78, Fuj92, SF93, Fuj00] as the smallest units of speech. Syllables should look more similar under different circumstances than phonemes do. However, there are many more syllables (on the order of thousands) than phonemes, so defining the set of syllables is a much more complex problem.

We are in the preliminary stages of testing the feasibility and realism of using syllables. Due to the sheer number of syllables that need to be defined, we choose to define them in stages. Using our current model of phonemes we can create the first definitions for the syllables. These definitions will slowly be evolved to increase realism of the syllables.

8.2.2 Cineradiographic Database

Munhall et al. [MVBT95] describe a collaborative effort to preserve cineradiographic (x-ray films) vocal track footage made in the 1960s and 1970s. These films allow a dynamic view of the vocal tract parts during speech. The films are quite valuable for studying the vocal tract. Currently, such films are rarely made, due to health risks to the subject of such high exposure to radiation. The films are in a format not easily accessible and there is fear of deterioration. To preserve the films and make them accessible to a wide range of researchers, Munhall et al. transferred the films to laser disc. The laser disc containing 25 films totalling 55 minutes of x-ray footage has been made publicly available. The films are of native speakers of Canadian French and Canadian English.

The films allow the rotation and transverse (forward and backward) movement of the mandible to be measured. In addition, 2D tongue and lip contours can be discerned. Many of the facial model parameters could be estimated from the films. The resulting data could be used to validate the coarticulation model as well as to tune the model. Collecting this data for the roughly 1600 frames is a daunting task. Although our eyes are able to detect the movement of the vocal tract parts, determining them in a single frame is not trivial. The effort involved in gathering the data by hand is tremendous. Instead, we are researching an automated method using computer vision and image processing techniques to track the jaw, tongue, and lips from frame to frame.

8.2.3 Prosody

The prosodic features of speech are generally taken to include length, accent and stress, tone, intonation, and potentially a few others. However, there is no universal consensus among phonologists about either the nature of prosodic features themselves or the general framework for their description [Fox00]. The term prosody comes from the Greek *prosodia* which means something like "song sung to music". Recently, it has come to refer to the principals of versification, covering such things as rhythmical patterns, rhyming schemes and verse structure. In linguistic contexts, prosody more frequently refers to such characteristics of utterances as stress and intonation, and moreover in prose rather than in verse.

Prosody is generally ignored in animated speech. The exception is Cassell et al. [CPB+94]. They describe a system that automatically generates animated conversations between multiple human-like agents with speech, intonation, facial expressions and hand gestures. A dialogue planner creates the text and intonation of the utterances that are used to create synchronized speech with facial expressions, eye gaze, head motion, and hand gestures.

Here we use the definitions of Fox [Fox00] for the prosodic features. *Length*, is the duration of the utterance. *Accent*, is the singling out of a particular element of the utterance with respect to its surrounding elements. Terms such as accent, accentuation, stress, prominence, emphasis, salience, intensity, and force, all describe this phenomenon. Accent can be divided into pitch-accent and stress-accent, where *pitch-accent* is a melodic manifestation of accent while *stress-accent* is a dynamic one. *Tone*, is an intrinsic property of a syllable, word or grammatical construction and

describes pitch. *Intonation*, is change in the pitch over the course of a phrase or sentence, for example a rising pitch for a question in English.

It is unclear just how prosodic functions may affect the visual manifestation of the speech. Obviously, length will affect the positioning of the targets, but should already be handled by the coarticulation model. Accent could possibly change target values, especially if super-enunciating a word or phrase. During super-enunciating the positions may be held longer and may be overly characterized. Without further data it is unclear if the coarticulation model is able to correctly reproduce this behavior. A change in pitch is usually accomplished by changing the position of the mandible. The pitch information is produced by the text-to-phoneme generation and possibly could be used directly to drive changes in the mandible.

We are currently evaluating the Converter and Distributor (C/D) model of Fujimura [Fuj00] for use in our system. The C/D model uses syllables as the basic units of speech (as describe above) and shows promise in increasing the realism of both the audio and video of synthetic speech. We believe the C/D model will be most beneficial when prosodic effects are considered. Stress is an important prosodic effect and it is an integral part of the C/D model. Therefore, implemented the C/D model should give us a foundation for realistic stress in our synthetic speech.

8.3 Improvements to TalkingHead

Adding a different character to the system involves fitting the lips and skull and picking the characteristic points. As well the eyes and nostrils need to be defined. This process needs to be streamlined and partially automated. In addition the lip,

tongue, and skull model are designed for human characters. Research into adapting the system for non-human characters would increase the utility of the system.

In order to use a new language, a voice for that language is needed. In addition, the visemes for all phonemes used by the voice must be determined. There are currently many different voices for many different languages that can be used and the definition of new voices is well defined. The visemes will still need to be determined to use these existing voices. The best results are obtained by using data from subjects that are native speakers of those languages. We have current plans for German and Chinese as we have access to native speakers of these languages.

8.4 Validation Study

Validating an animation method is not an easy task. There are rarely good explicit metrics that can be used to calculate the effectiveness. A common method is to have subjects give speculative answers to questions such as "Does it look realistic?", or "Does it look good?".

Since we are animating speech, we can measure comprehension rates as a metric. Our planned study is to measure rates of speech comprehension with just sound and with both sound and video. The rates will be measured for both real and synthetic speech and compared. As well we plan to compare speech-reading (video only) rates between hearing and non-hearing subjects.

8.5 Summary

We have created a facial model capable of representing realistic mouth shapes required during speech. We have also developed animation techniques that allow

realistic motion of the mouth during synthetic speech. Our research is in the early stages but still shows promise that individual realism can be obtained.

The synthetic speech generated by our system exhibits static and dynamic realism. We can quickly generate comprehendable synthetic speech using only text. A system such as ours can be used to create human computer interfaces using speech, which is a natural mode of communication for humans. With the addition of voice recognition techniques a user can converse with the machine using speech, which is a comfortable natural method of interaction for humans. Such a system would allow talking kiosks for visitor information, interactive talking catalogs, interactive talking web site greeters, and interactive talking help systems to name a few.

APPENDIX A

FACIAL ANIMATION OVERVIEW

A.1 Background

Using technology to produce human faces and animating human faces started in the early 1970s with 2D and 3D methods. Since then, there have been hundreds of research papers describing the efforts of numerous researchers in the diverse field of facial animation. The research has covered topics on facial modeling, rendering, animating speech, animating expressions, creating caricatures, modeling hair, rendering hair, modeling skin, modeling wound closure, simulating facial muscles, simulating surgery, compression of videos with faces, virtual avatars, and virtual teleconferencing to name just a few.

An interesting use of 2D faces is as a method to visualize multivariate data. Chernoff [Che71, Che73, CR75] developed the technique in the early 1970s with the idea that visualizing data with more than four dimensions could be cast into a familiar domain, that of recognizing differences in facial attributes. Cartoon faces with 18 parameters are used to graphically represent data with up to 18 dimensions. Chernoff conducted experiments that suggested the order of assignment of dimensions to the parameters could change the classification error rate by 25%. Flury and Riedwyl

[FR81] develop another method using asymmetrical faces effectively doubling the number of variables that can be displayed while also removing some of the problems with Chernoff faces.

Parke [Par72] creates the first animation of a 3D face by hand digitizing expressions and facial geometry then defining key positions. Animation results from interpolation of the keyframes. This is very tedious work, so Parke [Par74] later developed a parametric facial model based on empirical and traditional hand-drawn animation methods. The parameters, which define a linear interpolation between two extremes, come in two flavors: conformation parameters and expression parameters. The conformation parameters define the shape of the face and the expression parameters are used to create the animation.

Gillenson [Gil74] presents a method to create 2D drawings of individuals, such as police artist sketches. The resulting images can represent the attributes of an individual effectively enough for recognition. This system is strictly 2D and is only used for static poses.

Some of the earliest synthetic facial animation was created using oscilloscopes, and it was the first real-time facial animation. Boston [Bos73] used an oscilloscope to create mouth shapes presented to a lip reader who was then able to recognize a small vocabulary and sentences of speech represented by the oscilloscope.

Montgomery [Mon80] augments earlier work with the addition of nonlinear interpolation between frames as well as forward and backward coarticulation approximation. The system, which draws lip outlines on a CRT from data hand-captured from video frames, is designed to test lip reading ability.

A.1.1 Definition of Facial Animation

Figure A.1 is a schematic of a typical facial animation system. A facial animation system will contain one or more animating and modeling components. Animating consists of changing the model shape over time to produce animation. This is generally done by setting the parameters, be it high-level or low-level, of the facial model for each frame. An animation technique may define its own parameterization, generally high-level, with a mapping from that parameterization to the facial model's parameterization.

Modeling is the rest of the process. Facial modeling includes defining the geometry of the face, parameterizing, animation controls, and rendering. A facial model will therefore have some way to define the surface of the face, a way to specify a specific shape (a parameterization), a mapping of the parameterization to a specific shape, and rendering. The various parts of the facial model may vary in complexity with some possibly missing. A facial model might also have more than one of each of the parts.

The goal of the animation is to produce the desired motion. This motion could be realistic or artistic. The animation could be designed to produce intelligible speech, tell a story, entertain, teach, reproduce observed motion, virtual reality, etc. The goal of modeling is usually to support the animation at minimal cost. Its complexity is directly related to the goal of the animation. Occasionally, especially in medical applications, there are situations where realistic static views of the face are the goal.

This overview will give the reader the broad base of knowledge needed to create a facial animation system. The anatomy of the face will be discussed along with techniques for modeling and animating faces. Parameterizations of the motion of the

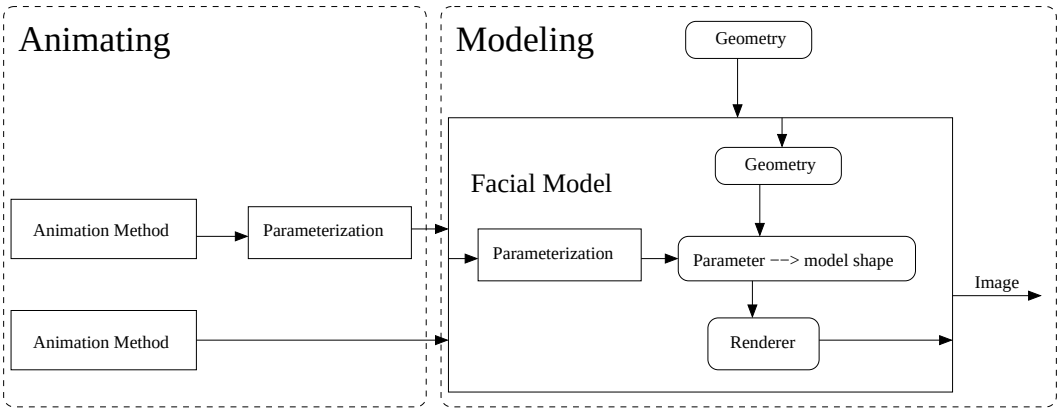


Figure A.1: A schematic of a facial animation system. The animation technique(s), which may or may not define a parameterization for the model, change the model over time. The facial model will accept a definition of a shape of the model in the form of a parameterization. The definition is mapped to a specific shape of the surface that is rendered to produce an image.

face are presented. The special requirements of specific types of applications such as speech and surgery will also be discussed. Vision techniques and their uses will be reviewed as well as facial animation standards.

A.2 Anatomy

Anatomical knowledge is extremely important for solving the problems of facial animation. The underlying physical structure determines the surface of the face and a facial animation system should consider this structure. The interaction between the soft tissues and bone determine how the facial surface deforms. The muscle contractions not only pull the skin over bone, muscle and fat, but also change the surface due to the displacement of their volume.

Consideration of these interactions is necessary for realistic facial animation. In computer graphics, we are generally interested in efficient simulations of reality. Modeling the tissues at the cellular level is computationally expensive, but we do need to understand how different tissues behave. Those viewing facial animation are very critical of the realism of the animation since we are trained from just after birth to read faces. We use this training to determine not only what message is being delivered, as in speech, but we also determine the emotion of the message and the emotion of the speaker.

For a good feel for how the muscles and bone affect the face, refer to a book such as Faigin [Fai90] that describes anatomy for the artist. Any good general anatomy reference (e.g. Gray [Gra77]), speech and hearing anatomy reference (e.g. Dickson and Maue [DM70] or Bateman and Mason [BM84]) or facial anatomy reference (e.g. Brand and Isselhard [BI82]) can be an invaluable source of information.

A.2.1 Skin

The skin is the principle organ of the sense of touch and is also a covering for the protection of the deeper tissue. The skin plays an important role in the regulation of body temperature and is an excretory and absorbing organ. It consists of a layer of vascular tissue, named the *derma*, *corium* or *cutis veru*, and an external covering of epithelium, called the *epidermis* or *cuticle*. On the surface of the derma are the sensitive *papillae* and embedded within the derma or just beneath it are certain organs with special functions, namely, the *sweat-glands*, *hair-follicles*, and *sebaceous glands*. The epidermis is non-vascular and consists of stratified epithelium that is molded on top of the papillary layer of the derma forming a defensive covering and limiting

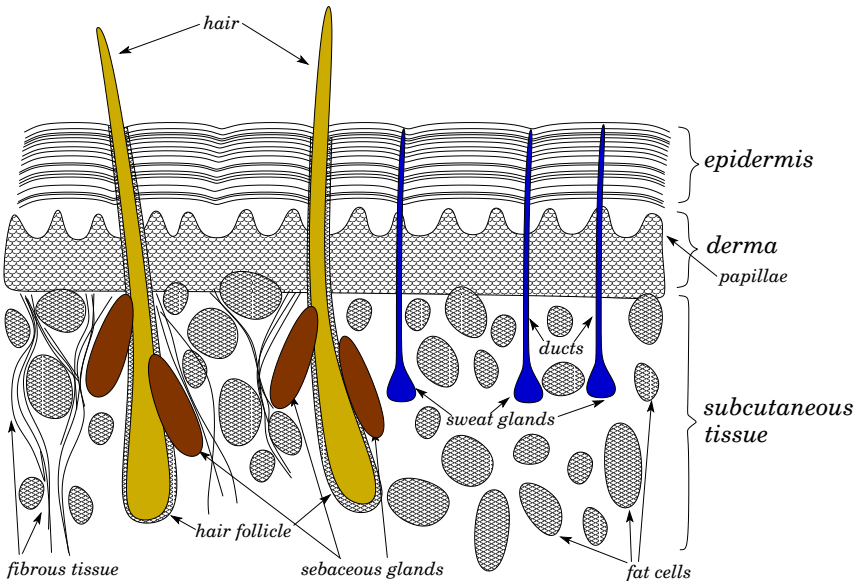


Figure A.2: A cross section of the skin.

the evaporation of water. The derma is tough, flexible and highly elastic in order to defend the parts beneath. Figure A.2 shows the different layers and parts of the skin.

In developing a finite element model of the skin, described in section A.3.2, Larrabee [Lar86b] goes into detail about the structure and behavior of skin. Skin is an anisotropic complex living organ that exhibits nonlinear stress-strain properties and viscoelastic or time-dependent behavior. An elastic object can be modeled by Hooke's Law (spring) and viscosity can be modeled by Newton's Law (dashpot), while a viscoelastic material can be modeled as a combination of the two.

Skin gets its mechanical behavior from elastin (4% dry weight) fibers, collagen (72%) fibers, and ground substance (20%). Elastin is the only known mammalian protein to have truly elastic properties. Elastin exists as a composite assembly of small tapering rope-like fibers, about $1.5nm$ in width, that form a network. They

are paired structures that are periodically linked longitudinally. Collagen is wavy and random while relaxed and, as it becomes strained, it orients along the strain direction. The amorphous matrix or ground substance in which collagen and elastin fibers are embedded seem to affect the time-dependent behavior of skin (creep, stress relaxation, hysteresis) and acts as a cementing connective tissue.

Initially as strain goes up, the stress stays at zero due to deformation of the delicate elastic network. As the random collagen fibers are oriented, the stress goes up slowly. Then once they align, the stress goes up fast almost linearly. Skin behaves differently by direction, corresponding to Langer's lines in a highly nonlinear fashion. Properties of skin such as hysteresis, creep and stress relaxation can be found in Larrabee [Lar86c].

A.2.2 The Skull

The skull consists of the cranium, which houses the brain, and the skeleton of the face. Except for the mandible and the auditory bones, the bones of the skull are connected in specialized joints know as *sutures*. The mandible is the only movable part of the skull. For facial animation, the names of the individual bones are only important when discussing the bony attachments of the muscles. Table A.1 lists the bones of the skull as well as the hyoid. The hyoid bone, which is located in the neck separate from the skull, can be considered the skeleton of the tongue and therefore included with the bones of the face.

Name	Description
Ethmoid	Forms part of the cranial base and part of the skeleton surrounding the nasal cavities.
Frontal	Forms the anterior portion of the cranium and the vaults of the orbital cavities.
Hyoid	Horseshoe-shaped bone in the neck above the larynx.
Lacrimal	Located on the medial surface of the orbital cavity between the maxilla and ethmoid bones. Paired
Mandible.	Forms the lower jaw.
Maxillary	Forms the upper jaw. Paired
Nasal	Forms the bridge of the nose between the orbital cavities. Paired
Parietal	Forms the vault and part of the side wall of the cranium. Paired
Occipital	Forms the posterior portion of the cranium and part of the posterior cranial floor.
Palatine	An "L"-shaped bone which forms the posterior one fourth of the hard palate and part of the lateral wall of the nasal cavity posteriorly. Paired
Sphenoid	Forms part of the base and lateral walls of the cranium and the vault of the pharynx.
Temporal	Forms part of the cranial wall and base and houses the middle and inner ear. Paired
Vomer	Flat, plowshare shaped bone that forms the posterior and inferior part of the nasal septum.
Zygomatic	Forms the bony prominence of the cheek and part of the lateral wall and floor of the orbital cavity. Paired
Inferior conchal	Extends horizontally along the lateral wall of the nasal cavity inferior to the ethmoid bone. Paired

Table A.1: Bones that comprise the skull.

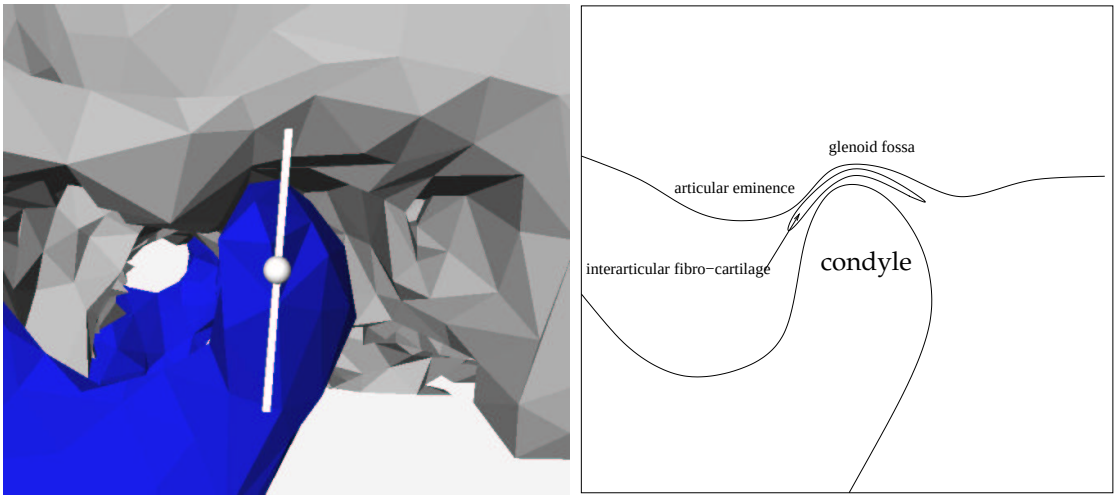
The temporomandibular joint

The temporomandibular joint is the only movable joint of the skull and allows the mandible (jaw bone) to move. Movement of the mandible allows the mouth to open for the intake of sustenance, mastication[4], and speech.

The temporomandibular joint is a diarthrodial ginglymous [NS62] (sliding hinge) joint that allows the mandible a large scope of movements. The mandible acts mostly like a hinge, with two separate joints acting together, and each of the joints having a compound articulation. The first joint is between the condyle and the interarticular fibro-cartilage while the second is between the fibro-cartilage and the glenoid fossa. The condyle of the mandible is nested in the mandibular (or glenoid) fossa of the temporal bone as seen in Figure A.3. The upper part of the joint enables a sliding movement of the condyle and the articular disk, moving together against the articular eminence. In the lower part of the joint, the head of the condyle rotates beneath the under-surface of the articular disk in a hinge action between the disk and the condyle.

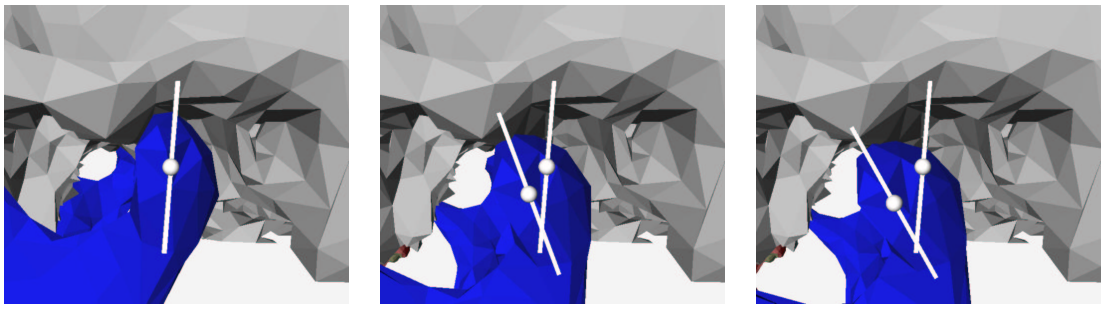
The mandible can be thought of as a joint with three degrees of freedom, with the jaw moving in (retraction or retrusion) and out (protraction or protrusion); side-to-side (lateral movements); and open and closed. During opening, the mandible rotates downward in a hinge action as the condyles move forward and does the reverse during closing. Figure A.4 shows how the condyle rotates and slides with the bar representing an up axis and the sphere representing the point of rotation. In figure A.4c it can be seen how the condyle has both translated down and forward as well as rotated. During protrusion, the condyles slide forward on the articular eminences with the teeth remaining in gliding contact. For retrusion, both condyles move rearward to

[4]Mastication is the mechanical division of food.



a b

Figure A.3: The temporomandibular joint a) shown on our 3D geometry of the skull and b) a cross-sectional drawing of the joint showing the bony parts of the joint as corresponding to a). The mandible is shaded differently from the skull and the condyle on the right side of the skull can also be seen.



a b c

Figure A.4: The temporomandibular joint and its center of rotation are shown a) in the rest position, b) about half open and c) full opening of the mouth. The joint is a sliding hinge joint with the condyle of the mandible rotating as it slides from the glenoid fossa to the articular eminence.

settle in the mandibular fossa with the teeth remaining in gliding contact. Lateral movement is achieved by fixing one condyle and drawing the other condyle forward.

Hyoid

The hyoid bone is considered the skeleton of the tongue. It has about 30 muscle attachments and attaches to the styloid process of the temporal bone by ligaments.

The teeth

The 32 permanent teeth are firmly anchored to either the maxillary bones or the mandible in sockets called *alveoli*. The upper teeth make up the *maxillary arch* and the *mandibular arch* contains the lower teeth. The biting surfaces do not conform to a flat plane. Most often the maxillary occlusal plane, as seen in profile, inclines slightly downward from the canine to the first molar and then rises again in the molar region. The resulting curve, which is slightly convex for the maxillary teeth and concave in the lower arch, is known as the *curve of Spee*. In the coronal plane, the posterior teeth are arranged so that the upper molars are tipped somewhat bucally and the lowers somewhat lingually; this is known as the *curve of Monson*. Normally in *occlusion* (contact between upper and lower teeth), the upper incisors cover the incisal one-third of the lower incisors (middle four teeth.)

A.2.3 The Tongue

The tongue consists of symmetric halves separated from each other in the mid line by a fibrous septum. Each half is composed of muscle fibers arranged in various directions with interposed fat and is supplied with vessels and nerves. The complex

arrangement and direction of the muscular fibers gives the tongue the ability to assume the various shapes necessary for the enunciation of speech [Gra77].

The muscles of the tongue are broken into two types: intrinsic and extrinsic. The extrinsic muscles have external origins and place the tongue within the oral cavity. The intrinsic muscles are contained entirely within the tongue with the purpose of shaping the tongue itself. Extrinsic muscle fibers interweave with intrinsic and intrinsic interweave with each other [BM84, Gra77].

A.2.4 The Lips

The lips, composed of muscles, nerves, vessels, areolar tissue, fat and glands, are covered by skin on the outside and by mucous membrane on the inside. The lips, *labium superius oris* and *labium inferius oris* (upper and lower lips respectively) create a highly variable valve for the anterior portion of the mouth. The lips assist in the intake of solid and liquid foods into the oral cavity and prevent the contents from escaping during mastication and swallowing. The lips also aid in the production of speech particularly in the consonants b, p, f, v, m, and w, known as labial sounds.

In the red or *vermillion zone* of the lips (a characteristic feature of mankind only), the skin (*epithelium*) covering the lips is thin with a rich vascular supply, which provides the red color. The point of contact of the lips during closure is known as the *rima oris*. The lateral margins of the mouth make up the *angle of the mouth* or *angularis oris*. A slight prominence at the midsection of the upper lip, the *tubercle*, appears in most subjects. The shallow depression extending from the upper lip to the nose is the *philtrum*, which along with the rounded parts of the lips forms *Cupid's bow*.

In youth, there is no lateral boundary, but with aging, a furrow beginning at or close to the mouth corner appears. This groove, the *labiomarginal sulcus*, runs in a posteriorly convex arch toward the lower border of the mandible. The *nasolabial groove*, present in almost all persons, begins at the wing of the nose and courses downward and laterally to pass at some distance from the mouth corner forming an upper border separating the cheek from the upper lip.

The point of the chin is known as the *mentum* and the sharp, deep groove that separates the lower lip and the chin is the *labiomental groove* with its depth determined by age, fullness of lower lip and prominence of the bony and soft chin.

The lip coverings are tightly bound to the connective tissue covering the muscle fibers causing the skin and mucous membrane to closely follow movements of the muscle. The skin of the face varies with sex, the male having the thicker and firmer skin. The chin does not move when the lower lip is depressed in most males, and the female is likely to show more of the upper teeth and gum during smiling.

A.2.5 The Cheeks

The cheeks, or *buccae*, serve as the lateral boundary of the oral cavity. Several muscles that act upon the lips at the angle of the mouth are just under the skin of the cheeks. The buccinator, which forms the mobile part of the cheeks, is made up of a deeper layer of muscle, with attachments to both upper and lower jawbones. The buccinator draws the cheek in during contraction. The variable *buccal fat pad of Bichat*, also called the *suckling pad*, is found in the cheeks giving it a fleshy feel.

A.2.6 Vocal Tract

Speech is produced using the vocal tract, which is made up of organs of the body whose functions include a number of biological roles. That is, there is no speaking apparatus apart from the structures used for the intake of air and food. Speech production can be broken into four overlapping aspects which are:

- an *energy source*: expiratory air;
- a *vibrator*: the vocal folds;
- *resonators*: oral, nasal, and pharyngeal cavities;
- *modifiers*: the articulators: the lips, tongue and soft palate;

The expiratory muscles send the inhaled air in the lungs under pressure up the windpipe to operate the vocal folds in a vibratory pattern for sound production. The sound stream (or air stream for whisper-like sounds) leaves the larynx and resonates in the cavities of the throat, oral and nasal chambers. The articulators of the oral cavity halt, constrict and redirect the sound stream to form chunks of sound that can be meaningfully interpreted. Figure A.5 shows a schematic of the vocal tract cavities.

A.2.7 Muscles

A brief description of the muscles involved in facial animation is given here. Only the muscles of the face and those of the vocal tract are considered. There is no clear consensus among the different anatomy texts on the names of the muscles and on the number of muscles. The muscle fibers can intermingle so much that the actions of the fibers becomes confusing when trying to determine when fibers belong to one muscle or another, or a part of a muscle or a different muscle altogether. For example, the

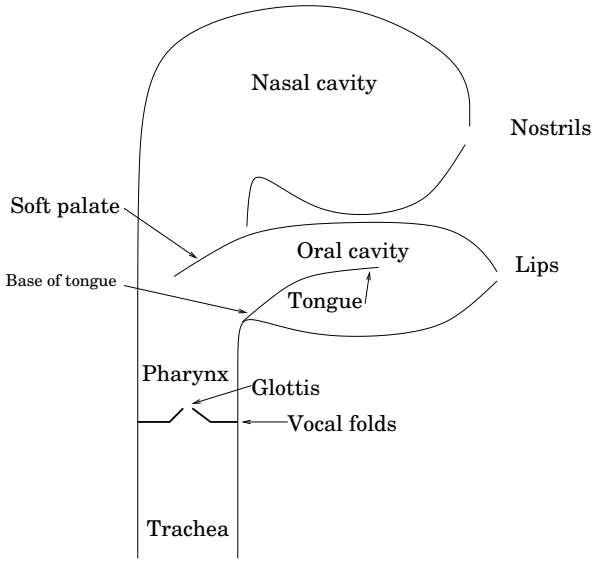


Figure A.5: A schematic of the vocal tract, which is made up of a series of valves and cavities to produce distinct sounds for communication.

orbicularis oris, which is made entirely of fibers from the other muscles that insert into the lips, has a clearly defined action separate from the primary purpose of the individual muscles.

The *mandible*, or *lower jaw*, is the only part of the skull that moves. Movement of the mandible allows access to the oral cavity for the intake of food and air, mastication and speech. Table A.2 lists the muscles that directly move the mandible.

In the cranial region the skin is the thickest of anywhere on the body with dense location of hair-follicles. The *occipito-frontalis* muscle covers the top of the skull and consists of two muscular slips with a tendinous aponeurosis between the two. The occipital portion is called the *occipitalis* and the frontal portion is called the *frontalis*. The ears in man are almost immovable and the three muscles surrounding them are rudimentary. Table A.3 lists the muscles of the scalp and describes their actions.

Name	Action
Digastric	Aids in lowering the mandible, raises and moves the hyoid bone forward or posteriorly, and indirectly raises the larynx.
Geniohyoid	Aids in lowering the mandible and draws the hyoid bone forward and slightly upward.
Mylohyoid	Aids in lowering the mandible and draws the hyoid bone forward and upward.
Masseter	Raises and may retract the mandible.
Platysma	Assists in lowering the mandible.
Lateral pterygoid	Lowers and protrudes the mandible by pulling the condyles forward with lateral movements by alternate contraction.
Medial pterygoid	Raises and protrudes the mandible with lateral movements by unilateral contraction.
Temporalis	Raises and retracts the mandible.
Stylohyoid	Draws the hyoid bone up and back.

Table A.2: The muscles of the mandible and their actions.

Name	Action
Occipitalis	Moves the scalp backward and works with frontalis by fixating its tendinous aponeurosis.
Frontalis	Raises the eyebrows and wrinkles the forehead.
Auricularis anterior	Draws the ear slightly forward and upward.
Auricularis superior	Slightly raises the ear.
Auricularis posterior	Moves the ear feebly backward.

Table A.3: The muscles of the scalp and their actions.

Name	Action
Orbicularis oculi	Closes the eyelid and tightens the skin of the forehead.
Levator palpebrae	Raises the upper lid.
Corrugator supercilii	Pulls the eyebrow medially.
Tensor tarsi	Draws the eyelids inward and compresses the lachrymal sac.
Superior rectus	Turns the globe of the eye upward.
Inferior rectus	Turns the globe of the eye downward.
Internal rectus	Turns the globe of the eye inward.
External rectus	Turns the globe of the eye outward.
Superior oblique	Aids in upward eye movement by correcting rotation.
Inferior oblique	Aids in downward eye movement by correcting rotation.

Table A.4: The muscles of the eyes and their actions.

The palpebral (eyelid) and orbital (cavity containing eyeball) regions contain muscles that aid in the use of the eyes. The eyelids protect and clean the eyes while the orbits protect and contain the eyes along with muscles to rotate the eyes. Table A.4 lists the muscles of the eyes and describes their actions. The eyelids also play a part in displaying emotion.

The muscles of the nasal region aid the nose in the intake of air and for input to the sense of smell. Movements of the nose also play a part in displaying emotion. Table A.5 describes the muscles of the nose and their actions.

The lip coverings are tightly bound to the connective tissues covering the *orbicularis oris*, which is the primary muscle of the lips, causing the lips to closely follow the movement of the muscles. The fibers of the orbicularis oris are made up of fibers from the other muscles inserted into the lips as well as fibers proper to the lips themselves. The muscles of the mouth are described in Table A.6. This area is often referred to as the maxillary (or jaw) region.

Name	Action
Compressor naris	Constricts nostrils.
Dilator naris	Dilates the nostril.
Levator labii superioris alaeque nasi	Elevates the upper lip and the wing (ala) of the nostril, also aids in forming the nasolabial groove.
Procerus	Draws the inner angle of the eyebrows downward and produces wrinkles over the bridge of the nose.
Depressor alae nasi	Draws the ala of the nose downward.

Table A.5: The muscles of the nose and their actions.

Name	Action
Buccinator	Compresses the cheek against the teeth and retracts the corner of the mouth.
Depressor anguli oris	Draws the corner of the mouth downward and medialward.
Depressor labii inferious	Depresses the lower lip.
Incisive inferior	Pulls the lower lip in towards the teeth.
Incisive superior	Pulls the upper lip in towards the teeth.
Levator anguli oris	Moves the corner of the mouth up and medially.
Levator labii superioris	Raises the upper lip and carries it a little forward.
Levator Labii Superioris Alaeque Nasi	Elevates the upper lip and the wing (ala) of the nostril while aiding in forming the nasolabial groove.
Mentalis	Raises and protrudes the lower lip.
Orbicularis oris	Closes the lips, compresses the lips against the teeth, and protrudes the lips.
Platysma	Pulls the corner of the mouth down and back.
Risorius	Pulls the corner of the mouth back.
Zygomaticus major	Draws the corner of the mouth laterally and upward.
Zygomaticus minor	Draws the outer part of the upper lip upward, laterally and outward.

Table A.6: The muscles of the mouth and their actions.

Extrinsic Muscles

Name	Action
Styloglossus	Retracts and elevates the tongue.
Hyoglossus	Depresses and retracts the tongue.
Genioglossus	Protrudes the tongue, raises and protracts the hyoid bone.
Palatoglossus	Elevates the tongue and narrows the fauces.

Intrinsic Muscles

Name	Action
Superior longitudinal	Shortens and curls the tip of the tongue upward.
Inferior longitudinal	Shortens and curls the tip of tongue downward.
Transverse	Elongates, narrows and raises the sides of the tongue.
Vertical	Flattens and broadens the tongue.

Table A.7: The muscles of the tongue and their actions.

The tongue is a muscular organ of the special sense of taste situated in the floor of the mouth. The base or root of the tongue is directed backward and connected via different tissues to the hyoid, the epiglottis, the soft palate and the pharynx. The tip of the tongue is directed forward against the lower incisor teeth. The tip of the tongue, its sides, dorsum and part of the under surface are free. There are two types of muscles, discussed in table A.7, associated with the tongue: extrinsic and intrinsic. The extrinsic muscles originate outside of the tongue and position the tongue within the oral cavity. The intrinsic muscles make up the tongue itself and shape the tongue surface. The musculature of the tongue creates a highly variable surface that can shape the vocal tract to aid in the production of sounds. With the tongue connected directly to the hyoid, when the hyoid is moved the tongue also moves, so muscles of they hyoid should also be considered with the tongue.

The *palate* forms the roof of the mouth and has two parts, the *hard palate* in front and the *soft palate* behind. The hard palate is covered by the mucous membrane.

Name	Action
Tensor palatini	Tenses the soft palate
Levator palatini	Raises the soft palate
Uvulus	Shortens the uvula
Palatoglossus	Elevates the tongue and narrows the palatine arches
Palatopharyngeus	Constricts the palatopharyngeal folds.

Table A.8: The muscles of the soft palate and their actions.

Name	Action
Superior constrictor	Constricts the pharynx and aids in velopharyngeal closure.
Middle constrictor	Constricts the pharynx.
Inferior constrictor	Constricts the pharynx.
Stylopharyngeus	Raises and opens the pharynx.
Salpingopharyngeus	Opens the auditory tube and raises the oropharynx and nasopharynx.

Table A.9: The muscles of the pharynx and their actions.

The soft palate is a movable fold suspended from the posterior border of the hard palate that forms an incomplete septum (dividing wall) between the mouth and the pharynx. The lower border is free and in the middle is a conical-shaped projection called the *uvula*. The soft palate aids in sucking and swallowing and in producing speech. It plays a vital role in the plosives, /p/, /d/, /t/, etc., especially during loud speech where high pressure is required. The muscles of the soft palate are itemized in table A.8.

A.2.8 Articulation of Sound

The articulation of sound involves the soft palate (velum) and pharynx in velopharyngeal closure and movements of the tongue, mandible, and lips. All of the musculature of these parts are also involved in chewing and swallowing.

The pharynx is a muscular tube incomplete in the front where it is continuous with the openings into the nasal, oral, and laryngeal cavities. It extends from the base of the sphenoid bone above to the entrance into the esophagus at the level of the cricoid cartilage below and is wider above than below.

The soft palate is composed of muscle, connective, and gland tissues. It is attached to the posterior rim of the hard palate anteriorly, to the side walls of the pharynx and oral cavity laterally, and hangs freely into the pharynx behind the tongue posteriorly.

Velopharyngeal closure for speech involves the upward and backward movement of the soft palate against the posterior pharyngeal wall about the level of the anterior tubercle of the atlas. Medial movement of the lateral pharyngeal walls against the sides of the soft palate completes the velopharyngeal closure.

The larynx

The larynx has two principle functions, 1) voice production and 2) a sphincter to prevent the entrance of foreign material into the lungs and to permit the build up of intrathoracic pressure for such activities as coughing, vomiting, urination, and defecation. Intrathoracic pressure is also necessary to bring the full force of contraction to bear on the thoracic muscle of the arms and shoulders for strenuous activity such as lifting.

A.3 Modeling

The techniques used to model a human head in facial animation are similar to techniques in other areas of computer animation. A major difference in facial animation, is that the face is highly deformable and even the smallest of these deformations can convey information. For instance, the slight movement of the lower eyelid can be the difference between truth and deceit [Ekm85].

The model must describe the geometry of the face to be animated. That is, if the goal is to animate a talking head of Napoleon Bonaparte, then the model should look like Napoleon, both in its physical geometry, and in its texturing. Texture maps can create the correct colors, and can hide incorrect geometry. But, of course, the more accurate the geometry, the better the results. In addition to similar appearance, the head should act like Napoleon. That is, the model should deform in a manner that Napoleon's real head would deform.

In this section, approaches to creating a facial model are described. First, the different model types will be discussed. Then, geometry acquisition and the use and acquisition of textures are described. Some of the more well known facial models will be presented and finally the modeling of specific parts of the face separately is detailed.

A.3.1 Geometric Representations

Facial animation models vary widely from simple geometry to anatomy-based, with the complexity generally based on the intended use. When deciding on the construction of the model, important factors are: geometric data acquisition, animation

data acquisition, animation methods, rendering quality desired, etc. This section describes the different geometric representations used for facial models. It is possible for more than one type of representation to be used by a model. It is common for the geometry of parts of the model, such as the tongue, eyes, or lips, to be represented differently than the rest of the face.

Polygonal

In facial animation, as in most computer graphics, polygons are by far the most common choice of representing the geometry of the model because they are easy to create, use, deform, and render in hardware. Acquisition of the geometry usually involves sampling the surface in a regular grid, which is trivially represented as a polygonal mesh. The surface can also be sampled randomly, which is more difficult to produce a polygonal mesh but robust methods exist. Polygons are easy and quick to render, easy to modify with existing editors and easy to represent. However, polygonal meshes suffer visually, particularly at the silhouette.

Spline

Spline representations create smooth surfaces for visually appealing images. They are often used in facial modeling and animation in various forms such as B-splines [DMS98, HFG94, NHS88, NHRD90, SVG95, Wai89], triangular B-splines [EG98], Bezier splines [Ree90, Tao97], Coons patches [Wil90b] and hierarchical B-splines [PB95, WF94]. Splines have local control over the surface allowing local deformations with changes in a control point. However, it is a complex problem to shape the local surface in an exact manner. Other drawbacks of splines are that they are difficult to create, edit, and interpolate a set of points. Spline surfaces are best kept

at genus 0, so the holes for the mouth, nostrils, eyes and ears present problems. To model the holes, the surface can be pushed inward to give the perception of a hole at the cost of extra complexity or alternatively, trim curves can be used to cut the surface.

Another serious drawback of spline surfaces is that local detail requires additional global complexity. This problem can be overcome with hierarchical B-Splines [FB88], which allow complexity to be added locally without adding global complexity. Hierarchical splines have been used by Wang and Forsey [WF94] and Provine and Burton [PB95] for facial animation.

Subdivision

Subdivision surfaces are a type of smooth surface. They have the advantage of being able to create local complexity without global complexity, and make it easy for a modeler to sculpt a new model. However, they are difficult to interpolate to specific data. Subdivision surfaces have been used successfully by Elson [Els90a] [Els90b] and also by Pixar [DKT98] for the animated short *Geri's Game*.

Implicit

A surface can be defined as all the points that have the same value of an implicit formula. An example function is the distance from a point c, which defines a sphere for all points of a distance r from c. Implicit primitives can be added together in a well defined manner to create more complex objects. Implicit surfaces are easy to evaluate and detecting collisions is easy. However, rendering is extremely slow and modeling objects with sharp edges, surface detail, and fine deformations is difficult. These drawbacks keep implicit objects from being used widely in facial animation.

However, for situations where photorealism is not required, such as cartoon characters and caricatures, implicit functions work well.

Muraki [Mur91] presents a method to fit a blobby model to range data by minimizing an energy function that measures the difference between the isosurface and the range data. Examples are given using range data of human faces. By splitting primitives, the fit is refined until a global tolerance is reached.

Implicit surfaces can be successfully used for parts of the face and for certain functions such as collision detection and avoidance. Pelachaud [PvOS94] describes a method of modeling and animating the tongue using soft objects, taking into account volume preservation and penetration issues. Guiard-Marigny [GMTA+96] use point primitives for fast collision detection and use contact surfaces for modeling lips. We use an implicit function for collision avoidance between the skin and skull described in Section 2.2.1.

A.3.2 Instance Definition

The geometric representation has a space of possible face shapes. This space generally is very large. For instance, a polygonal model with N vertices in $\Re^3$ has a $3N$-dimensional shape-space. Often, a method to define an instance of the shape-space which is easier than defining all of the $3N$ degrees of freedom is used. This section describes such instance definitions.

Parameterized

The concept of parameterizing is to define a k-dimensional space with a set of k parameters where k is usually small. For example, a facial mesh with n vertices in $\Re^3$has $3n$ degrees of freedom, which is more than we wish to define. Defining a

parameterization where k is much smaller than $3n$ allows the shape of the mesh to be specified by much fewer numbers. A facial model can have a parameterization that defines the shape of the model and one that defines the motion of the model. The difference is very subtle since they both define shape in some sense. Parke [Par74] calls them conformation parameters for the shape and expression parameters for the deformations. Generally, the shape parameters differ between individuals and remain constant for that character over time. Deformation parameters change over time specifying some behavior, such as a smile, and they may be the same for different characters.

Parameterizing the deformations is a very common animation method and will be discussed in Section A.4. Parameterizing the shape of the head owes its popularity to Parke [Par74] who did the seminal work in computer 3D facial animation. Parke created a set of parameters empirically from traditional hand drawn animation methods to define both shape and deformations. The Parke model will be described more fully in Section A.3.5. Parameterization of shape is done in combination with a model type such as polygonal or B-spline. The parameterization defines a mapping between the reduced set of parameters and the underlying surface definition.

Using a parameterization to define shapes has numerous benefits including compression, ease in defining a character, intuitive definition of a character, straight forward morphing between characters, ease in making small changes in shape features, etc. The major drawback of defining the shape of a characters head with parameters is that the number of possible face shapes is limited.

Another benefit of using parameters for defining the shape is that a space of feasible face shapes can be defined allowing a computer to quickly generate numerous

plausible shapes. DeCarlo et al. [DMS98] use such a concept to modify a prototype shape to generate new models (different facial shapes.) The new models are created with variational techniques and are constrained by facial anthropometry so the resultant geometry is only influenced by the prototype. The underlying model is a B-spline surface, with each face defined by anthropometric parameters. This technique is used to create numerous plausible facial shapes, not to generate a specific face.

Point-Mass Systems

In physics-based models the model is treated like a point-mass system and motion is calculated using Newtonian physics or $F = ma$. Often, spring forces are used defining the skin surface as a mesh of points connected with springs [KGC+96, KGB98, KMCT96, Lar86d, LTW93, Lee93, LTW95, PVY93, SMP81, Pie89, RCI91, TW90, VY92, Wai89, Wat87, Wat89, WT91, WMTT95]. The representation can be given volume by connecting the surface to underlying soft tissue and bone. The surface is then moved by applying forces resulting from muscle contraction. The muscles can be defined as spring forces as well or as constraints on node positions. The resulting differential equations are solved using some form of integration, such as fourth-order Runge-Kutta [PTVF92].

Platt and Badler [SMP81] describe work on a system designed to recognize and generate American Sign Language. Point masses are connected via springs while muscles apply forces to the points to achieve deformations. They use a very simple physics model due to the lack of computer processing power. This system is described in more detail in Section A.3.5.

Waters [Wat87] describes a system to simulate muscles, particularly muscles of the face, by giving each muscle a zone of influence. Waters and Terzopoulos [Wat89,

TW90] create a better model of the skin for more realistic deformations. Lee et al. [LTW93, Lee93, LTW95] make further modifications to the model. The Waters model is described in detail in Section A.3.5.

Finite Elements

The finite element method [HO79, OH77, OH80a, OH80b, Seg84] can be used for accurate simulation of the propagation of forces through an object. However, finite element methods are complex to use and expensive to calculate. For facial animation, where very small deformations are important, the finite element mesh must be very fine causing both increased computation time and difficulty in defining the mesh. Although it has not been a popular method for facial animation, it has been used [Gue89a]. Finite element methods have more utility in applications where very high accuracy of the tissue movement is required such as surgery simulation [Den88, KGC+96, KGB98, RGTC98], simulation of the tongue [WT95, WTP97] simulation of the TMJ [DGZ+96], and simulation of skin closure [Lar86a].

Larrabee [Lar86d] develops a finite element model of skin deformation that ignores extreme stress and strain ranges and also ignores viscoelastic properties in initial wound closure. Skin is assumed isotropic with zero tension and modeled as a 2D system utilizing only a few elastic constants that can be estimated in vivo. The skin is idealized as an elastic membrane with nodes at regular intervals and springs are attached to an immobile subsurface representing the subcutaneous attachments. The stress and strain properties of the skin are related by the standard equations of classical elasticity, which are theoretically accurate only up to 10-15% strain.

Guenter [Gue89a] describes a system for simulating human facial expressions. Guenter creates a mesh of nodes and uses finite element analysis to calculate the displacements of the nodes based on muscles forces.

Research by Wilhelms-Tricarico [WT95, WTP97] on a physiologically based model of speech production led to the creation of a finite element model of the tongue with 22 elements and 8 muscles. This initial model does not completely simulate the tongue, but it does show the feasibility of the method.

Koch et al. [KGC+96, KGB98] use volume data (Visible Human Project[NLH99]) to create a facial model with a C^1 finite element model to describe the facial surface that is attached to the skull with springs. The skull is either obtained with computed tomography, along with tissue stiffness values, or a default skull, non-proportionally scaled to fit into the face, and tissue stiffness values are used. To support springs through various tissue types they use springs attached in series. Muscles [LTW95] are attached to the skull and surface and facial motion is described using FACS. The effect of the muscles on the finite elements is determined and then stored as displacement fields that are used in their emotion editor for real-time animation.

Roth et al. [RGTC98] use a finite element approach for volumetric modeling of soft tissue for accurate facial surgery simulation. They extend linear elasticity to include incompressibility and nonlinear material behavior.

DeVocht et al. [DGZ+96] create a finite element model of the temporo mandibular joint (TMJ) using data from the Visible Human Project. The model is used to study the biomechanics of the TMJ.

A.3.3 Geometry Acquisition

This section discusses issues and methods of geometry acquisition, which is how the base shape of the surface is initially defined. The base shape is the initial shape before deformations are applied to achieve expressions. The geometry is just the surface definition, it must be deformed over time to create animation, it may need to be deformed to look like a particular person or character, it needs animation controls added to it, and it needs to be rendered. The source of geometry may be the most important factor in an animation system so it will play an important role in other aspects of the model. On the other hand, the other aspects of the model may force the geometry to be acquired in a specific manner.

Hand in hand with the source for the geometry is how the model will be created. Another issue is how many different characters must be modeled and their importance. If there are many different characters, the time to create the model must be minimized so modifying the geometry of a prototype model to fit each new character may be the way to proceed. Medical applications typically require data for both the surface and the underlying tissues. This data must be specific to the patient. The most common methods of creating models are modifying existing models and digitizing geometry and the methods are discussed in the following sections.

Laser Scanners

A popular choice for acquiring geometry is to digitize the geometry of a particular person or sculpture of a character using a laser scanner. This allows quick creation of geometry, but all you have is geometry. The rest of the model, such as deformation control must still be added.

Laser scanners use a laser to calculate distance to the model surface and can create very accurate models for use in facial animation [deG89, Kle89, LTW93, Lee93, Mur91, NHS88, Ree90, WT91, Wil90a]. The scanners sample the surface at regular intervals and the surface can be reconstructed in many different ways, with polygonal representation the most popular. Another advantage of laser scanners is their ability to also capture color information concurrently resulting in a texture map that matches the geometry well. However, laser scanners are bulky, expensive, the subject must be available to scan, and scan time is measured in seconds necessitating the subject be still for several seconds. Laser scanners scan in a cylindrical or elliptical pattern that works well for the front and back of the head, but fails to capture the chin and the top of the head. They also have problems with surfaces that are not roughly perpendicular to the sensor such as the sides of the nose and the ears. Another problem for laser scanners is hair, which disperses the laser giving erroneous data for areas covered in hair such as the top of the head, eyebrows, beards and mustaches.

Modification of Existing Models

The steps in creating a model of a new character vary greatly depending on the choices made in the model type, how deformations are done, how the geometry is acquired, etc. In general, creating a new character involves acquiring the geometry and defining deformations to that geometry for motion such as expressions, speech, or emotion. The deformations can be based on muscle action, parameters, finite elements, or some other technique. For each new character, the deformations must somehow be mapped onto the new geometry. Generally, this is done by either a human or a human-assisted system. Fully automated methods are very complex to create. If numerous different characters are required, creating the model for each character

can be an expensive endeavor. For that reason, many researchers instead develop a prototype model with the animation controls built in and then adapt its geometry to match the geometry of the new character [AWS90, AS92, DMS98, EPMT98, LTW93, Lee93, Pat91, PW91, PALS97, PHL+98, RvBSvdL95, Tao97, MTMdAT89, WT91, XANK89].

Taking existing characters and making modifications or interpolating between multiple characters is a common approach. Magnenat-Thalmann et al. [MTMdAT89] describe methods to create new facial models from existing models with local transformations, interpolation between models, or composition of separate parts. Patel [Pat91] and Patel and Willis [PW91] allow interpolation between existing models to create new models. DiPaola [DiP89, DiP91] modified Parke's model to have more parameters and to allow for the creation of a larger number of realistic faces as well as stylistic or cartoon-like faces. DeCarlo et al. [DMS98] modify a prototype head constrained by facial anthropometry using variational techniques to generate new models. Escher et al. [EPMT98] describe how to fit a prototype facial model using Dirichlet free form deformations in MPEG-4.

Using vision to fit a generic model to a new person has also proved successful. Akimoto et al. [AWS90, AS92] adapt a facial model prototype by extracting facial features from images of a front and side view using image processing techniques. Reinders et al. [RvBSvdL95] use vision techniques to modify the CANDIDE model to images. Xu et al. [XANK89] describe a system that automatically obtains a 3D model by modifying a base model from two orthogonal images using a priori knowledge, best fit, and interpolation. Pighin et al. [PALS97, PHL+98] develop an interactive system to select correspondences in images to calculate a rough generic

model fit and an estimate of camera poses. Then a scattered data-fitting algorithm interpolates unconstrained (no correspondence) vertices to complete the fit. This new fit is then refined by interactively making new correspondences that do not change the pose and reinterpolating unconstrained vertices until a satisfactory fit is made.

A collection of surface points, such as those collected by a laser scanner can also be used to fit a generic model. Waters and Terzopoulos [WT91] report on a method to adapt their mesh to laser scanned data giving a physically-based model of a particular person with muscles interactively painted on the model. Lee et al. [LTW93, Lee93] extend this work by developing a predominately automatic method.

Human Modeler

One method of model creation is for a human modeler (artist) to shape a model by hand. This is useful when the model is of a non-human, a caricature or not a specific person. Although the artist has the most freedom in this type of system, it requires the most skill. It is very popular in the entertainment industry where such talent exists. There are several paradigms for such a system. Hofer [Hof93] describes a system for creating character facial animation by sculpting. The user interface allows the model to be modified using free form deformations. Patel [Pat91][PW91] creates a system that allows a user hierarchical control over the model and its animation. DiPaola [DiP89] [DiP91] creates new models by modifying parameters of Parke's model interactively. Kleiser [Kle89] sculpts in clay then uses a 3D digitizer to get the geometry.

From Images

Another option is to extract the information from video or photographs. This allows model creation of subjects that are no longer available and has application for video conferencing and compression. The common approach is to use vision techniques to create a correspondence between two images. Each image is a projection of the 3D world onto the 2D image plane and each square pixel corresponds to an infinite pyramid in $\Re^3$. If the same point on an object can be detected in more than one image, creating a correspondence, an intersection volume can be calculated. The more images used and the wider the angle between the images the more accurate the location (smaller the volume). The most common method is to identify feature points such as mouth corners, eye corners, nose tip, chin tip, etc. and use those to fit a generic head model to the input photographs.

This technique requires at least two images and they should be as far apart angularly as possible with the largest possible overlap of the face seen in the different views. Two orthogonal images maximizes the above relationship. Because of the shape of the face, a front and side view allow for all points on one quarter of the face to be seen in both images. By assuming symmetry about the midline, the entire front of the head can be extracted. For a full head, a top and back image are also necessary. Using two orthogonal images is a common method since it is easy to code and can be achieved with minimal hardware. Akimoto et al. [AWS90, AS92] adapt a generic facial model by extracting facial features from images of a front and side view using image processing techniques. They use pixel color clustering to find large regions (background, face, hair) and match a profile curve to the side view to get height information. This information is then used to localize a search for facial

features using edge (Sobel) detection. Xu et al. [XANK89] describe a system that automatically obtains a 3D model by modifying a base model from two orthogonal images using a priori knowledge, best fit, and interpolation.

As the cost of cameras and computing power decreases, systems using more cameras, video sequences, and high-resolution cameras have emerged. Guenter et al. [GGW+98] determine the geometry and texture from a bank of six cameras. The Michelangelo [SMC00] project uses numerous high-resolution cameras and vision techniques to capture high-resolution geometry at 30Hz.

Reinders et al. [RvBSvdL95] extract a face from a video sequence by adapting a generic 3D facial model to fit the video sequence. The algorithm uses robust automatic localization of semantic features using a knowledge-based selection mechanism. It adapts a generic face model to the extracted facial contours by globally transforming the generic model to account for pose and scale. The mismatch is then refined by applying local adaptation based on graph matching. Their method makes heavy use of heuristics and a priori knowledge of the makeup of a human head.

Ezzat and Poggio [EP96] use an image-based modeling technique to create photo-realistic models of human faces from a set of images of a face without a 3D model. A learning network is trained to associate the example images with pose and expression parameters. The network can synthesize a novel, intermediate view using a morphing approach, which can model both rigid and non-rigid motion. An analysis-by-synthesis algorithm extracts a set of high-level parameters from an image sequence using embedded image-based models. The model parameters are then perturbed in a local and independent manner for each image to minimize an error metric based on correspondence.

Pighin et al. [PALS97, PHL+98] develop an interactive system to select correspondences in images in order to calculate a rough generic model fit and an estimate of camera poses. Then a scattered data-fitting algorithm interpolates unconstrained (no correspondence) vertices to complete the fit. This new fit is then refined by interactively making new correspondences that do not change the pose and reinterpolating unconstrained vertices until a satisfactory fit is made. Texture maps are created using a weighted sum of the input images based on visibility, positional certainty, and closeness to the new viewing direction. Animation is achieved with liner interpolation of keyframes.

The technique is also common in model-based communication [AHS89, MAH90] with generally only a rough shape extracted. If parameters for a human head such as shape and pose can be extracted from a frame of a video sequence, only those parameters need to be sent instead of the entire image, resulting in high compressions ratios.

Automatic

Sometimes one simply wants to generate realistic geometry of a human head for crowd scenes, nonimportant background characters, or as a starting point for creating a final character. DeCarlo et al. [DMS98] use facial anthropometric statistics and proportions to generate constraints on a facial surface. Using variational techniques, they create realistic facial geometry that fits the constraints from a deformed prototype. Their method allows quick creation of realistic facial geometries. However, the resultant model is influenced by the prototype.

Volumetric Data

Another method is to create a model from computed tomography [Wat92] MRI [Tao97], Ultrasound [SL96] or other medical devices. These devices allow the acquisition of the internal structure as well as the surface and are invaluable in medical applications such as plastic or facial surgery. However, they are invasive techniques that cannot be used in the general case. The Visible Human Project [NLH99] is a common source of volumetric data [KGC+96, KGB98] to test such techniques.

Waters and Terzopoulos [Wat92] describe a system that uses computed tomography data to derive a facial model obtaining a polygonal mesh of the exosurface of the skull and skin by utilizing marching cubes. The model is a spring lattice that uses the Lagrange equations of motion and Euler integration. A single spring, k, connects node i to node j with a natural length of l_k, a spring stiffness of c_k, a node separation (spring length) of $r_k = x_j - x_i$, and a spring deformation of $e_k = ||r_k|| - l_k$. The force, s_k, the spring exerts on node i is calculated as:

$$s_k = \frac{c_k e_k}{||r_k||} r_k$$

The discrete Lagrange equation of motion is the system of coupled, second-order ordinary differential equations:

$$f_i = m_i \frac{d^2 x_i}{dt^2} + \gamma_i \frac{dx_i}{d_t} + g_i \qquad\qquad i = 1, \ldots, N$$

where,

$$g_i(t) = \sum_{j \in N_i} s_k$$

A muscle layer is constructed equidistant between the skull and skin, with cross struts to avoid shearing, and with a spring constant lower than that of the skin. Contraction of linear muscles [Wat87] attached throughout the model results in animation.

Koch et al. [KGC+96, KGB98] use volume data to determine tissue stiffness values and depth information for their finite element model. Their system is a simulation of facial surgery, which requires an accurate method to obtain the internal structure of the intended patient. Using medical diagnosis equipment gives them the pre-operative information they require so that the surgery can be practiced and the results simulated.

Konno et al. [KMCT96] describe a facial surgery simulation system designed to aid in surgery in cases of facial paralysis. They acquire the 3D geometry from CT scans of the patient. And Roth et al. [RGTC98] use a finite element approach for volumetric modeling of soft tissue for accurate facial surgery simulation.

A.3.4 Rendering

Animation achieves apparent motion by making small changes in successive still images. The quality of the individual stills affects the overall quality of the animation. This section discusses techniques to render human faces. We first discuss rendering techniques for realistic skin. We then discuss the use of texture maps for high quality renderings. And finally, we discuss methods of producing wrinkles.

Rendering Skin

Hanrahan and Krueger [HK93] present a model for the subsurface scattering of light in layered surfaces using one-dimensional transport theory. This method replaces Lambertian diffuse reflection with a more accurate model of diffuse reflection that considers how the light interacts with different layers. This method works well for human skin, which is a multi-layered tissue.

Wu et al. [WKMT96] develop a technique to simulate static and dynamic wrinkles on a skin surface and also to render the skin. They model the skin with both a micro and macro structure. At the micro level, they create a tile texture pattern using a planar Delaunay triangulation with a hierarchical structure and render the micro furrows using bump mapping. At the macro level, they model wrinkles, which will be discussed in more detail later.

A method such as that of Hanrahan and Krueger is computationally expensive, complex, and difficult to set all the parameters to look like a particular person. The method of Wu et al. can also be expensive. Instead, the common technique of applying texture maps can be used. Texture maps allow the model to be rendered with increased visual realism and can hide many of the inadequacies in the representation of the facial surface. Another important aspect of rendering a human face is wrinkles, wrinkles are not only needed for visual realism, but their appearance also conveys expression and emotion. Textures and wrinkles will be discussed further in the following sections.

Textures

Laser scanners are capable of collecting intensity information [NHS88, NHRD90, WT91] concurrently with distance to an object. This makes them a good choice for collecting the geometry since texture information can also be obtained concurrently. Laser scanners produce high-resolution surfaces with high-resolution textures. Parts of face that the laser scanner fails to collect geometry for, such as under the chin and parts of the ear, will also be missing texture information. However, intensity can be collected for hair even if the depth information is missing.

The same vision techniques used to create the geometry can also be used to collect texture information [AS92, Tao97, XANK89]. However, each image will have different specular lighting since the angle to the light source is different. The ambient lighting may differ so the texture information must somehow be combined. Assuming symmetry, two views is enough to get a texture map with a technique such as that by Akimoto et al. [AWS90, AS92]. They use front and side view images to create a texture map using pixel blending where the textures overlap. More views can be used for a higher resolution texture, with either multiple cameras or a video sequence. Lighting effects, model registration, and deformation of the subject must be considered.

Ishibashi and Kishino [IK91] present an efficient color/texture analysis and synthesis method for model based coding of human images. Using the HSV (hue, saturation, value) color model, the image is broken up into uniform color regions using first and second order statistics determining the region's color independently of shape and shading. The texture and color are then synthesized onto a 3D model of the human body. The image is divided into high chromatic, low chromatic, and achromatic regions then subdivided using segmentation methods. The hue is corrected in the low chromatic regions then joined with the high chromatic regions and subdivided by thresholding the hue. The achromatic regions are subdivided based on brightness. Color is represented by hue and saturation while texture is represented by distributions of value intensity. The intensity of the value component comes from texture and shading. The shading is removed by correcting the value component of pixel i in region r (V_{ri}). $V'_{ri} = V_r + (V_{ri} - V_{riw})$, where V_r is the average intensity in region r and V_{riw} is the average intensity of $w \times w$ pixels, including pixel i.

Pighin et al. [PALS97, PHL+98] use input images to interactively fit a generic model and estimate camera poses. Texture maps are created using a weighted sum of the input images based on visibility, positional certainty, and closeness to new viewing direction. Textures are extracted from the images by combing the input images into one texture image. A cylindrical projection of the face establishes a mapping of the face to a 2D texture giving a texture coordinate for each 3D point. For each point $p = (x, y, z)$, they calculate the image coordinate (x_j, y_j) in image j of p, and extract the color $I_j(x_j, y_j)$. The texture color at (u, v) is the sum of all $w_j(u, v)I_j(x_j, y_j)$ where w_j is a weight of the image at that location based on visibility, positional certainty (surface normal at p dotted with direction to the camera), and similarity between camera direction and direction of the new view.

Wrinkles

Guenter [Gue89a, Gue89b] uses wrinkles as surface detail by mapping a 2D spline curve onto the surface of the model. The curve can then be used to create a discontinuity in the surface normal during shading calculations.

Pelachaud et al. [PVY93] present a system that produces automatic facial expressions with wrinkles from spoken input. They focus on the integration of expressive wrinkles and the generation of synchronized animated speech. The facial model is capable of facial muscle deformations and bulges. They modify an early system [PBS91] by replacing the Platt model with the model of Viaud and Yahia [VY92], which is capable of wrinkles. The wrinkles are created by post processing the model based on the muscles and creating tangency discontinuities between two isolines of the spline surface, which are perpendicular to the action line of the muscular force. Age is also used in determining wrinkles.

Wu et al. [WMTT95] describe the generation of wrinkles for expression and aging. Expressive wrinkles appear during expression and become visible over time, due to changes in the elastin and collagen fibers. As the elastin fibers stretch over time they lose elasticity and the collagen fibers collect into sheaves making the skin less homogeneous. To model these effects they define the skin as a 2D lattice over 3D points connected with springs. Muscles are interactively attached to the mesh and are implemented as spring forces between insertion points. Expressive wrinkles are formed during muscle contraction and micro wrinkles (aging) are created by changing the rest constants of the springs over time.

Wu et al. [WKMT96] develop methods to simulate static and dynamic wrinkles on a skin surface and also to render the skin. They model both a micro and macro structure of the skin. The micro lines are defined by a Delaunay triangulation of the 2D texture image plane. The triangle edges represent furrows and the ridge is created by ramping the height based on barycentric coordinates with the maximum height at the center of the triangle. This can be done hierarchically with triangles being subdivided into smaller ridges and furrows. The micro furrows are rendered using bump mapping. Macro lines are hand painted onto the model as a B-spline curve that is projected onto the 2D texture and constrained to lie on the micro structure. The space between macro lines is parameterized and the parameters are used along with a shape function to create bulges between the macro lines. Varying the shape function creates different types of bulges. For the dynamic behavior of the skin, they calculate the strain of a triangle using the vertical and horizontal line through the center along with one of the edges. Hooke's law is used for computing the stress components on the triangle, which in turn are used to find the in-plane forces along

the edges. Expressive wrinkles are defined as curves transverse to the muscle fibers with the height of the bulge mapped from the calculated strain.

A.3.5 Famous Models

Numerous facial models have been created over the nearly three decades of research on facial animation. This subsection list models that have had the most impact in the area. The impact is judged on the contribution made by the model, how often the model is referenced and how often it is incorporated into other systems.

The Parke Model

While a graduate student at the University of Utah, Frederic Parke was the first person to develop a 3D computer graphics model of a human head for the purpose of facial animation. 3D graphics was a new frontier of research and the University of Utah was one of a handful of institutions where it was being conducted. Parke's parametric model [Par74] is probably the most used model in facial animation. Parke's model is used by so many independent researchers because it was the first facial model, is easy to use and modify, and is freely available. Parke released the code for his model to several researchers that shared their modifications with others and in 1996 Parke and Waters [PW96] both released software for their respective models to support their book.

In his masters thesis, Parke [Par72] creates facial animation by keyframing a polygonal model using the brand new technique of Gouraud [Gou71] shading, and the model is digitized by hand using two orthogonal views, both ground-breaking techniques. Polygons are chosen over patches because a shading algorithm for patches did not exist, and patches were hard to use and implement and were extremely slow.

The keyframing was tedious, involving collecting data for the face and expressions for each keyframe. To reduce the workload, Parke [Par74, Par75] develops a parametric facial model.

The model is capable of representing different individuals via conformation parameters and different expressions with expression parameters. Animation can be achieved by changing the parameters reducing the workload on the animator drastically. Interpolation between poses (expressions, phonemes) is the main concept. By defining two extremes, a value (parameter) can be used to interpolate between the two extremes. Defining appropriate parameters and extremes makes it possible to animate the face.

The parameters are divided into conformation and expression parameters, with symmetry between sides of the face assumed. There are many scaling parameters to create the correct shape including: X,Y,Z of face, forehead size, nose bridge and lower nose width. Eyelids are done by a combination of linear interpolation and spherical mapping. Different eyelid positions result from linearly interpolating a closed eyelid mesh and an open eyelid mesh. The interpolation is then mapped onto a sphere 1.1 times the size of the eyeball with the same center. There are 2 parameters for the eyebrows, 9 for the eyes, and 10 for the mouth region. The mouth region includes the lips, teeth, and a jaw that rotates but has no lateral movement. The parameters are chosen empirically and from traditional hand drawn animation methods. Using Madsen [Mad69], Parke determines that the important capabilities to achieve lip animation are:

1. Open lips for /a/,/e/,/i/

2. Closed lips for /p/,/b/,/m/

3. Oval mouth for /u/,/o/,/w/

4. Tuck lower lip under front teeth for /f/ and /v/.

5. Move between these lip positions as required.

6. The remaining sounds are formed by the tongue and do not require precise animation.

Parke built an animation by going through the following steps:

1. Set the parameters frame by frame resulting in too much motion.

2. Allowed only smoothed jaw movement which created a smoother animation, but still did not look natural nor match the speech well.

3. Added /f/ and /v/ tucks with no improvement.

4. Added variations to the mouth width with some improvement.

5. Modified the model to allow upper lip movement giving improved animation.

6. Finally, added eye, eyebrow and mouth expression motion resulting in more convincing animation.

Parke found that he needed just ten parameters to produce animated speech: Jaw rotation, upper lip position, mouth width, mouth expression, eyebrow arch, eyebrow separation, eyelid opening, pupil size and eye tracking.

Parke noticed that more conformation parameters were needed, such as eyebrow shape and nostril size, and that the model should not be symmetric. His initial model lacked neck pivots, hair, neck, ears, and a tongue. Parke [Par82] later added more parameters and refined the model.

Lewis and Parke [LP87] process recorded speech using linear prediction to determine the spoken phonemes which are converted into keyframes (visemes) of a modified Parke model that now has a full head.

Pearce et al. [PWWH86] modify the Parke model by adding a scripting interface within the Graphicsland system. The scripting interface gives them the ability to produce synchronized speech automatically by sending the phoneme information to both the speech synthesizer and the facial model. The scripting interface allows for a library of expressions, such as smile, blink, and phonemes to be built and accessed by the animator at a high level. Hill et al. [HPW88] describe the system from a speech production point of view by extending speech synthesis by rules to also create the facial model parameters. Wyvill and Hill [WH89] describe the system from a graphics point of view and gives parameter values for generating the phonemes.

In their work with the deaf and speech reading, Cohen and Massaro [CM93] extend the modifications of the Parke model made by Pearce et al. [PWWH86]. They add more parameters to support speech synchrony and include tongue geometry. They are interested in creating speech that can be understood by a deaf person, so they are more interested in the visual cues of speech than in the auditory ones.

DiPaola [DiP89, DiP91] adds new conformation parameters to the Parke model to create a larger space of possible face shapes. The new parameters allow more realistic as well as stylistic or cartoon-like faces. DiPaola also adds the ability to have skin creases and facial hair (eyebrows, mustache and beard) using hair globs.

The Waters Model

Keith Waters [Wat87] presents a parameterized muscle model for creating facial animation. The muscle model can be adapted to any physical model; a polygonal

mesh is used in his work. The parameterized muscles are given zones of influence and the nodes of the facial model are displaced within these zones using a cosine falloff function. Work using the muscle system by defining ten muscles based on FACS is presented.

Waters uses two types of muscle linear/parallel and sphincter. Each muscle creates spring like forces and has a zone of influence with a falloff function. Each muscle has a direction (towards a point of attachment) and a magnitude. A simple example is a linear muscle with a circular zone of influence and a cosine falloff function. In this case points close to the muscle attachment will move the most while points near the edge of the circle will barely move with the influence on the points in between following a cosine function. Cosine acceleration and deceleration is used as a first order approximation since the displacements are small and the rate of motion occurrence is very fast.

A linear muscle is defined as a muscle that is inserted into the bone, which does not move, and its point of attachment in the skin is where the maximum displacement occurs with the displacement towards the bony attachment. Figure A.6 describes how Waters defines a linear muscle. Here, we describe the latest incarnation of the formulas [PW96], which is a slight refinement over the earlier work [Wat87]. For a linear muscle with a zone of influence with an angle $< 180°$, the muscle pulls along the vector between v_1 and v_2, Ω is the half angle of the zone of influence and Rs and Rf define the start and finish of the falloff respectively. The zone of influence is the region $v_1 p_r p_s$ with the falloff zone the region between $p_n p_m p_r p_s$. Any mesh node $p = (x, y, z)$ located within the zone of influence $v_1 p_r p_s$ is displaced towards v_1 along

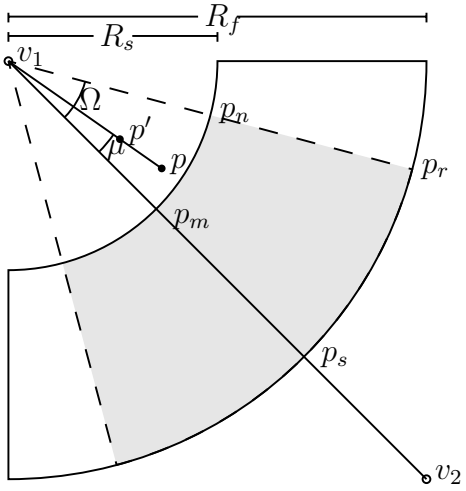


Figure A.6: Waters vector muscle model. The vector muscle that acts along $\overrightarrow{v_1v_2}$ influences the sector $v_1p_rp_s$ with R_s and R_f defining the start and finish of the falloff.

the vector $\overrightarrow{pv_1}$ to $p' = (x', y', z')$ where

$$p' = p + akr\frac{\overrightarrow{pv_1}}{\|\overrightarrow{pv_1}\|}$$

and $a = cos(\mu)$ is the angular displacement, μ is the angle between $\overrightarrow{v_1v_2}$ and $\overrightarrow{v_1p}$, k is the muscle spring constant and r is the radial displacement calculated as

$$r = \begin{cases} cos\left(\frac{1-D}{R_s}\right), & \text{if inside } v_1p_np_m; \\ cos\left(\frac{D-R_s}{R_f-R_s}\right), & \text{if inside } p_np_rp_sp_m; \end{cases}$$

where $D = \|v_1 - p\|$,

A sphincter muscle squeezes towards a single point of contraction and moves the skin towards the center similar to the tightening of a bag with a string. The new point p' is calculated as

$$p' = 1 - \frac{\sqrt{l_y^2p_x^2 + l_x^2p_y^2}}{l_xl_y}$$

where l_x is the major axis and l_y the minor axis of the ellipse.

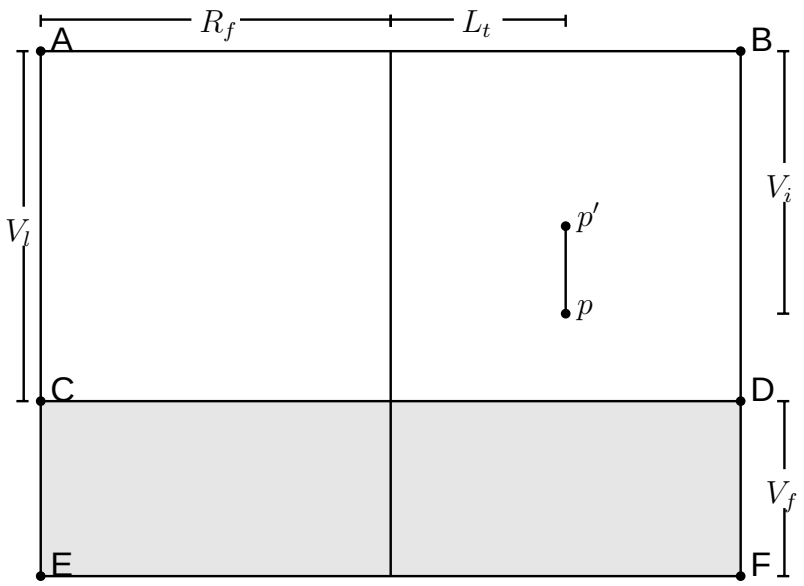


Figure A.7: Waters sheet muscle model.

Later, Waters [Wat89] defines a sheet muscle. The sheet muscle consists of strands of fibers that lie in flat bundles. The sheet muscle does not contract to a localized point nor emanate from a single source like the linear muscle, but instead it acts as a series of almost-parallel fibers spread over an area as shown in figure A.7. The new location p' of node p inside the zone of influence is $p' = p + d$ where

$$d = \begin{cases} cos(1 - \frac{L_t}{R_f}) & \text{for p inside sector ABCD} \\ cos(1 - \frac{L_t}{R_f}(\frac{V_i}{V_l} + V_f)) & \text{for p inside sector ABEF} \end{cases}$$

Waters [Wat89] describes work with Terzopoulos on a dynamic model of facial tissue in which the face is modeled as a tri-layer mesh of springs, one layer each for the skin, muscle, and bone. Here the muscles are attached to the middle layer and the deformation on the surface layer is propagated through the mesh of springs. A node, x_j, is affected by muscle vectors, m_i. The new location, x'_j, is calculated by

$$x'_j = x_j + \sum_{i=1}^{m} c_i b_{ij} m_i$$

where c_i are weights, m_i is the muscle rest length, $r_{ij} = m_i^t - x_j$ is the distance to the tail of the muscle, and b_{ij} is the muscle blend. b_{ij} is calculated with

$$b_{ij} = \begin{cases} cos\left(\frac{\|r_{ij}\|}{a_i}\frac{\pi}{2}\right); & \text{for } \|r_{ij} \leq a_i\| \\ 0 & \text{otherwise} \end{cases}$$

where a_i is the radius of influence of the cosine blend profile.

The displacement is propagated throughout the mesh using second-order Lagrangian dynamics [TPBF87] and second-order Runge-Kutta integration of

$$m_i \frac{d^2 x_i}{dt^2} + \gamma \frac{dx_i}{dt} + f_i^e = 0$$

where x_i is the position of the nodes, m_i is the mass, γ is the coefficient of velocity-dependent damping and f_i^e are the net forces at node i. f_i^e is calculated by

$$f_i^e = f_i^s + f_i^g + f_i^m$$

where f_i^m is the force due to muscles, f_i^g is the effect of gravity, and f_i^s is the force due to the springs connected to node i. f_i^s is calculated by

$$f_i^s = \sum_{k \in C_i} (c_k(\|x_i - x_j\| - l_k))$$

(C_i). where l_k is the rest length of the spring, c_k is the spring constant, and C_i is the set of springs connected to node i.

Terzopoulos and Waters [TW90] add a volume preservation force and present a method to estimate muscle parameters from video by applying makeup and using snakes to track the makeup. The volume preservation force is added to the total force on node i, however, they do not give details on how to calculate this force. In this work, they modify slightly the way spring forces are calculated as well as use biphasic springs.

A single spring k connects node i to node j with a natural length of l_k and a spring stiffness of c_k. $r_k = x_j - x_i$ is the node separation (spring length) and the spring deformation is $e_k = ||r_k|| - l_k$. The force, s_k, the spring exerts on node i is calculated as

$$s_k = \frac{c_k e_k}{||r_k||} r_k$$

To more accurately model tissue, which is readily extensible at low strains, but exerts rapidly increasing restoring stresses after reaching a strain e^c, biphasic springs are used. The spring constant c_k for a biphasic spring is

$$c_k = \begin{cases} \alpha_k & \text{if } e_k \leq e_k^c \\ \beta_k & \text{if } e_k > e_k^c \end{cases}$$

where the small-strain stiffness α_k is smaller than the large-strain stiffness β_k.

Waters and Terzopoulos [WT91] report on a method to adapt their mesh to laser scanned data to get a physically-based model of a particular person. Muscles are interactively painted onto the model. Waters [Wat92] uses marching cubes on computed tomography data to obtain a polygonal mesh of the exosurface of the skull and skin. The muscle layer is constructed equidistant between the skull and skin with cross struts to avoid shearing, a spring constant lower than that of the skin and deformations calculated as above.

Waters and Levergood [WL93] describe DECface, which is a system that generates lip-synchronized facial animation from text input. Using DECTalk to generate phonemes and a waveform, they associate a keyframe with each phoneme and interpolate between them. To simulate acceleration and deceleration they use $s' = s * (1 - cos(\pi(s_0 - s)))/2$ for the interpolation. However, in fluent speech the target position is rarely achieved so they assign a mass to the nodes and then use

Hookean spring calculations, with simple Euler integration, to approximate the elastic nature of the tissue. Expressions are created with Waters' muscle model [Wat87].

Lee et al. [LTW93, Lee93] extend [WT91] by developing a heuristic method that is predominately automatic for creating a physics-based model of a particular person's head from laser-scan data by fitting a generic model. The physics of the model are also modified. The spring stiffness constant, c_s, for biphasic dermal-fatty springs is

$$c_s = \begin{cases} \frac{I_s K_2}{l_s^r} & \text{for } l_s > l_s^r(1 + B_p^s) \\ \frac{I_s K_1}{l_s^r} & \text{for } l_s^r(1 + B_p^s) > l_s > l_s^r \\ \frac{I_s K_1 \Phi(l_s, l_s^r)}{l_s^r} & \text{for } l_s^r > l_s \end{cases}$$

where K_1 and K_2 are the spring constants for a biphasic spring, l_s is the length of spring s, l_s^r is the rest length, B_p^s is the springs biphasic point given in terms of $\frac{l_s}{l_s^r}$, I_s is the multiplicative factor (-1 when the spring is inverted and 1 otherwise.)

$$\Phi(l_s, l_s^r) = tan\left(\frac{\sigma\pi\left(1 - \frac{l_s}{l_s^r}\right)}{2}\right)$$

is the spring compression penalty where $0 < \sigma < 1$ is a scaling factor. For other springs the stiffness constant is found using

$$c_s = \begin{cases} \frac{I_s K}{l_s^r} & \text{for } l_s > l_s^r \\ \frac{I_s K \Phi(l_s, l_s^r)}{l_s^r} & \text{for } l_s^r > l_s \end{cases}$$

where K is the spring constant for a spring.

Lee also develops a new linear muscle model where the force acting on node i by muscle j is

$$f_i^j = \Theta_1\left(\epsilon_{j,i}\right)\Theta_2(\omega_{j,i})m_j$$

where $\epsilon_{j,i}$ is the width ratio (perpendicular distance of the node from the normal of the muscle), $\omega_{j,i}$ is the length ratio (distance between the end of the muscle and the closest point on the muscle vector to the node), and m_j is the unit muscle vector. Θ_1

and Θ_2 are scaling function calculated as

$$\Theta_1(\epsilon_{j,i}) = sin\left(\frac{\pi\epsilon_{j,i}^k}{2}\right) + \frac{1}{8}\left(sin\left(2\pi\epsilon_{j,i}^k - \frac{\pi}{2}\right) + 1\right)$$

$$\Theta_2\left(\omega_{j,i}\right) = sin\left(\frac{1}{2}\left(cos\left(\pi\left(\frac{\omega_{j,i}}{w_j}\right)^{1.8}\right) + 1\right)\right)$$

where w_j is the width of muscle j.

The discrete Lagrange Equation of motion for node i is now calculated as

$$m_i\frac{d^2x_i}{dt^2} + \gamma_i\frac{dx_i}{d_t} + g_i + q_i + s_i + h_i = f_i;$$

where g_i is the spring force, q_i is the volume preserving force, s_i is the skull penetration force, h_i is a force that restores node i to its rest location, and f_i is the muscle force.

The Platt Model

Platt and Badler [SMP81] describe work on a system designed to recognize and generate American Sign Language. This was the first work to apply a physics-based force model to facial animation and they introduced the Facial Action Coding System (FACS) [EF78a] to the facial animation community.

FACS is used to describe the motion and tension nets to describe the geometry of the face with a net made up of:

- *Points*, which are 3D locations such as skin, muscle or bone.
- *Arcs*, which connect two points together with a spring.
- *Muscle fibers*, which connect bone or muscle points with one or more skin points.
- *Muscles*, which are made up of many muscle fibers.

The skin surface is a polygonal mesh with bone and muscle points under the surface without any other interior structure. Muscle forces are determined by hand based on FACS Action Units and when the forces are applied to the muscles, the skin deforms. When force f is applied to point p, the change in location is $dl = \frac{df}{k}$. To animate for N time steps a force of $\frac{f}{N}$ is applied at each step.

Platt [Pla85] abandons the springs for a displacement method. The work is based on a hierarchical definition of the face with the face divided into seven areas: forehead, left/right eye, left/right cheek, nose and mouth. The face is divided into landmarks and transient regions. Landmarks are permanent locations on the face and can be moved or masked with two types of landmarks: structural (eye openings) and surface (eyebrows). Transient regions are locales of change such as the area between the brows. Instead of muscle forces, the FACS AUs give rise to MOVE commands, which do the actual deformations. They also modified FACS to create a more generative (as opposed to analytical) system.

The Platt model has not been as well traveled as the Parke model or the Waters model but it has been used by other researchers. Pelachaud [Pel91] and Pelachaud et al. [PBS91] use the Platt model to create animated speech with expressions from a speech source. This research also deals with coarticulation issues in speech. Essa [Ess94] also uses the Platt model in his research on creating and analyzing facial expressions.

CANDIDE

Rydfalk [Ryd87] describes a system called CANDIDE designed at the University of Linköping in Sweden to display a human face quickly. CANDIDE was designed when graphics hardware was slow and the number of polygons was very important.

When designing the system they had four constraints: use of triangles, less than 100 elements, static realism, and dynamic realism. He defines static realism as when the motionless face looks good and dynamic realism as when the animated motion looks good. FACS AUs are used to define the animation. The final geometry has less than 100 triangles and 75 vertices resulting in a very rough model that lacks adequate complexity in the cheeks and lips. With so few triangles, it is very difficult to make the model geometrically look like a particular person.

This technical report describes an implementation with no novel research and because of the lack of complexity, it has little utility in facial animation. However, it is popular among vision researchers [Red91, RvBSvdL95], mostly those in Europe, for applications such as tracking, model based communication, and compression where a low polygon count model is sufficient and desired.

A.3.6 Other Facial Parts

Facial animation geometry can be complex or simple. Many of the early systems and even some of the later ones, only use a facial mask, which is only the front part of the face. The biggest reason for this is that the rest of the head is covered with hair and hair is very complex to model. If it is modeled as a surface, it must be very coarse to keep the model complexity down. The tongue, eyeballs, teeth, and skull are other parts of the face that are often ignored. This section presents methods that deal with just parts of the face, generally to add in some of this missing geometry that is not part of the skin of the face. Specifically modeling hair, the tongue, the lips, and the skull will be discussed.

Hair

Creating a character with a full head of hair that realistically moves has been one of the hard computer graphics problems. A head of hair has many tens of thousands of individual hairs, with each hair a complex surface that has anisotropic reflectance properties. Developing a system that can represent 100,000 hairs, represent particular styles, and render the hair with realism while the hairs move freely over time is a difficult problem. Calculating collisions between the hairs and the hair and the head, attractive static forces, and dynamic motion due to head movement is a daunting task.

Hair has two distinct tasks: modeling and rendering. Modeling involves styling and motion, which includes dynamics and collision detection. Solutions generally solve just a subset, usually sacrificing dynamics, to get reasonable speeds. One approach is to use a low-resolution polygonal surface model [DiP89, DiP91] with no dynamics to represent the hair. Another approach is to use procedural texturing methods to create visually realistic hair sacrificing dynamics [PH89, SPS96]. The individual strands of hair can also be modeled as a polygonal shape [AUK92, RCI91, WS89] or as lines [Mil88].

A simple method is to model the hair surface as a set of polygons. The easiest method is to use hair color for the polygons defining the hair. This is often used for eyebrows and mustaches [Par74]. An extension is to model many polygonal shapes to give the hair a less solid look. DiPaola [DiP89, DiP91] allows for specifying eyebrows, mustache and beard parametrically with the hair implemented using hair globs, which are small irregularly shaped polygonal surfaces.

Perlin and Hoffert [PH89] define hypertextures, which are volumetric solid textures. One of the applications they describe is a hair model. This method is computationally expensive, and there is no clear method to add dynamics to the hair. However, this method could be used for eyebrows, beards and mustaches.

Sourin et al. [SPS96] model hair using noise and implicit functions. A sphere that represents the location of the hair is used to locate the noise. For any point $p = (x, y, z)$ the point $p' = (x', y', z')$ is the location on sphere that the ray from p to the center of the sphere intersects the sphere. p' is used to calculate the hair value, $hair(x, y, z) = noise(x', y', z')$, thus all points along that ray have the same hair value. Offsetting the hair controls its thickness. Displacing p first as in

$$hair(x, y, z) = noise(projection(displacement(x, y, z))) - offset$$

allows the hair to be styled other than straight.

Miller [Mil88] uses pseudo-reflectance maps to create a fast anisotropic lighting model of cylindrical objects that is applied to line segments to create fur. The anisotropic shading model uses tangent vector information as well as surface normal vectors. To model fur, hairs are generated based on an equally spaced grid in parametric space, with jitter added to each hair. The root of the hair is placed on the surface and the end of the hair is a linear combination of the tangent vectors and surface normal.

Watanabe and Suenaga [WS89] use small trigonal prism elements to form curved hairs that are combined together in wisps and are rendered quickly using a z-buffer. Combining the individual hairs into bundles called wisps allows them to decrease the number of parameters needed to define a full set of hair. Individual hairs, as well as wisps, are controlled by a small number of parameters to facilitate quick modeling.

Hair can then be modeled with a small number of parameters: number of wisps, hairs per wisp, hair length, thickness and color. With visibility algorithms, two-thirds of the hair can be removed to reduce complexity.

Rosenblum et al. [RCI91] model hair as segments with point masses, springs and hinges with strand to head collision detection and shadowing effects. They describe a system for placing, rendering and simulating the motion of hair. The system employs z-buffer and shadow buffer techniques for quick rendering with self-shadowing effects. Hairs are modeled as segments of point masses connected with springs and hinges between segments. Newton's second law ($f = ma$) is solved to calculate the motions of the hair strands. Strand to head collision detection is also performed. Hair is interactively placed on the polygonal model by the animator, with support for jittering the follicle placements and mirroring.

Anjyo et al. [AUK92] present a method to style, render and animate hair. The head surface is approximated with an ellipsoid to reduce computation of pore locations and collision detection. Hair is styled by placing hair pores on the ellipsoid and growing the hair outward. Bending of the hair is calculated based on a simplified solution of a cantilever beam with collision detection between the hair strands and the ellipsoid. Hair can then be cut and adjusted to get the final desired shape. The cantilever beam is borrowed from the field of material strengths and is a straight beam with a one-sided fixed support. Considering the hairs to be rigid sticks, dynamic behavior is reduced to solving a simple one-dimensional differential equation of angular momentum for each hair. Hair to head collision detection is performed under a pseudo force field to give visually pleasing results but not highly accurate. They also present fast rendering of anisotropic reflection.

Tongue

In order to achieve more realism, several researchers recognized the need to represent the tongue in a facial animation system. These systems have used a range of methods to simulate the tongue including a simple geometric tongue with rigid motion, a human sculpted tongue in keyframe positions, finite elements, and a highly complex model using soft objects.

Some of the first uses of a tongue in computer facial animation were in the creation of animated shorts. Reeves [Ree90] describes the use of a teardrop shaped collection of 12 bi-cubic patches to model the tongue in *Tin Toy*, Pixar's academy award winning short. Although the tongue was modeled, it was usually left in the back of the mouth. Kleiser [Kle89] sculpts a face in phonemic positions then interpolates them, paying particular attention to the tongue, eyes, teeth, eyebrows and eyelashes.

In their work with the deaf and speech reading, Cohen and Massaro [CM93] modified the Parke [Par74] model to work with phoneme input and added a simple tongue model to increase intelligibility of the generated speech. The tongue model they use is quite simple and animates very stiffly with parameters for tongue length, angle, width, and thickness. A simple rigid tongue may be adequate if only the tip is visible. However, in general facial animation and excited conversation where the mouth is opened wide, the limitations are apparent.

In studying the tongue as part of the vocal tract, Stone [Sto91] proposes a 3D tongue model by dividing the tongue into 25 functional segments, five lengthwise segments and five crosswise. Stone and Lundberg [SL96] reconstruct 3D tongue surfaces during the production of English consonants and vowels using a developmental 3D ultrasound. Electropalatography (EPG) data was collected providing tongue-palate

contact patterns. Comparing the tongue shapes between the consonants and vowels revealed that only four classes of tongue shapes are needed to classify all the sounds measured. The classes are: front raising, complete groove, back raising, and two-point displacement. The first three classes contained both vowels and consonants while the last consisted only of consonants. The EPG patterns indicated three categories of tongue-palate contact: bilateral, cross-subsectional, and a combination of the two. Vowels used only the first pattern and consonants used all three. There was an observable distinction between the vowels and consonants in the EPG data, but not in the surface data.

Maeda [Mae90] creates an articulatory model of the vocal-tract that uses 4 parameters for the tongue by studying 1000 cineradiographic frames of spoken French. The four parameters are: jaw position, tongue-dorsal position, tongue-dorsal shape and tongue-tip position.

Wilhelms-Tricarico [WT95] presents a finite element model of the tongue with 22 elements and 8 muscles from work on a physiologically based model of speech production. This initial model does not completely simulate the tongue but shows the feasibility of the methods. In later research [WTP97], he creates a way to control the model.

Research into computer facial animation led Pelachaud et al. [PvOS94] to develop a method of modeling and animating the tongue using soft objects with consideration of volume preservation and penetration issues. This is the first work on a highly deformable tongue model for the purpose of computer animation of speech. Based on the research of Stone [Sto91], they create a tongue skeleton composed of nine triangles. The deformation of the skeletal triangles is based on 18 parameters: 9 length, 6 angle,

and a starting point. During deformation, the approximate areas of the triangles are preserved and collision with the upper palate is detected and avoided. The skeletal triangles are considered charge centers of a potential field with an equi-potential surface created by triangulating the field using local curvature. Their method allows for a tongue that can approximate tongue shapes during speech, as well as effects on the tongue due to contact with hard surfaces.

Lips

Most of the visual signal for speech comes from the mouth area, which is shaped by the lips. In order to achieve realistic speech the lips must be capable of producing the shapes during speech. Despite this, the lips have received little special attention in modeling and instead have just been considered as part of the face geometry. However, animation techniques have considered the lips as important for both speech and facial expressions.

Fromkin [Fro64] reports on a set of lip parameters that characterize lip positions for American English vowels. The parameters are developed from the study of frontal and lateral photographs, lateral x-rays, and plaster casts of the lips. The lip parameters identified are:

1. Width of lip opening.
2. Height of lip opening.
3. Area of lip opening.
4. Distance between outer-most points of lips.
5. Protrusion of upper lip.

6. Protrusion of lower lip.

7. Distance between upper and lower front teeth.

This parameterization of the lips is very good for speech but it does not allow for other lip motions, such as expressive ones.

In Parke's [Par74] ground breaking work, he used around 10 parameters to control the lips, teeth, and jaw during speech synchronized animation. The parameters are chosen empirically and from traditional hand drawn animation methods. From Madsen [Mad69] Parke determines the important capabilities to achieve lip animation are:

1. Open lips for a,e,i

2. Closed lips for p,b,m

3. Oval mouth for u,o,w

4. Tuck lower lip under front teeth for f and v.

5. Move between these lip positions as required.

Finally, The remaining sounds are formed by the tongue and do not require precise animation.

Bergeron [BL85] reports that for the film short "Tony De Peltrie" keyframes were defined then interpolated using curves. The keyframes included those for phonemes. This method of shaping the lips for each phoneme is a common approach. However, it is difficult to create a complete set of poses that encompass the entire space of motion desired. For example, to have a pose of smiling and saying /a/ at the same

time requires having to create that specific pose. By defining shapes that differ from the neutral in a single way, such as smiling, these shapes can be used to define a space of possible shapes. Each base shape is calculated as a displacement [Ber87] from the neutral. By defining a percentage of each base shape, animations that are more complex can be achieved. Our work is similar to this concept; however, we choose our bases using muscle displacements instead of phonemes and expressions, which we believe gives a larger space of possible lip shapes. Another difference is in how we combine the displacements.

Research by the speech community on lip reading involved drawing lip outlines to represent speech. These works concentrate on creating good motion for the purpose of speech intelligibility and not on visual realism. Boston [Bos73] reports that a lip reader was able to recognize a small vocabulary and sentences of speech represented by mouth shapes drawn on an oscilloscope.

Erber [Erb77] also uses an oscilloscope to create lip patterns and claim to create motion that is more natural. Their studies show that lipreading the display gives similar results to those of reading a face directly. Erber and De Filippo [EF78b] drew eyes, nose and a face outline on cardboard, cut out the mouth area and placed it over the oscilloscope. Erber et al. [ESF79] uses high-speed cameras to capture speech and determine lip positions by hand.

Brooke [Bro79] draws outlines of the face, eyes and nose with movable jawline and lip margins. Positional data is hand-captured from a video source. Brooke and Summerfield [BS83] report on a perception study to determine if a hearing speaker can identify the utterances. The natural vowels were identified 98% of the time while the

the synthetic vowels /u/ and /a/ were identified at rates of 97% and 87% respectively. The vowel /i/, which was usually confused with /a/, had a recognition rate or 28%.

Montgomery [Mon80] draws lip outlines on a CRT using data that was hand captured from video frames in a system designed to test lip reading ability. They augment by adding nonlinear interpolation between frames as well as forward and backward coarticulation approximation.

Guiard-Marigny [GM92] measures the lip contours of French speakers articulating 22 visemes in the coronal plane. Assuming symmetry, the vermillion region of the lips is split into three sections and mathematical formulas are created to approximate the lip contours. From polynomial and sinusoidal equations, the 14 coefficients are reduced to 3 using regression analysis. The three parameters are internal lip width, internal lip height and lip contact protrusion. With the same technique on lip contours in the axial plane, Adjoudani [Adj93] identifies two extra parameters to extend the lip model to 3D. The two new parameters are upper and lower lip protrusion. Guiard-Marigny et al. [GMAB94] describe the 3D model in English.

Guiard-Marigny et al. [GMTA$^+$96] replace the polygonal lip model with an implicit surface model using point primitives for fast collision detection and contact surfaces. For the polygonal model, to increase the speed of computation they use a keyframe animation technique. The inbetweens are calculated as the barycenter of a set of extreme lip shapes. There are two extreme shapes per parameter and the barycentric coordinates are the parameters. They build the implicit surface from point primitives for each of the 10 key shapes which are interpolated to create lip shapes. Implicit surfaces give an exact contact surface [Gas93] that allows modeling the interaction of the lips with other objects.

Guiard-Marigny et al. [GMAB97] added a jaw with 6 degrees of freedom using motion captured with a sensor device attached to the lower teeth. They determined that during speech the jaw uses only 3 degrees of freedom: pitch angle, horizontal and vertical position. The original sensor could not be used to determine jaw movement during speech without interfering with the lips. Instead, they assume correlation of the jaw parameters during running speech and place makeup on the chin and use image processing techniques to extract the parameters for the jaw. Their method does not distinguish between head motion and jaw motion. They compared the lip/jaw model to the lip model to just audio and there is a noticeable improvement in perception using the lip/jaw model over just the lip model and both over just audio.

Skull

The underlying bone structure of the human head plays an important part in not only the static shape of the skin, but also in how it deforms via muscle contractions, as the skin and underlying tissue slide over the bone. Possibly, you have heard the phrase about models that "it's all in the bones." In fact, from only the skull, forensic artists can reconstruct the surface. Krogman and Iscan [KI86] report on some of the latest advances in forensic medicine. Therefore, the skull should not be ignored when considering the final shape of the face. However, it has been mostly ignored in facial animation due in part because the skull is internal and difficult to acquire noninvasively.

Koch et al. [KGC+96] report on a system to simulate facial surgery, which requires the data to be accurate for that patient. The data is collected from volumetric medical scans of the patient which includes the skull, skin, and the tissue between. The system

can be used to determine not only facial shape, but also facial function due to muscle action. The system was extended [KGB98] to create facial expressions.

Waters [Wat92] describes a system that uses computed tomography data to derive a facial model obtaining a polygonal mesh of the exosurface of the skull and skin utilizing marching cubes. A muscle layer is constructed equidistant between the skull and skin with cross struts to avoid shearing and a spring constant lower than that of the skin. Linear muscles [Wat87] are then attached throughout the model to achieve animation.

A simple method to simulate a skull is to create a skull that is a constant depth from the surface. The resulting skull is by no means accurate but it does have a roughly accurate shape that allows for interaction between skin, muscle and bone. Waters and Terzopoulos [WT91] have a constant depth skull and muscle layer. Lee extends the Waters model [Lee93, LTW93, LTW95], still using a constant depth skull with skull penetration forces. For skull penetration a normal is calculated for each node, based on a single face, in cylindrical coordinates then converted to the generic mesh coordinates. The force on the node in that normal direction is then negated to simulate muscles sliding across bone.

A.4 Parameterizations

A parameterization is a way to abstract the shape or movement of the surface in $\Re^3$ to something easier to use by an animator or director. Parameterizations are not solely in either the modeling or the animating domain. Parameterizations that are used to define movement may belong on the animation side, while parameterizations that define shape may belong to the model. The issue is very gray, and therefore, we

discuss parameterizations separately from the animation methods and the modeling techniques. An example parameterization is FACS, which is a descriptive definition of the changes on a face, so therefore should be a modeling issue. However, it is also used as a method to generate expressions so it belongs on the animation side. Another example is the parametric Parke model, which has conformation parameters that describe static shape and expression parameters that are used to define motion.

Several layers of parameterizations or several separate parameterizations can be used in a facial animation system. The facial model may be as simple as pure geometry without a parameterization, such as CANDIDE, or it could be very complex with separate parameterizations for parts of the face, such as the facial model describe in this dissertation. The system may use several methods to create animation and each of those methods may define a new parameterization between the animation and the model.

A.4.1 Empirical Methods

A common early method was to create a parameterization based on empirical study of faces or using techniques from traditional hand-drawn animation. If a new expression needed to be added or slightly modified, a new parameter would be added.

Parke [Par74] creates a set of parameters empirically and from traditional hand drawn animation methods to make animating a face simpler and more intuitive. The parameters interpolate vertex positions between two extremes and changing the parameter values over time creates animation. Conformation parameters define the shape of the model and expression parameters are used to create motion.

Nahas et al. [NHS88] use empirical methods to create a set of parameters to create expressions and speech. The parameters move characteristic points, which causes deformations in the B-spline surface.

Pelachaud et al. [Pel91, PBS91, PVY93] create automatic facial expressions with wrinkles from spoken input. They create a high-level language to drive 3D expression animation from speech as a modified Pierrehumbert [PH87] notation.

A.4.2 Muscle Based

Simulated muscles have been a popular method of specifying deformations [Ess94, Gue89a, Gue89b, KMCT96, PW91, Pat91, PBS91, PVY93, Pie89, PB95, Ree90, SVG95, Wai89, WF94, WMTT95] of the model surface. Muscles are generally given an insertion and attachment with a zone of influence. The zone of influence specifies the area of the facial surface that moves during contraction of the muscle. This concept is based on anatomy where a muscle is made up of many fibers and the attachments of all the fibers create an area, which is where the zone of influence is one. Most systems further define another surrounding area where the influence drops from one to zero. This simulates the stretching of the skin near where the muscle is attached. Muscles have been used to modify a polygonal mesh [PW91, Pat91, PBS91, PVY93, WMTT95], a mesh attached with springs [KMCT96, Pie89], finite elements [Ess94, Gue89a, Gue89b], or even splines [Ree90, SVG95, Wai89, WF94].

Muscle based parameterizations use anatomical knowledge of the face as a basis for defining shape or motion. Muscle contractions directly cause the deformations on the surface of the face making them a good choice to base a parameterization on, especially at the model level. At the animation level, it is difficult for an animator

to specify all the muscle values. Instead, another high-level parameterization such as emotions or expressions is often used. Muscle parameters are often normalized between no contraction and full contraction, but specification of forces can also be used.

Waters [Wat87, Wat89] develops a muscle model that can be used to create facial expressions. Muscles are modeled by spring forces with zones of influence and cosine acceleration and deceleration. This method of defining muscle action can be used in a facial model, a human full-body model or even an animal model. This muscle model has been used in many systems [KGB98, Lee93, LTW93, LTW95, PALS97, TW90, UGO98, WT91, Wat92].

Magnenat-Thalmann et al. [MTPT88] describe a system using Abstract Muscle Action (AMA) procedures. An AMA is defined for roughly the action of a single muscle and acts on a specific area of the face. AMAs are combined together to create expressions that can be added to a script track for an actor.

A.4.3 FACS

The Facial Action Coding System [EF78a] (FACS) developed by psychologists Paul Ekman and Wallace Friesen, is designed to describe the changes in facial appearance. One application of this system is to have an observer score the actions of the subject face in order to categorize the expression or emotion. FACS breaks the movements of the face into Action Units (AUs) that roughly correspond to individual muscle movements. Although the system was designed as a descriptive system, facial animation researchers [AHS89, CPB+94, Gue89b, Gue89a, HFG94, KGB98, Pat91, PW91, PBS91, PVY93, SMP81, Pla85, PB95, Ree90, Wai89, Wat87, Wat89, WT91,

AU	Name	Action
1	Inner Brow Raiser	Raises inner portion of the eyebrows and creates horizontal wrinkles on the forehead.
2	Outer Brow Raiser	Raises the outer portion of the eyebrows.
4	Brow Lowerer	Lowers the eyebrows and pulls them closer together while producing deep vertical wrinkles between them.
5	Upper Lid Raiser	Raises the upper eyelid while widening the eye opening.
6	Cheek Raiser and Lid Compressor	Draws the skin from the temple and cheeks towards the eyes while narrowing the eye opening and bagging the skin below the eye.
7	Lid Tightener	Tightens the eyelids while narrowing the eye opening.
41	Lid Droop	The eyelid droops down reducing eye opening.
42	Slit	The eye opening is very narrow while the lids appear relaxed.
43	Eyes Closed	The eyes are closed with no sign of tension.
44	Squint	The eye opening becomes very narrow while the lids appear tensed.
45	Blink	The eyes close and open quickly with no pause while closed.
46	Wink	One eye closes briefly with a slight pause while closed.

Table A.10: FACS Action units for the Upper Face group.

WMTT95, WKMT96] have embraced FACS as a way to generate facial expressions by using the AUs to describe the desired expressions. It has also been used in vision [Red91, AHS89] systems for model-based tracking.

AUs are grouped based upon location and/or type of action. One group is for the upper face and the remaining five groups affect the lower face. The lower face groups are: Up/Down, Horizontal, Oblique, Orbital and Miscellaneous Actions.

The first region is the upper face and the AUs describe changes to the eyebrows, forehead, eye cover fold, and the upper and lower eyelids. Table A.10 lists the AUs for the upper face and describes them.

AU	Name	Action
9	Nose Wrinkler	Pulls the skin along the sides of the nose upwards creating wrinkles along the sides and on top of the nose, increases the infraorbital furrow, causes bagging under the eyes, lowers the brows, pulls up the center of the upper lip. May flare the nostrils and widen the nasolabial furrow.
10	Upper Lip Raiser	Raises the upper lip, deepens the nasolabial furrow, and widens and raises the nostril wings. May increase the infraorbital furrow or part the lips.
15	Lip Corner Depressor	Pulls the corner of the mouth down and may bag or wrinkle the skin under mouth corner.
16	Lower Lip Depressor	Pulls and stretches the lower lip down and laterally. May protrude the lower lip, cause wrinkles on the chin or part the lips.
17	Chin Raiser	Raises the chin and the lower lip. May cause wrinkling on the chin, produce a depression under the lower lip or protrude the lower lip.
25	Lips Part	Parts the lips while the teeth are together or almost together.
26	Jaw Drop	The jaw lowers while the lips may or may not part.
27	Mouth Stretch	Stretches the mouth open as the jaw moves down.

Table A.11: FACS Action units in the Lower Face Up/Down group.

The AUs of the Lower Face Up/Down group move the skin and features in the center of the face upward towards the brow or downward towards the chin. This group of AUs is listed in Table A.11.

The AUs of the Lower Face Horizontal group, which are described in Table A.12, move the skin sideways either from the midline laterally towards the ears or vice versa.

The Lower Face Oblique group consists of AUs that move the skin angularly both upwards and outwards towards the cheekbone. The Lower Face Oblique group is enumerated and described in Table A.13.

AU	Name	Action
14	Dimpler	Tightens and narrows the mouth corners pulling them inward causing wrinkles and/or a bulge while stretching the lips laterally. May cause a dimple or deepen the nasolabial furrow.
20	Lip Stretcher	Moves the lips laterally elongating the mouth and stretching the lips and nostrils. May causing wrinkles near the mouth corners or move the nasolabial furrow laterally.

Table A.12: FACS Action units in the Lower Face Horizontal group.

AU	Name	Action
11	Nasolabial Furrow Deepener	Pulls the upper lip up and out deepening the nasolabial furrow.
12	Lip Corner Puller	Pulls the lip corners back and up deepening the nasolabial furrow.
13	Sharp Lip Puller	Pulls the lip corners sharply up tightening and narrowing them while puffing the cheeks. May deepen the nasolabial furrow or tighten the upper lip.

Table A.13: FACS Action units in the Lower Face Oblique group.

AU	Name	Action
18	Lip Puckerer	Pulls the mouth forward and medially puckering the lips while making the mouth smaller and rounder.
22	Lip Funneler	The lips are pushed outward as in /ur/.
23	Lip Tightener	Tightens and narrows the lips rolling them inwards.
24	Lip Pressor	Presses the lips together, while tightening and narrowing them, without pulling up the chin.
28	Lips Suck	Sucks in the red parts of the lips covering the teeth and stretching the skin above and below the lips.

Table A.14: FACS Action units in the Lower Face Orbital group.

The Lower Face Orbital group involves the muscles around the mouth opening and moves the lips and skin adjacent to the mouth. These Action Units are detailed in Table A.14.

The Lower Face Miscellaneous group contains actions which affect the lower face that do not belong in the other groups. These miscellaneous Action Units are described in detail in Table A.15.

FACS also allows scoring of the positions of the head and eyes. The Action Units used for scoring of the eyes and head positions are listed in Table A.16. Each of these positions should have a level of intensity except for AUs 63-66.

FACS is used by scoring a static picture or movie clip using AUs to describe the observed changes in the face. Most of the AUs are on/off, that is, they are scored if the movement is seen. However, some AUs score as a low (X), medium (Y) or high (Z) level of intensity. AUs can be bilateral or unilateral and are scored with an L or R for the left or right side of the face respectively. There are rules for which AUs can be active at the same time and how to combine them properly. These rules are in the FACS manual [EF78a] and will not be listed here. Scoring is done in three passes:

AU	Name	Action
8	Lip Towards Each Other	The upper lip is pulled down but not over the teeth.
19	Tongue Show	The tongue protrudes showing at least its tip.
21	Neck Tightener	Wrinkles and bulges the skin of the neck and under the chin.
29	Jaw Thrust	The jaw is pushed forward sticking the chin out with the lower teeth in front of the upper teeth.
30	Jaw Sideways	The chin, lower lip and lower teeth are displaced from the midline.
31	Jaw Clencher	A bulge appears along the jaw bone where it is hinged.
32	Bite	The teeth bite the lip.
33	Blow	Air is blown out and the cheeks expand.
34	Puff	The cheeks puff out with the lips closed.
35	Suck	The cheeks are sucked into the mouth producing a crevice in the cheek.
36	Bulge	The tongue is pushed against the cheeks or the lips causing a bulge.
37	Lip Wipe	The tongue licks the lips.
38	Nostril Dilator	Flares out the nostril wing changing the shape of the nostril.
39	Nostril Compressor	Compresses the nostril wing decreasing the nostril opening.

Table A.15: FACS Action units in the Lower Face Miscellaneous group.

AU	Name	AU	Name
51	Head Turn Left	58	Head Back
52	Head Turn Right	61	Eyes Turn Left
53	Head up	62	Eyes Turn Right
54	Head Down	63	Eyes Up
55	Head Tilt Left	64	Eyes Down
56	Head Tilt Right	65	Walleye
57	Head Forward	66	Crosseye

Table A.16: FACS head and eye positions.

first the lower face, then the head and eye positions, and then finally the upper face. An example FACS scoring is, **1C or 1+4C**, where the **C** signifies that minimum requirements are not met, and the **or** means that either AU or AU combination could be met. The + signifies a combination of AUs. Other special codes are: @ for alternative to, < for subordinate to, and > for dominates. Co-occurrence rules have a few exceptions: ESQ - except when sequential: a rule applies unless two actions occur sequentially; EMO - except when motion observed: a rule applies unless motion is seen; and APEX - the apex of an action is the point of greatest change during that action.

Ekman and Friesen [EF78a] give suggested combinations of AUs that translate into emotional terms. Table A.17 lists the AU or AU combinations that describe the universal emotions. Ekman [Ekm73] describes emblems about emotion, in which a person uses the AU or AUs for an emotion to describe an emotion just as if using words without experiencing the emotion. Ekman and Friesen [EF78a] suggest that some of the major emotions listed in Table A.17 may occur as emblems with different timing than for actual emotions. They further suggest that the prototypes are less likely to be used as emblems without quantitative evidence. In addition, certain AUs and AU combinations are used as conversational signals tied to speech rhythm or content and can be confused with emotion. The AUs 1+2, 1+2+5 and 4 are the most common seen [EF78a].

Although FACS is a descriptive coding scheme and has no temporal information, is it used heavily to generate facial animation. Essa [Ess94] removes these deficiencies in FACS and creates FACS++ as a more generative system. Platt [Pla85] determined

Emotion	Prototypes	Major Variants
Surprise	1+2+5X+26, 1+2+5X+27	1+2+5X, 1+2+26, 1+2+27, 5X+26, 5X+27
Fear	1+2+4+5*+20*+[25,26,27] 1+2+4+5*+[25,26,27]	1+2+4+5+[L,R]20*+[25,26,27], 1+2+4+5*, 1+2+5Z+(25,26,27), 5*+20*+(25,26,27),
Happy	6+12*, 12Y	
Sadness	1+4+11+15x(54+64), 1+4+15*+(54+64), 6+15*+(54+64)	1+4+11+(54+64), 1+4+15X+(54+64), 1+4+15X+17+(54+64), 11+15x+(54+64), 11+17
Disgust	9, 9+16+[15,26], 9+17, 10*, 10*16+[25,26], 10+17	
Anger	4+5*+7+(10*,22)+23+[25,26], 4+5*+7+(17)+[23,24]	Any prototype without any one of: 4,5,7, or 10.

Table A.17: Translation of FACS AUs into emotions. An * indicates the AU can be at X, Y or Z level of intensity. [a,b,c] stands for a or b or c. (a,b,c) means a and/or b and/or c or none. For sadness 25 or 26 may appear with all prototypes.

a more generative, as opposed to analytical, set of AUs from FACS by excluding AUs with a temporal component and those AUs with eye, tongue, or head movements.

A.5 Animation Techniques

The human face is very expressive and is used to communicate ideas as well as emotion. Humans have been practicing interpreting faces since birth and we are very sensitive to inaccuracies. This makes realistic animation a very difficult problem. Animating faces involves speech, expressions (which includes emotion, gestures, etc.), and life motions. Life motions include: breathing, non-expressive blinking, head motion and hair motion; they are all the little things that make the model look alive, but do not convey specific information.

Facial animation can be used to tell a story [DKT98, Ree90], for low bandwidth communication [AHS89], for education, to teach the deaf to speechread [Bro79, CM93], for a more familiar human-computer interface, in games, to give inanimate objects human characteristics [PLG91], and with avatars - to name a few.

There are several methods to create animation and all forms of animation have been applied to facial animation at one time or another. It is very common to use several of these methods together to create an animation. This section describes the different techniques that have been used to create facial animation.

A.5.1 Image Based Animation

Image-based methods have been successfully used [AHS89, BS94, CG98, EP97, EP98, YD88a, YD88b] to create facial animation. In a strictly 2D approach, images are used as keyframes and they are morphed between to create animation. The other basic method is to blend the texture on a 3D model using morphing, while also interpolating the 3D model between keyframes. This gives motion that appears more realistic because the small deformations that occur in the skin, such as wrinkles and bulges, will appear since they are in the new texture. Texture information that may have previously been missing, such as inner mouth and eyelids, will now appear. The major drawbacks are that not only must textures for all possible articulations, and all possible combinations of articulations be acquired (and this may not be possible), but they must also be stored. In addition, the appearance and disappearance of wrinkles and shadows will be unrealistic.

Yau and Duffy [YD88b, YD88a] use textures along with a dynamic model to create facial animation. Very simple model deformations are combined with blending textures to create facial animation.

Aizawa et al. [AHS89] describe model-based analysis image coding (MBASIC), which is designed to compress video streams of faces. Facial expressions are synthesized by a combination of clip-and-paste and facial structure deformation. The expressive regions are extracted from the input image, transmitted, and then clipped into place. The 3D model is deformed using deformation rules for each of the AUs of FACS. Control points are defined on the model and the AUs move them based on a parameterization with other points interpolated from the control points.

Brooke and Scott [BS94] describe a system that uses principal component analysis (PCA) and a hidden Markov model (HMM) to create animated speech of 2D images. PCA is performed on a standard test spoken set of 100 3-digit numbers. Fifteen principal components that account for over 80% of the variance are identified. The image data is hand-segmented with the encoded frame sequences phonetically labeled and the 15 principal components describe the image of any phoneme. The data is used to train the HMM model and for each triphone HMMs are constructed with between 3 and 9 states. Animation is achieved by fitting a quadratic through the states which results in a continuous stream of PCA coefficients. The coefficients do not exactly represent the statistics of the HMMs but they only deviate slightly and produce smoother transitions.

Cosatto and Graf [CG98] use a sample based method to achieve facial animation. They break the head into regions and store image samples for each region based on a parameterization for expressions. The parameterization includes visemes in order to

create animations of speech. To produce animation, they determine the appropriate parameters for each frame and then build the image based on the stored samples.

Ezzat and Poggio [EP97, EP98] extract visemes from a spoken corpus, then, based on optical flow, concatenate morphs between the visemes together to produce speech. They record a speaker enunciating one word for each of the 42 phonemes they use. The visemes are extracted by hand from the video sequence, reduced to 16 visemes and the optical flow between the frames is calculated. The optical flow for all frames of a viseme are concatenated together and used as a correspondence for a warp with inverse optical flow used for the reverse warp. The two warps are then blended together with any holes filled in with a special background color.

Pighin et al. [PALS97, PHL+98] describe a system that uses image processing techniques to extract the geometry and texture information. To achieve animation, they create a face mesh and texture for each keyframe and interpolate them. The geometry is linearly interpolated while the textures are blended. To improve the blending, the textures are warped using optical flow to make it fit closer to the texture it is to be blended with, thereby reducing discontinuities.

A.5.2 Hierarchical Control of Animation

Creating a hierarchy of motion, description, surfaces, etc. is a technique that can be used for animating, parameterizing and modeling. The concept is to create several simpler layers to replace a single complex layer to facilitate increased performance or decreased resource requirements. For example, characteristic points, points whose movement are a good indicator of the movements of nearby locations, can be defined

on the surface. Moving the characteristic points directly and the points nearby based on their relationship to the characteristic points simplifies the deformation process.

Platt [Pla85] divides the face into landmarks and transient regions. Landmarks, which can be moved or masked, are permanent locations on the face. The two types of landmarks are structural, such as an eye opening, and surface, such as an eyebrow. Transient regions are locales of change such as the area between brows. Each region of the face is a frame and the entire face is a frame. By connecting the frames together, a model is created. The 7 areas of the face are: forehead, left/right eye, left/right cheek, nose and mouth.

Magnenat-Thalmann et al. [MTPT88] use a hierarchical framework to control a facial model using three levels of control. At the first level, abstract muscle action (AMA) procedures are defined for roughly the action of a muscle and each AMA acts on specific areas of the face. At the next level, AMAs can be combined together to create expressions, where expressions are emotions (smile, cry, etc.) or phonemes. Finally, expressions are included in a script track for an actor to produce animation.

A.5.3 Animation Using Scripting

Scripting gives greater control over the characters to the director or animator and facilitates the reuse of animation. A script is usually a human-readable text file that is created with a text editor, and then sent to the animation system to create animation. Generally, a script that animates one character can be reused to animate another in the same manner. Scripting has been a popular method in facial animation [Els90b, Pat91, PW91, PWWH86, MTT87, MTPT88] as an interface to the animation system.

A.5.4 Procedural Animation

Procedural animation attempts to define motion or modeling as a mathematical description that can be written in a procedural language. Reeves [Ree90] describes how animation shorts at Pixar, such as *Tin Toy* were made. The models are controlled with menv, which is a procedural language. In a typical virtual reality application, one or more users control a character but the many supporting characters must be controlled by the computer. This is typically done with procedural and artificial intelligence methods.

A.5.5 Animating Expressions

Animating facial expressions has been the major focus of research in facial animation. Facial expressions are movement or movements of part or parts of the face. Expressions are used to convey emotion and ideas and they include head nods, eye blinks, raising eyebrows, eye gaze, and high-level emotions. During communication, expressions [Ekm73, EF84] are used to accentuate speech, for turn taking, to convey understanding, and to deceive [Ekm85]. Speech is sometimes considered an expression, other times as separate. Animating speech will be discussed later as a separate topic in Section A.6. As well as creating facial expressions, there has also been research on detecting expressions [EG98, Ess94, Red91, RvdGG96].

Since animating expressions is the main goal of most facial animation systems instead of discussing all the work again in this section, only selected works will be discussed. When learning how to animate expression, it is good to consult books on the psychology of expressions and emotion [Ekm73, EF84, Ekm85] books for hand-drawn animation or even a book for animators using 3D computer graphics [FD99].

Animating the expressions involves a model capable of deformations and a method to describe the deformations of the model to specifically create expressions. Nahas et al. [NHS88] use empirical methods to form a set of parameters for creating expression. Guenter [Gue89a, Gue89b] creates expressions by combining AUs. Patel and Willis [Pat91, PW91] create expressions from existing expressions, from AUs or from muscles.

The expression can also be determined directly from input speech. Pelachaud et al. [Pel91, PBS91, PVY93] present a facial animation system that correlates facial expressions with the intonation of speech. An input utterance is parsed to find intonational patterns, phonemes and timing. Lip shapes and eye blinks are created at the phoneme level with conversational signals and punctuators at the word level. Animation is created by finding the AUs for emotions, adding the AUs for the phonemes, then computing conversational signals, punctuators, head and eye movement and eye blinks. Conversational signals (such as blinking, or raising eye brows) occur on stressed segments of speech with the rate and type affected by the emotion. Punctuators occur at the end of pauses or to signal punctuation, such as smiling and head movements, and are emotion dependent. Manipulators (they only consider blinks) are added for conversational signals and punctuators for regularity. They consider pupil size for both showing emotion and lighting conditions. Pelachaud et al. [PVY93] adds the generation of expressive wrinkles with a spline surface to the system.

The expression and the speech itself can be generated by the computer. This is extremely useful for computer-controlled characters in virtual worlds and games and well as for background characters. Cassell et al. [CPB+94] describe a system for automated generation of conversations with expressions.

A.5.6 Physically Based Animation

In physically based animation, motion is produced by solving Newton's equation of motion, $F = ma$. Vertices of a mesh are treated like particles with mass connected together with springs. The springs are given a natural rest length of the model in its neutral position. External forces, usually to simulate muscles, are applied to nodes of the mesh and the connective springs cause other nodes to move. The system can be numerically integrated with a method such as Euler integration or a higher order Runge-Kutta scheme for much more accuracy. This is considered more of a modeling technique than an animation technique and is discussed in Section A.3.

A.5.7 Keyframing

The keyframing animation technique comes from traditional animation in which a master animator would draw key locations in an animation, such as changes in direction or grabbing an object. Animation would then be produced by drawing all of the frames in between the key frames by interpolating the two keys. Computers are extremely helpful in such a technique as they can produce the in between frames quickly and accurately with a variety of interpolation methods. Keyframing [BL85, deG89, deG90, EP97, EP98, Gou89, Kle89, Pat91, PW91, PBS91, Pel91, PVY93, PALS97, PHL$^+$98, PB95, Wai89] is an extremely popular animation technique in facial animation research. It is often combined with other techniques with the way in which the keyframes are produced, combined, and interpolated varying widely. For instance, Kleiser [Kle89] sculpts the face in key positions, digitizes the sculptures, then interpolates the appropriate keyframes to produce animation. The different positions can be combined together in a single keyframe. deGraf [deG89, deG90] digitizes a

model in extreme positions then uses a puppeteering interface to quickly create the interpolation between the extremes. Gould [Gou89] creates multiple facial expressions then interpolates the expressions, termed multi-target interpolation. Patel and Willis [Pat91, PW91] allow multiple interpolation schemes between keyframes. Pelachaud et al. [PBS91] use B-spline interpolation of keyframes. Pighin et al. [PALS97, PHL+98] create a face mesh and texture for each keyframe and linearly interpolate the geometry and blend the textures.

Track Based

Specifying all of the parameters for a model for each keyframe can be quite expensive. A method to ease this burden is to create tracks for an animation. A track controls a portion of the model such as the position of the eyes, eye blinks, speech, lip positions, etc. allowing the animator to specify only those parts of the animation that are important to that keyframe. The track for a specific behavior, such as a smile, will be similar not only within an animation sequence, but across animations and characters. This promotes the reuse of animation. A more useful method allows tracks to control overlapping sets of parameters which creates the necessity for a method to combine parameter values, such as addition, max value, min value, average value, etc. Many facial animation systems have used tracks. Hofer [Hof93] allows tracks for each facial feature, allowing complex behavior to be built up hierarchically. Magnenat-Thalmann et al. [MTPT88] put expressions down in a script track and use spline interpolation. We use a track based approach in this work to avoid defining all model parameters for each keyframe and to allow specifying a parameter for more than one purpose such as for a phoneme and an expression.

A.5.8 Performance Animation

Performance animation is capturing the motion of some performance, and applying it to a facial model to create animation with many different possible methods. The first method, puppeteering, is lower tech and requires a skilled puppeteer. Some sort of input device that maps input motion to facial motion is used by a skilled puppeteer to create animation. Puppetry allows for producing animations quicker, giving immediate feedback to actors and directors. Usually this is for a real-time system, but it could also be captured for later playback. deGraf [deG89, deG90] uses a puppeteering interface that controls a model that has been digitized in extreme positions. Sturman [Stu98] describes the use of puppetry interfaces to create computer animation. At Media Lab, they produce facial animation by using input devices, such as gloves, pedals and joysticks to control multi-target interpolation. They achieve lip synchronization by the skill of the puppeteer.

Another method is to use image processing and vision techniques to extract motion from video. Motion from a sequence of images is extracted by tracking feature points on the face across frames. Sometimes markers or makeup is used. Terzopoulos and Waters [TW90] use snakes along with make-up to track facial features, which they tie to muscle actuation. Williams [Wil90a] tracks reflective dots on a face and animates corresponding control points on a facial model. Williams [Wil90b] and Patterson et al. [PLG91] improve realism and describe using the method to animate a dog in the film short "*The Audition*."

Saulnier et al. [SVG95] describe a near real-time system that processes a video signal and extracts shape and motion parameters for a human face then uses the parameters to control a virtual clone that acts similar, but not necessarily looks

similar, to the subject. The entertainment industry (movies, TV, games, etc.) uses other forms of motion capture including magnetic and vision systems. The vision systems often use infrared wavelengths.

Animation can also be driven directly from a speech waveform. This technique is performance based, but it does not obtain motion directly from the performance. Simons and Cox [SC90] report on a real-time system with a delay that generates mouth shapes from an input utterance. Pelachaud et al. [PBS91, PVY93] drive facial animation from speech and correlate facial expressions with the intonation of the speech.

A.5.9 Displacement

In displacement animation, deformations are defined as the difference from some neutral state. Bergeron and Lachapelle [BL85, Ber87] define key frames then interpolate them using curves achieving speech-sync by hand. These notes describe the TAARNA system used to animate the title character in the film short "Tony De Peltrie." TAARNA is a command driven interface that allows the creation of keys that can then be interpolated for animation. The system allows facial expressions and body movement. The character in the film short is a caricature with realistic expressions. Five elements are animated independently: left and right eyebrows, left and right eyelids, and general expressions (with mouth). Eyeballs are controlled with the skeletal control. 28 general expressions, including phonemes, made by a live model were photographed and of these 20 were digitized. They painted contour lines on a live model who was photographed making expressions that were then digitized. A model of the characters head was made and digitized and a correspondence from the

live model to Tony was made. Using this correspondence, they created expressions on Tony's face. New expressions are created by interpolating existing ones. Animation is achieved by defining keyframes and using curved interpolation.

Elson [Els90a, Els90b] describes an animation technique using displacement states, either absolute or relative of a base model. Animation is achieved by scripting when each displacement needs to be done. They are layered such that the phoneme displacements for speech can be done in one layer and the displacements for expressions in another.

A.5.10 Natural Language Animation

Takashima et al [TST87] describe an animation system that creates animation from stories written in natural language. The system works on Japanese and is written in Lisp. There are three modules in the system: the story understanding module, the stage directing module and the action generating module. The first module attempts to understand the story, based on actions, and creates a scenario. The stage directing module adapts the story based on heuristics, creating scenes that can be animated by the action module. The action generation module takes the adapted scenario and creates models and motion for the characters and scene.

A.6 Animating Speech

A natural function for human faces is speech. Speech allows us to convey information quickly. Synchronizing animation to audial speech is very important to create smooth, understandable speech. McGurk [MM76] reported that when the lips are showing one sound, the audio a second, a third sound is interpreted. Massaro and Cohen [MC83] show that the visual part of speech is very important. Therefore, it is

crucial to synchronize properly. Without proper synchronization the message can be misinterpreted or ignored. In addition, when the intention is to create a believable human, our familiarity with human speech allows us to immediately know something is wrong, even if we do not know exactly what. It is interesting to note that when using a caricature or cartoon interface that is not intended to look completely realistic, the audience will quickly fill in the missing information and just pay attention to the message, but when attempting to be realistic, the slightest imperfection is quickly noted.

A.6.1 Contribution of the Visual Signal in Speech Perception

McGrath [McG85] describes various experiments to study visual and audio-visual speech perception using both natural and synthetic faces. A study of sentence intelligibility while delaying the acoustical element of an audio-visual signal showed that a delay of $80ms$ or more causes a disruption in understanding. This experiment implies that disruption is on a syllabic and not a phonetic time-scale and lipreading could handle a $40ms$ delay. In another set of experiments, lighting and makeup were used to control visibility of the teeth, tongue, lips, and facial frame of a natural face. The lipreading results were compared to computer generated outlines of faces [BS83] and showed: 1) under these circumstances synthetic is lipread just as well as natural; 2) multidimensional scaling analysis accounted for confusions among natural and synthetic vowels with the same three perceptual dimensions; 3) visibility of the teeth

aided in distinguishing close-front vowels from other vowels, and rounded from unrounded vowels; and 4) appropriate consonant-vowel transitions improved lipreading accuracy. Finally, the synthetic model was able to elicit the McGurk Effect [5]

Massaro and Cohen [MC83] present experiments to determine the contribution of audial and visual information during speech perception. They used recorded video from an actual speaker and generate speech using the Klatt system with the recorded formants of the recorded speaker. They conclude that the visual portion of speech has a strong contribution in what is perceived. They point to other research that confirms their conclusions and research which disagrees with them.

Guirad-Marigny et al. [GMOB95, GMAB97] evaluate speech intelligibility of a lip model [GMAB94] alone and of the same lip model superimposed upon a synthetic jaw and skull, versus just audio. The lip model was able to restore about 1/3 of the missing information when the acoustic signal was degraded and there was a noticeable gain in speech intelligibility when the synthetic jaw was added to the lips.

Brooke and Templeton [BT90] report on research concerned with determining the lowest resolution where it is still possible to discern visual cues of speech. The preliminary study only considers 5 British vowels and the early results show that an area of 16x16 pixels of the oral area with just two levels of intensity is adequate to detect the visual cues of speech. This is good news for cell phone and talking wristwatch applications.

[5]McGurk and MacDonald [MM76] took the visual signal of /ga-ga/ and dubbed it with the audial signal for /ba-ba/. Subjects viewing the video often reported hearing /da-da/.

A.6.2 Head Movement During Speech

Hadar et al. [HSGR83] present a study of head movement during speech and conclude that the head is almost constantly moving during speech and is still during pauses and while listening. They measured the velocity of the head at .4 second intervals and considered a point below a threshold (4° and .2Hz) to be at zero velocity. During speaking turns, 75.7% of the data points were non-zero and during pauses, only 12.8% were non-zero giving strong evidence that during speech the head is constantly moving. When removing pauses greater than 1 second during speech, non-zero velocities were 89.9% and the pauses accounted for 58.8% of the still positions during speech. They noted that as the frequency of the movement increases the amplitude lowers. They define three classes of movements: slow movements (.2 - 1.8 Hz, ordinary movements (1.8-3.7Hz) and rapid movements (3.7-7.0Hz).

A.6.3 Lip Synchrony

Synchronizing the motion of the lips with the audio to create convincing speech is an extremely difficult problem. Figure A.8 is a schematic of the most common method of creating animation of synchronized speech. This method is an extension of the practices of traditional hand-drawn animation [Mad69]. The sound track is broken up into phonemes along with their timing, location and duration. This can be done by hand, automatically, or a combination. A very simple automated method equates loudness with jaw rotation. Other methods include spectrum matching and linear prediction [LP87, Lew91]. The sound track is either a speech waveform [deG89, deG90, MAH90, Pel91, PBS91, PVY93, PB95, SC90] or text [HPW88, MAH90, PWWH86, WL93, WH89, WH90]. The phonemes are

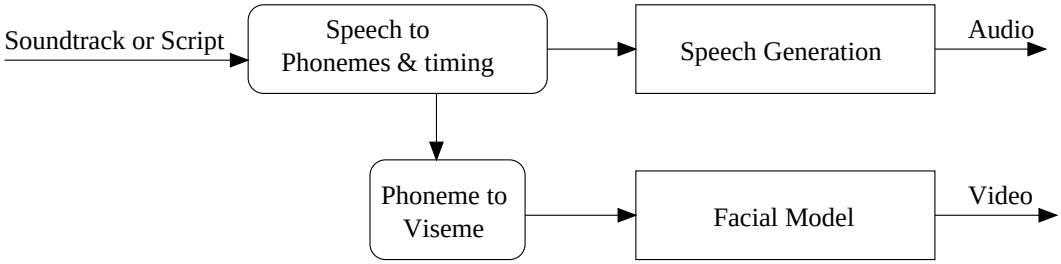


Figure A.8: A typical design for speech synchronized animation.

then converted into speech, if required, and are converted into visemes. To create the visemes a mapping between the phonemes and model deformations must be made. This mapping is usually done in advance and reused. The mapping can be created by extracting information from photographs, video, or using speechreading [Fro64, KBG85, NNT30, Wal82] techniques. Animation is achieved by interpolating between the visemes. Unfortunately, it really is not that simple. During speech, the same phoneme does not always visually look the same, but instead depends on the phonemes before and after. This phenomenon is known as coarticulation and is discussed in Section A.6.4.

In Parke's seminal works [Par72, Par74, Par75], traditional methods [Mad69] are applied to his 3D model by hand. Parke built an animation by going through the following steps:

1. Set the parameters frame by frame resulting in too much motion.

2. Allowed only smoothed jaw movement which created a smoother animation, but still did not look natural nor match the speech well.

3. Added /f/ and /v/ tucks with no improvement.

4. Added variations to the mouth width with some improvement.

5. Modified the model to allow upper lip movement giving improved animation.

6. Finally, added eye, eyebrow and mouth expression motion resulting in more convincing animation.

This technique has stiff requirements on the effort and talent of the animator. The existence of these requirements have led to the development of new automated techniques.

Lewis and Parke [LP87] obtain facial animation by processing a recorded utterance using linear prediction. They use 9 vowels and 3 consonants as phonemes in the system. They determine the parameters for the Parke model for these phonemes creating keyframes to interpolate.

Pearce et al. [PWWH86] modify the Parke model to have a scripting interface that allows a phoneme as input. Hill et al. [HPW88] describe how they modify a speech-synthesis-by-rules system to include the facial model parameters, which are obtained from photographs. Wyvill et al. [WH89, WH90] explain the graphics portion of the research. By inputting the phonemic information to both the facial model and the speech synthesizer, animated speech is generated.

Nahas et al. [NHS88, NHRD90] describe a system that moves attracting points to deform the shape of a B-spline surface. They produce synchronized French speech using 20 phonemes.

deGraf [deG90] [deG89] captures a model in phonemic positions then use voice recognition to detect the phonemes and a puppeteer interface to control expressions and head movements.

Morishima et al. [MAH90] describe a method to generate real-time visual speech by computer from either text or a speech waveform. Text-to-image conversion is accomplished by breaking speech into phonemes, with parameters to aid in handling coarticulation effects. They control 8 feature points, located on the lips and chin, for each phoneme and rely on a muscular structure to deform the rest of the mesh. The phoneme parameters are gathered by processing images of the person voicing the phonemes. Key frames for the phonemes are used with spline interpolation. For speech to image conversion, the waveform is processed by one of two methods, vector quantization or a neural network. For vector quantization, speech is broken up into a code word index. Using the codebook, which has the parameters for the 8 feature points, visual speech is generated. Eye movement, eyebrow movement and blinking are random to give the appearance of realism.

Pelachaud et al. [Pel91, PBS91, PVY93] describe a system that combines intonation of speech with facial expressions. The system produces automatic facial expressions with wrinkles from spoken input. They extract the phonemic information and use rule-governed translation from speech and utterance meaning to facial expressions, primarily those expressions that convey information from the intonation of the voice. Coarticulation is handled by using a look-ahead model.

Waters and Levergood [WL93] describe DECface, which is a system that generates lip-synchronized facial animation from text input. Both the speech and animation are computer generated and synchronized in real-time. The system uses 45 phonemes and for each phoneme, they have an associated mouth position or viseme. To generate speech, the text is converted to a series of phonemes and times, t. The visemes

become keyframes at the specified time and using

$$s' = \frac{s(1 - cos(\pi(s_0 - s)))}{2}$$

they interpolate the keyframes giving acceleration and deceleration near the keys. However, during fluent speech the mouth rarely reaches the ideal viseme shape due to the short interval between positions and the physical properties of the mouth. To simulate this behavior the nodes are assigned a position, mass, and velocity and using spring forces between them, Euler integration is performed giving the system elastic behavior. The visemes were defined form studying snapshots of a person mouthing CvC and VcV strings.

Provine and Bruton [PB95] describe a method of synchronizing lips with audio designed for model-based communication. The speech is broken up into fundamental speech units (FUs): phonemes and diphthongs, Phonemes have one keyframe and diphthongs have two positions associated with them using $\frac{1}{3}$ of the time for the FU of the initial position and the rest of the time for the final position. Instead of interpolating between the geometric positions, they interpolate the muscle parameters based on FACS for each FU allowing the speech to be encoded along with other facial motion. When a muscle is involved in both an expression and an FU the maximum pull factor of the two actions is used. Coarticulation effects are modeled by looking ahead when doing parametric interpolation.

An approach that bypasses the coarticulation problem is to capture the motion of the lips from the person actually speaking the sound track. The resulting motion is synchronized with the speech and very convincing if the motion is captured with sufficient resolution. However, novel speech and speech modification become issues. The motion can be captured by hand or with automated techniques.

Patterson et al. [PLG91] film an actor giving the performance, capture the motion from the video sequence, and use it to drive animation of another character along with the original soundtrack, giving lip-sync for free. They are able to animate any character by creating a correspondence between the captured points and the character. For instance, they can animate a talking fence, statue or dog.

Early work in the speech and hearing community involved drawing lip outlines extracted from photographs by hand on CRTs [Bro79, Bro82, BS83] or oscilloscopes [Bos73, Erb77, EF78b, ESF79].

A.6.4 Coarticulation

Coarticulation is one of the most important issues for computer generated visual speech. Coarticulation is the blending affect that surrounding phonemes have on the current phoneme. Both the previous and future phonemes can affect the current phoneme. Coarticulation is a byproduct of the finite acceleration and deceleration of the tissues involved in the vocal tract. For example, when saying "to", the lips round during the production of /t/ in anticipation of /oo/ but when saying "ta" there is no rounding. Another example is the /h/ in "how", which has the start of the rounding of the lips to produce the /o/ sound, whereas in "hat" there is no rounding. The segment of animation that produces /h/ in "how" and the /h/ in "hat" should look different, but if a system only interpolates between visemes they will look very similar. If speech is produced at a fast rate, the blending will be increased.

Pelachaud et al. [Pel91, PBS91, PVY93] handle coarticulation using a look-ahead model [KM77] that considers articulatory adjustment on a sequence of consonants followed or preceded by a vowel. They integrate geometric and temporal constraints

into the rules to make up for the incompleteness of the model. The algorithm works in 3 steps: 1) apply a set of coarticulation rules to context dependent clusters, 2) consider relaxation and contraction time of muscles in an attempt to find influences on neighboring phonemes, and 3) check the geometric relationship between successive actions is checked.

Cohen and Massaro [CM93] extend the Parke model for better speech synchrony by adding more parameters for the lips and a simple tongue. They include a good discussion of coarticulation research by the speech and hearing community. They use the articulatory gesture model of Löfqvist [Löf90], which uses the idea of dominance functions. Each segment of speech has a dominance function for each articulator that increases then decreases over time during articulation. Adjacent segments have overlapping dominance functions that are blended. They use a variant of the negative exponential function

$$D_{sp} = \begin{cases} \alpha_{sp} e^{-\theta_{\leftarrow sp}|\tau|^c} & \text{if} \tau \geq 0 \\ \alpha_{sp} e^{-\theta_{\rightarrow sp}|\tau|^c} & \text{if} \tau < 0 \end{cases}$$

where D_{sp} is the dominance of parameter p for speech segment s, α gives the magnitude of the dominance function, τ is the time distance from the segment center, c controls the width of the dominance function, $\theta_{\leftarrow sp}$ is the rate on the anticipatory side, and $\theta_{\rightarrow sp}$ is the rate after the articulation.

Waters and Levergood [WL93] describe DECface, which is a system that generates lip-synchronized facial animation from text input. Using DECTalk to generate phonemes and a waveform, they associate a keyframe with each phoneme and interpolate between them. To simulate acceleration and deceleration they use

$$s' = \frac{s(1 - cos(\pi(s_0 - s)))}{2}$$

for the interpolation. However, in fluent speech the target position is rarely achieved so they assign a mass to the nodes and then use Hookean spring calculations, with simple Euler integration, to approximate the elastic nature of the tissue. Expressions are created with Waters' muscle model [Wat87].

A method to deal with coarticulation, popular in 2D image techniques, is to use triphones instead of phonemes. A triphone is a combination of three phonemes. More information must be stored but the resulting transitions are more realistic. This is popular in vision techniques where triphones can be input as train data then reused for animation.

Brooke and Scott [BS94] describe a system that uses principal component analysis (PCA) and a hidden Markov model (HMM) to create animated speech of 2D images. PCA is performed on a standard test set of 100 3-digit numbers spoken by a native British-English speaker and 15 principal components were identified which accounted for over 80% of the variance. The image data is hand segmented with the encoded frame sequences phonetically labeled and the 15 principal components describe the image of any phoneme. The data is used to training the HMM model and for each triphone HMMs are constructed with between 3 and 9 states. Animation is achieved by fitting a quadratic through the states resulting in a continuous stream of PCA coefficients. The coefficients do not exactly represent the statistics of the HMMs, however they only deviate slightly and produce smoother transitions.

A.6.5 Emotion in Speech

Ekman [Ekm73] describes emotional emblems, in which a person uses the AU or AUs for an emotion to describe an emotion, just as if using words, without experiencing the emotion. An example emblem is smiling when saying, "I've had a great day." Ekman and Friesen [EF78a] suggest that some of the major emotions may occur as emblems with timing different than for actual emotions. In addition, certain AUs and AU combinations are used as conversational signals tied to speech rhythm or content and can be confused with emotion.

van Bezooyen [vB84] describes research into how emotion is voiced. The aim of the work was to establish the characteristics and recognizability of vocal expressions of emotion in Dutch as a function of sex, age, and culture. Vocal refers to how things are said as opposed to what is said. The study uses neutral, disgust, surprise, shame, interest, joy, fear, contempt, sadness, and anger as the different emotions. A speaker enunciates 4 phrases using the 10 different emotions several times and the best sample for each emotion is collected. This procedure is repeated twice for 4 different speakers giving 320 recorded emotions. The database of emotions is used to test perceptual characteristics, determine the base set of perceptual characteristics, find acoustic correlates of the perceptual parameters, determine recognition based on age and sex and also by native language. The perceptual characteristics used are: lip rounding, lip spreading, laryngeal tension, laryngeal laxness, creak, tremulousness,whisper, harshness,pitch level,pitch range,loudness,tempo,precision of articulation.With statistical analysis, van Bezooyen determines that loudness, laryngeal laxness, pitch level, pitch range, and laryngeal tension are all that are needed to classify the emotions.

Murray and Arnott [MA96] develop the Helpful Automatic Machine for Language and Emotional Talk (HAMLET), which uses DECTalk for rule-based speech synthesis. The system controls parameters (underlying voice quality, pitch, and timing of phonemes) to change the basic intonation contour of the utterance.

A.6.6 Animating Conversations

Cassell et al. [CPB+94] describe automated generation of conversations with expressions. Their system automatically generates animated conversations between multiple human-like agents with speech, intonation, facial expressions and hand gestures. A dialogue planner creates the text and intonation of the utterances that are used to create synchronized speech, facial expressions, lip motion, eye gaze, head motion, and hand gestures. Communication is achieved using language and nonverbal communication such as facial expressions and hand gestures.

Movements of the head and facial expressions are characterized by their relationship to the linguistic utterance and their significance in transmitting information, and can be separated into five groups. *Syntactic functions* accompany speech flow and are synchronized at the verbal level. These facial movements, such as raising the eyebrows, nodding, or blinking, appear on an accented syllable or a pause. *Semantic functions* emphasize what is being said, substitute for a word or refer to an emotion, such as wrinkling the nose when talking about something disgusting. *Dialogic functions* regulate speech flow and depend on the relationship between the participants of the conversation. *Speaker and listener characteristic functions* convey the speaker's social identity, age, emotions and attitude. For instance, when speaking with a friend

there will be more eye contact than when lying to someone. Finally, *listener functions* correspond to reactions to the speaker such as signals of agreement, attention or comprehension.

Hand gestures include iconics, metaphorics, deictics, and beats. Iconics represent a feature of the speech such as creating a circle with your hands to describe something round. Metaphorics represent an abstract feature of the speech such as making a grabbing motion with a hand while talking about taking something. Deictics indicate a point in space and refer to people, places and other spatializeable entities. An example deictic is to point to someone and say "her." Beats are small formless waves of the hand for heavily emphasized words, turn taking, and other special linguistic work. An example would be to briefly raise a hand up and down while saying "all right."

Synchronization of hand movements and gaze is handled by Parallel Transition Networks (PaT-Nets) that allow coordination rules to be encoded as simultaneously executing finite state automata. The gesture generator will generate gestures during the utterance and a coarticulation model will guarantee there is enough time to complete the gestures by shortening or aborting them altogether. Eye gaze, head movement and facial expressions are specified separately. Facial expressions connected to intonation are automatically generated, while other kinds of expressions, such as emblems, are specified by hand. Gaze is separated into planning, comment, control and feedback. Facial expressions are clustered into functional groups: lip shape, conversational signal, punctuator, manipulator and emblem.

Takeuchi and Nageo [TN93] modify a speech dialogue system to add facial displays. They then analyzed the effectiveness of the dialogue system with and without the

facial displays. They found that conversations between users and the dialogue system were more successful when facial displays were used.

Facial displays are communicative signals that help coordinate conversation, and are primarily social. They use three types of displays based on Chovil [Cho89]:

- *Syntactic Displays.* Facial displays that mark stress on words or clauses, are connected with the syntactic aspects of the utterance, or are connected with the organization of the speech. The syntactic displays are:
  - Exclamation Marks - eyebrow raising.
  - Question marks - eyebrow raising or lowering.
  - Emphasizers - eyebrow raising or lowering.
  - Underliners - longer eyebrow raising.
  - Punctuations - eyebrow movements.
  - End of an utterance - eyebrow raising.
  - Beginning of a story - eyebrow raising.
  - Story continuation - avoid eye contact.
  - End of a story - eye contact.
- *Speaker Displays.* Facial displays that illustrate the idea being verbally conveyed or add additional information to the ongoing verbal content. The speaker displays are:
  - Thinking/Remembering - Eyebrow raising or lowering, closing the eyes, and pulling back one side of the mouth.

  - Facial shrug/"I don't know" - Eyebrow flashes, mouth corners pulled down, and mouth corners pulled back.
  - Interactive/"You know?" - eyebrow raising.
  - Metacommunicative/Indication of sarcasm or joke - eyebrow raising and looking up and off.
  - "Yes" - Eyebrow actions.
  - "No" - Eyebrow actions.
  - "Not" - Eyebrow actions.
  - "But" - Eyebrow actions.

- *Listener Comment Displays.* Facial displays made by the listener that are made in response to the utterances of the speaker. Listener comment displays are:
  - Backchannel/Indication of attendance - eyebrow raising, mouth corners turned down.
  - Indication of loudness - eyebrows drawn to center.

The listener may also use displays to acknowledge their level of understanding to the speaker. The understanding level displays that Takeuchi and Nageo [TN93] use are:

- Confident - Eyebrow raising, head nod.
- Moderately confident - Eyebrow raising.
- Not confident - Eyebrow lowering.
- "Yes" - Eyebrow raising.

The listener may also make displays that portray how they have evaluated the utterance. The evaluation of utterance displays that Takeuchi and Nageo [TN93] use are:

- Agreement - eyebrow raising.
- Request for more info - eyebrow raising.
- Incredulity - longer eyebrow raising.

A.7 Surgery

Another application for facial animation is in surgical planning [DWD+83, KAA83, KGPG96], training, [DWD+83, Pie89], and simulation [KAA83, KGPG96, KGC+96, KMCT96, Pie89, RGTC98]. A computer simulation of surgery can allow a surgeon to practice an upcoming surgery and simulate the post-operative shape and movement to use in planning the surgery. A simulation of wound closure allows a surgeon to practice different closure techniques to gain experience as well as planning for a particular operation.

Dev et al. [DWD+83] describe a commercial system that takes CT images and creates a 3D model of the bone, which can be used to mill a model. Karshmer et al. [KAA83] describe work in progress on a system designed to view pre-operative patient data and simulate post-operative results over time. Keeve et al. [KGPG96] describe a system to aid in surgical planning that uses CT data semi-automatically registered to laser scanned surface data. The data is fit to a generic finite element mesh and then the bone is modified and the resultant surface is simulated.

Koch et al. [KGC+96] report on a system to simulate facial surgery, which requires that the data be accurate for each patient. They use a higher-order finite element surface model with a mesh of springs attached to the skull to get greater accuracy. The data is collected from volumetric medical scans of the patient, which includes the skull, skin and the tissue between. The system can be used to determine not only facial shape, but also facial function due to muscle action. The system was extended [KGB98] to allow for the creation of facial expressions to further evaluate the surgical results.

Pieper [Pie89] discusses a system for simulating plastic surgery as well as the results of the surgery. The skin model has a tetrahedral structure connected via springs with three layers: the epidermis, the dermis and subcutaneous cellular tissue. Springs are also used to model constraint forces on the skin, which include muscles, gravity and draping. The three important properties of facial tissue Pieper is concerned with are: non-linear viscoelastic behavior, nonhomogeneity, and anisotropic response due to irregular anatomy.

Konno et al. [KMCT96] describe Computer Aided Facial Expression Simulation (CAFES), a facial surgery simulation system designed to aid in the surgery of cases with facial paralysis. Determining how to correct problems of paralysis currently is done by experience of the surgeon. This system is designed to aid in planning by creating a tool to determine post operative facial motion. To create the 3D geometry a Bezier contour curve is fit to the skin and skull and radial lines are projected at regular intervals from the center of the skull in each CT slice. The intersections of the lines with the contour curves form a set of vertices in each sliced that, when joined together, form a polygonal mesh. The mesh is treated as a mass-spring lattice and

spring muscles are added with a GUI by indicating the insertion points. They use fourth order Runge-Kutta integration to animate the system.

Roth et al. [RGTC98] use a finite element approach for volumetric modeling of soft tissue for accurate facial surgery simulation. They extend linear elasticity by adding incompressibility and nonlinear behavior with a Bernstein-Bezier formulation. The Bernstein-Bezier form is a higher-order interpolation and suitable for rendering and modeling. In addition it has an integral polynomial form of arbitrary degree and an efficient subdivision.

A.8 Vision

Many of the applications and problems of facial animation overlap with those of computer vision. For instance, in order to provide motion capture data, the subject must be tracked, and one way to track is with computer vision techniques. Using a 3D model to aid in tracking provides more robust algorithms. This 3D model should be deformable with a small number of parameters to describe it. Therefore, facial modeling can aid in facial tracking, which gives data to create facial animation of the facial model.

Another area of overlap is in user interfaces. A system that provides a talking head to provide information to a user would greatly benefit by a vision system that can read the users emotions and speech as input. Carrying the paradigm of speech as a natural way to communicate for humans, it makes sense for the human to provide information as well as receive it with speech. This section provides a glimpse at some of the research that is primarily in the computer vision domain, but has crossover with facial animation.

A.8.1 Tracking

Yuille et al. [YCH88] use deformable templates to track facial features. An energy function tying edges, peaks, and valleys in intensity and the intensity in the image to the parameters of the deformable templates is minimized to find the best fit of the template. The function is updated by steepest descent instead of sampling for speed. The templates use a priori knowledge of the feature being searched for, have global structural forces, and have a finite number of parameters that can be used as a compact representation of the feature.

Reddy [Red91] presents algorithms for tracking facial features using the CANDIDE model and interpreting the motion as a linear combination of global and local motion. A predictive method, based on FACS, was developed for local motion tracking. Feature points are detected and tracked with deformable curves and search regions.

Blake and Isard [BI94] present a method of tracking the outline of a non-polyhedral object at frame-rate, without special hardware, using only a workstation with a video camera and framestore. The algorithms are a synthesis of B-spline curve representation, deformable models, and control theory. As an example of their tracking algorithms, they present results from tracking rigid and non-rigid lip motion. They present two algorithms: one employs Kalman filtering and the second is a learning algorithm that builds a dynamic model from examples.

Saulnier et al. [SVG95] describe a near real-time system that processes a video signal and extracts shape and motion parameters for a human face then uses the parameters to control a virtual clone that acts similar, but not necessarily looks

similar, to the subject. The system tracks head and lip movements, closing and opening of the eyelids, gaze direction and eyebrow contractions

A.8.2 Face Detection and Recognition

Face recognition is an important topic in computer vision. There are two goals: detecting [SNK72] a human face or faces in an image, and recognizing [Kan77, MP95] a particular person from an image of their face. Face detection and recognition have applications in human computer interaction and security.

Sakai et al. [SNK72] present a method to detect human faces using a detection-after-prediction method with feedback. The input consists of binary images of edges from a Laplacian filter. The algorithm detects eyes, nose, mouth, chin, and chin contour and does not proceed in a predetermined order, but instead uses the feedback of previous steps along with a priori knowledge of the organization of a face to choose the best order for the current input. Their method works well for simple faces and for faces that are looking forward, however, faces with glasses and beards are problematic.

Moghaddam and Pentland [MP95] describe a system for model-based coding of faces. They determine the distance from face-space (DFFS) for each pixel, which is the error rate for that pixel. The image is linearly scaled and the DFFS for each pixel is calculated. By finding the global minimum, they find the best match for a face. Subdivision of the image allows them to find eigeneyes, eigennoses, eigenmouths, etc. The subdivision along with heuristics and special knowledge allows a more robust method of finding the face. This method allows them to detect faces, detect a particular face and compress the signal.

A.8.3 Compression

Very high compression ratios of video and images can be achieved if the contents of the image(s) are considered instead of the pixels. For instance, in video or images of faces extracting high-level information about the face and just transmitting or storing that high-level information results in a large reduction of storage requirements. The two common methods are to describe the face in terms of a face or to fit a parameterized facial model to the image and and use those parameters to describe the face.

Sirovich and Kirby [SK73] define eigenpictures, which can represent a picture of a face by a low dimension vector to within some error bound. Moghaddam and Pentland [MP95] use eigenfaces to define a face as 100 bytes of data allowing for high compression. Reinders et al. [RvBSvdL95] describe a system that fits the CANDIDE model to a video sequence so that only parameters for the 3D model need to be sent instead of the entire video stream. Saulnier et al. [SVG95] extract shape shape and motion parameters of a 3D model for a human face then uses the parameters to control a virtual clone that acts similar, but not necessarily looks similar, to the subject. Provine and Bruton [PB95] break a video sequence of a talking person into fundamental speech units (FUs): phonemes and diphthongs, and transmit only the model parameters to recreate the speech.

A.8.4 Speech Reading

Chandramohan and Silsbee [CS96] describe a method for modeling the shape of a speakers mouth. They use a hybrid matching system using a pixel-based algorithm to find a coarse location of the mouth to select an appropriate deformable template,

and then use the selected template to find the mouth. Templates are sensitive to lighting conditions and to initialization and are generally good at recognizing only a subset of the total possible shapes. Preprocessing the image to remove lighting biases and the pixel based template matching algorithm allows good initialization. Using a set of templates representing the different mouth shapes increases accuracy.

A.8.5 Expression Detection

Reinders et al. [RvdGG96] describe a system that uses a Bayesian Belief Network to detect facial expressions. They use different modalities (audio and visual) and different detection methods within each modality to improve overall system robustness. Since each method is just an estimation, they use the framework of a Bayesian Belief Network to increase reliability. This algorithm is part of a man-machine interaction system that uses facial analysis and facial synthesis.

A.8.6 Interfaces

Speech is a natural way for humans to communicate, in fact it is a common form of communication in everyday life. Speech is therefore an excellent choice for communication between humans and machines. Giving information to a user in the form of a talking head makes the receipt of that information natural for the user. Conversely, attaining information from a users face or speech completes a natural way for a human to interface with a computer. Using speech may also make the experience more comfortable for users, especially ones that have a phobia of machines. Some of the promising early results of this young science are discussed in this section.

Reinders et al. [RvdGG96] describe a system that uses both audio and video along with multiple detection methods to create a man-machine interface. Ohmura et al. [OTK88] track three markers to detect face direction for a menu selection interface.

King [Kin93] proposes a holistic model of human-computer interaction that interprets gaze, pupil size and facial expressions. They also report on an experiment to determine if humans express facial emotion during human computer interaction. Subjects were given six tasks to do at a computer and were secretly videotaped during the two-minute tasks. The video was analyzed using FACS and it was determined that emotions were a part of human-computer interaction. They also found a high rate of covert and overt masking.

A.9 Standards

Standards allow researchers who tackle different areas of the same problem to easily integrate their systems by using a common interface, data stream, communication protocol, etc. The coming of standards also signifies the importance of an area and signals its coming of age. Facial animation is still relatively young, but it is very important in human communication. To date, only one standard has been created for facial animation, and that is for the compression of multimedia applications.

The first attempt at creating a standard was a workshop hosted at the University of Pennsylvania on facial animation sponsored by the National Science Foundation and the Institute for Research in Cognitive Science at the University of Pennsylvania. The workshop [PBV94] brought together a group of researchers from computer graphics, linguistics and psychology to get a view of the state of the art and define one or more "standards" in facial models. "The goal was to define scientifically defensible,

computationally reasonable, and experimentally useful computational facial models as the basis for future research and development. Common facial models within as well as across discipline boundaries will accelerate applications, accessibility, interoperability, and reduce redundant developments, software costs, and animation control incompatibility."

They attacked three areas: facial modeling, recognition/data mapping, and control. Modeling deals with facial modeling, applications, the face and its attributes, facial deformation, data control and validation of modeling results. Recognition/data mapping considers facial feature acquisition and processing, general facial data processing techniques, and visual speech feature processing. While control considers manipulation and control frameworks for facial animation, with a goal to provide a complete and succinct set of guidelines for controlling facial animation.

Although no standard was developed, the effort produced dialog between many researchers in the field of facial animation. The final report [PBV94] contains a good foundation for future work on a standard.

A.9.1 MPEG-4 SNHC FBA

The Motion Pictures Expert Group (MPEG) [Gro00] is a working group of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC) developing standards for the coded representation of digital audio and video. MPEG-4 is the standard for multimedia applications. A subgroup of MPEG, Synthetic - Natural Hybrid Coding (SNHC), develops standards for the integrated coded representation of audio and moving pictures of natural and synthetic origin with concentration on the coding of synthetic data. The Facial and Body

Animation group (FBA) is a group within SNHC that developed a standard for coding facial and body animation applications.

The standard specifies how parameters that calibrate and animate synthetic faces can be coded without standardizing the models themselves [ISO99a, ISO99b]. The Face Definition Parameters (FDPs) calibrate the model while the Face Animation Parameters (FAPs) animate it. A decoder has a generic face that can be shaped via the FDPs, or a new mesh can be coded. An optional texture map can also be applied. The standard has codes for visemes, as well as expressions. Feature points on the face can also be controlled. Moreover, the mesh itself can always be controlled.

A coded stream can use the default face model of the decoder, can modify it using FDPs or send a new face model along with information on how to animate it. The Face Animation Table (FAT) is a functional mapping from incoming FAPs to the feature control points of the face mesh. Face Animation Parameter Units (FAPUs) are defined as distances between key facial features (iris diameter, eye separation, mouth width, etc) and are used to scale the FAPs. FAPs are used to animate the face and can be high or low-level. High-level FAPs for expressions and visemes exist. There are low-level parameters to control head rotations, eye rotations, and key feature points on the face.

Ostermann [Ost98] describes how facial animation is achieved within the MPEG-4 standard. Escher and Magnenat-Thalmann [EPMT98] describe how a generic mesh (or a calibration mesh) in MPEG-4 can be deformed to be personalized using Dirichlet free form deformations. Eisert and Girod [EG98] estimate SNHC Facial Action

Parameters (FAPs) along with objects points for facial expression tracking for teleconferencing applications. Tekalp and Ostermann [TO99] give an overview of the synthetic visual objects for faces and 2D meshes.

APPENDIX B

MEDICAL APPENDIX

B.1 Terms of location and orientation

Anterior Toward the front.

Afferent For a vessel or nerve, towards the central structure.

Buccal Pertaining to the mouth or inner surfaces of the cheeks.

Caudal Toward the tail.

Cervial Towards the gum (for teeth).

Cervical Pertaining to the neck.

Cephalic Pertaining to the head. Also toward the head, the same as rostral.

Cephalocaudal From the head toward the tail.

Contralateral The opposite side.

Coronal Plane Named after the coronal suture of the skull, a vertical plane from side to side, made at right angles to the medial plane, dividing the body into anterior and posterior portions.

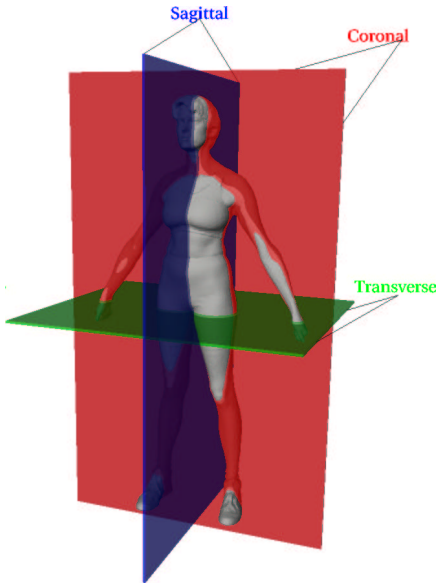


Figure B.1: The planes that describe location and orientation in human anatomy.

Costal Pertaining to the ribs.

Cranial Toward the head; upper or higher.

Deep Away from the surface.

Dorsal In the human, the same as posterior.

Efferent For a vessel or nerve, away from the central structure.

Frontal Plane The same as coronal plane.

Inferior Towards the feet; under or lower. Below. In the human, the same as caudal.

Ipsilateral On the same side.

Lateral Toward the side. Away from the midline axis of the body or structure.

Medial Toward the midline axis (mid-sagittal plane) of the body or structure.

Medial Plane Same as the sagittal plane.

Occlusally Towards the biting surface (for teeth).

Parasagittal plane Any plane parallel to the medial plane.

Posterior Toward the back.

Rostral Toward the head.

Sagittal Plane Named for the sagittal suture of the skull, is synonymous with the medial plane and is a vertical cut dividing the body into left and right portions.

Superior Toward the head. Upper or higher. Above. In the human, the same as rostral or cranial.

Transverse Plane A horizontal cut at any level made at a right angle to the axis of the body, dividing the body into upper and lower portions.

Superficial Toward the surface.

Ventral In the human, the same as anterior.

B.2 Body tissues

Cells of like form and function are grouped together into *tissues*. Combinations of tissues that together serve a specific body function are called *organs*. Primary life functions are carried on by groups of organs called organ *systems*.

Bone. Dense and rigid framework of the body.

Cartilage Similar to bone, except slightly or extremely flexible and softer than bone.

Hyaline cartilage. Relatively rigid.

Elastic cartilage. Nonrigid, Easily Deformed.

Connective tissue. Binds together the other tissues and organs of the body

Ligaments. Slightly elastic strands of tissue that interconnect the bones and cartilages.

Membranes. Broad, flat ligaments.

Tendons. Inelastic strands of tissue that connect muscles to bones.

Aponeuroses. Broad, flat tendons.

Fascia. Bands of connective tissue that lie deep to the skin and invest the muscles and organs.

Muscle. Uniquely capable of contraction upon neural stimulation.

Striated muscle. Capable of rapid, voluntary contraction. Has attachments that provide for movement of body parts.

Smooth muscle. Capable of slow, involuntary contraction. Found primarily in the digestive tract and vascular system.

Epithelial tissue. Performs the functions of protection and secretion.

Skin. Covers the external body surfaces.

Mucous membrane. Covers the internal surfaces of body cavities and passages that communicate with the exterior.

Nervous tissue. Uniquely capable of transmission of electrical impulses.

B.3 Anatomical Terms of the Face

Vermillion zone The red part of the lips.

Epithelium The skin of the lips.

Labium superius oris The upper lips.

Labium inferius oris The lower lips.

Rima oris The point of contact of the lips during closure.

Angularis oris The lateral margins of the mouth, also known as the angle of the mouth.

Angle of the mouth The lateral margins of the mouth, also known as the angularis oris.

Tubercle A slight prominence at the midsection of the upper lip which appears in most subjects.

Philtrum The shallow depression extending from the upper lip to the nose.

Cupid's bow Formed by the philtrum and the rounded parts of the lips.

Labiomarginal sulcus A groove running in a posteriorly convex arch from the corner of the mouth toward the lower border of the mandible, which appears with age.

BIBLIOGRAPHY

[Adj93] Ali Adjoudani. Élaboration d'un modèle de lèvres 3d pour animation en temps réel. Master's thesis, Mémoire de D.E.A. Signal–Image–Parole, Institut National Polytechnique, Grenoble, France, 1993.

[AHS89] K. Aizawa, H. Harashima, and T. Saito. Model-based analysis-synthesis image coding (MBASIC) system for a person's face. *Signal Processing: Image Communication*, 1(2):139–152, October 1989.

[AS92] Taka-aki Akimoto and Yasuhito Suenaga. 3d facial model creation using generic model and front and side view of face. *IEICE Transactions on Information and Systems*, E75-D(2):191–197, March 1992.

[AUK92] Ken-ichi Anjyo, Yoshiaki Usami, and Tsuneya Kurihara. A simple method for extracting the natural beauty of hair. In *Computer Graphics (SIGGRAPH '92 Proceedings)*, volume 26, pages 111–120, July 1992.

[AWS90] Taka-aki Akimoto, Richard Wallace, and Yasuhito Suenaga. Feature extraction from front and side views of faces for 3d facial model creation. In *IAPR Workshop on Machine Vision Applications, Tokyo*, pages 291–294, November 28-30 1990.

[Ber87] Philippe Bergeron. 3-d character animation on the symbolics system. *SIGGRAPH '87 course notes: 3-D Character Animation by Computer*, July 1987.

[BI82] Richard W. Brand and Donald E. Isselhard. *Anatomy of orofacial structures*. C.V. Mosby, St. Louis, 1982.

[BI94] Andrew Blake and Michael Isard. 3D position, attitude and shape input using video tracking of hands and lips. In Andrew Glassner, editor, *Proceedings of SIGGRAPH '94 (Orlando, Florida, July 24–29, 1994)*, Computer Graphics Proceedings, Annual Co, pages 185–192. ACM SIGGRAPH, ACM Press, July 1994.

[BL85] Philippe Bergeron and Pierre Lachapelle. Controlling facial expressions and body movements in the computer-generated animated short 'tony de peltrie'. *SIGGRAPH '85 Advanced Computer Animation seminar notes*, pages 1–19, July 1985.

[BM84] Hugh E. Bateman and Robert M. Mason. *Applied Anatomy and Physiology of the Speech and Hearing Mechanism*. Charles C. Thomas, Springfield, IL., 1984.

[Boo91] Fred L. Bookstein. *Morphometric Tools for Landmark Data*. Cambridge University Press, 1991.

[Bos73] D. W. Boston. Synthetic facial communication. *British Journal of Audiology*, 7:95–101, 1973.

[Bro79] N. M. Brooke. Development of a video speech synthesiser. In *Speech communication, recent developments in audiology, noise in industry : Autumn conference 1979 : Windermere Hydro Hotel*, pages 41–44. Institute of Acoustics, July 1979.

[Bro82] N. M. Brooke. Video speech synthesis for speech perception experiments. *Journal of the Acoustical Society of America*, 71:S77(A), 1982.

[BS83] N. M. Brooke and Quentin Summerfield. Analysis, synthesis and perception of visible articulatory movements. *Journal of Phonetics*, 11(1):63–76, January 1983.

[BS94] N. M. Brooke and S. D. Scott. Computer graphics animations of talking faces based on stochastic models. In *Proceedings of the International Symposium on Speech, Image-processing and Neural Networks, Hong Kong*, pages 73–76. IEEE, 1994.

[BT90] N. Michael Brooke and Paul D. Templeton. Visual speech intelligibility of digitally processed facial images. *Proceedings of the Institute of Acoustics (Autumn Conference)*, 12(10):483–490, 1990.

[BTCC00] Alan W. Black, Paul Taylor, Richard Caley, and Rob Clark. The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival/, 2000.

[CG98] Eric Cosatto and Hans Peter Graf. Sample-based synthesis of photo-realistic talking heads. In *Computer Animation '98, Philadelphia, University of Pennsylvania*, pages 103–110. IEEE Computer Society, June 1998.

[Che71] Herman Chernoff. The use of faces to represent points in n-dimensional space graphically. Technical Report Technical Report 71, Project NR-042-993, Office of Naval Research, 1971.

[Che73] Herman Chernoff. The use of faces to represent points in k-diomensional space and graphically. *Journal of the American Statistical Association*, 68(342):361–368, June 1973.

[Cho89] Nicole Chovil. *Communicative Functions of Facial Displays in Conversation*. PhD thesis, University of Victoria, 1989.

[CM93] Michael Cohen and Dominic Massaro. Modeling coarticulation in synthetic visual speech. In Nadia Magnenat-Thalmann and Daniel Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, Tokyo, 1993.

[CPB+94] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Bechet, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: Rule–based generation of facial expression gesture and spoken intonation for multiple converstaional agents. In Andrew Glassner, editor, *Proceedings of SIGGRAPH '94 (Orlando, Florida, July 24–29, 1994)*, Computer Graphics Proceedings, Annual Co, pages 413–420. ACM SIGGRAPH, ACM Press, July 1994.

[CR75] Herman Chernoff and M. Haseeb Rizvi. Effect of classification error of random permutations of features in representing multivariate data by faces. *Journal of the American Statistical Association*, 70(351):548–554, September 1975.

[CS96] Devi Chandramohan and Peter L. Silsbee. A multiple deformable template approach for visual speech recognition. In *Proceedings of ICSLP 96 the 4th International Conference on Spoken Language Processing, October 3-6*, pages pp 50–53, Philadelphia, PA, October 1996.

[deG89] Brad deGraf. 'performance' facial animation. *SIGGRAPH '89 Course Notes 22: State of the Art in Facial Animation*, pages 8–17, July 1989.

[deG90] Brad deGraf. 'performance' facial animation. *SIGGRAPH '90 Course Notes 26: State of the Art in Facial Animation*, pages 9–20, August 1990.

[Den88] Xiao Qi Deng. *A Finite Element Analysis of Surgery of the Human Facial Tissues*. PhD thesis, Columbia University, 1988.

[DGZ+96] J. W. DeVocht, V. K. Goel, D. L. Zeitler, D. Lew, and E. A. Hoffman. Development of a finite element model to simulate and study the biomechanics of the temporomandibular joint. In *The Visible Human Project Conference, Bethesda*, pages 89–90. National Institutes of Health, October 7-8 1996.

[DiP89] Steve DiPaola. Implementation and use of a 3d parameterized facial modeling and animation system. *SIGGRAPH '89 Course Notes 22: State of the Art in Facial Animation*, pages 18–33, July 1989.

[DiP91] Steve DiPaola. Extending the range of facial types. *Journal of Visualization and Computer Animation*, 2(4):129–131, 1991.

[DJ77] Donald Dew and Paul J. Jensen. *Phonetic Processing: The Dynamics of Speech*. Charles E. Merrill Publishing Company, Columbus, Ohio, 1977.

[DKT98] Tony DeRose, Michael Kass, and Tien Truong. Subdivision surfaces in character animation. *Proceedings of SIGGRAPH 98*, pages 85–94, July 1998.

[DM70] David Ross Dickson and Wilma M. Maue. *Human Vocal Anatomy*. Charles C Thomas, Springfield, Illinois, 1970.

[DMS98] Douglas DeCarlo, Dimitris Metaxas, and Matthew Stone. An anthropometric face model using variational techniques. In *Proceedings of SIGGRAPH 98*, pages 67–74, July 1998.

[DWD+83] Parvati Dev, Sally Wood, James P. Duncan, David N. White, and Stuart W. Young. An interactive graphics system for planning reconstructive surgery. In *Proceedings National Computer Graphics Association*, pages 130–135, Chicago, Ill., June 1983.

[EF78a] Paul Ekman and Wallace Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Inc, Palo Alto, CA, 1978.

[EF78b] Norman P. Erber and Carol Lee De Filippo. Voice/mouth synthesis and tactual/visual perception of /pa, ba, ma/. *Journal of the Acoustical Society of America*, 64:1015–1019, 1978.

[EF84] Paul Ekman and Wallace V. Friesen. *Unmasking the Face*. Consulting Psychologists Press, Palo Alto, CA, 1984.

[EG98] Peter Eisert and Bernd Girod. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics and Animation*, 18(5):70–78, September/October 1998.

[Ekm73] Paul Ekman. Darwin and cross cultural studies of facial expression. In Paul Ekman, editor, *Darwin and Facial Expression: A Century of Research in Review*. Academic Press, New York, 1973.

[Ekm85] Paul Ekman. *Telling lies: clues to deceit in the marketplace, politics, and marriage*. Norton, New York, 1985.

[Els90a] Matt Elson. Displacement animation: Development and application. *SIGGRAPH 90 Course Notes, Course 10, Character Animation by Computer*, pages 15–36, 1990.

[Els90b] Matt Elson. Displacement facial animation techniques. *SIGGRAPH 90 Course Notes, Course 26, State of the Art in Facial Animation*, pages 21–42, 1990.

[EP96] Tony Ezzat and Tomaso Poggio. Facial analysis and synthesis using image-based models. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, Vermont*, October 1996.

[EP97] Tony Ezzat and Tomaso Poggio. Videorealistic talking faces: A morphing approach. In *Proceedings of the Audiovisual Speech Processing Workshop, Rhodes, Greece*, September 1997.

[EP98] Tony Ezzat and Tomaso Poggio. Miketalk: A talking facial display based on morphing visemes. In *Computer Animation '98, Philadelphia, University of Pennsylvania*, pages 96–102. IEEE Computer Society, June 1998.

[EPMT98] Marc Escher, Igor Pandzic, and Nadia Magnenat-Thalmann. Facial deformations for MPEG-4. In *Computer Animation '98, Philadelphia, University of Pennsylvania*, pages 56–63. IEEE Computer Society, June 1998.

[Erb77] Norman P. Erber. Real-time synthesis of optical lip patterns from vowel sounds. *Journal of the Acoustical Society of America*, 62:S76, 1977.

[Eri89] Joan Good Erickson. *Speech reading : an aid to communication*. Interstate Printers & Publishers, Danville, Ill., 1989.

[ESF79] Norman P. Erber, Richard L. Sachs, and Carol Lee De Filippo. Optical synthesis of articulatory images for lipreading evaluation and instruction. In D. L. McPhearson, editor, *Advances in Prosthetic Devices for the Deaf: A Technical Workshop*, pages 228–231, Rochester, NY: NTID, 1979.

[Ess94] Irfan A. Essa. *Analysis, Interpretation and Synthesis of Facial Expressions*. PhD thesis, MIT Media Lab, 1994.

[Fai90] Gary Faigin. *The Artist's Complete Guide to Facial Expression*. Watson-Guptill Publications, New York, 1990.

[FB88] David R. Forsey and Richard Bartels. Hierarchical b-spline refinement. In *Computer Graphics (SIGGRAPH '88 Proceedings)*, volume 22, pages 205–212, August 1988.

[FD99] Bill Fleming and Darris Dobbs. *Animating Facial Features and Expressions*. Charles River Media, Inc., Rockland, Massachusetts, 1999.

[FL78] O. Fujimura and J. Lovins. Syllables as concatenative phonetic units. In Alan Bell and Joan Bybee Hooper, editors, *Syllables and Segments*. North Holland, Amsterdam, 1978.

[Fox00] Anthony Fox. *Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals*. Oxford University Press, Oxford, 2000.

[FR81] Bernhard Flury and Hans Riedwyl. Graphics representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association*, 76(376):757–765, December 1981.

[Fro64] Victoria Fromkin. Lip positions in american english vowels. *Language and Speech*, 7:215–225, 1964.

[Fuj92] Osamu Fujimura. Phonology and phonetics – a syllable-based model of articulatory organization. *J. Acoust. Soc. Japan (E)*, 13, 1992.

[Fuj00] Osamu Fujimura. The C/D model and prosodic control of articulatory behavior. *Phonetica*, 57, 2000.

[Gas93] Marie-Paule Gascuel. An implicit formulation for precise contact modeling between flexible solids. In *SIGGRAPH '93*, pages 313–320, August 1993.

[GGW+98] Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Fredrick Pighin. Making faces. In *SIGGRAPH '98*, August 1998.

[Gil74] ML Gillenson. *The Interactive Generation of Facial Images on a CRT Using a Heuristic Strategy*. PhD thesis, The Ohio State University, 1974.

[GM92] Thierry Guiard-Marigny. Animation en temps réel d'un modèle paramétrique de lèvres. Master's thesis, Mémoire de D.E.A. Signal–Image–Parole, Institut National Polytechnique, Grenoble, France, 1992.

[GMAB94] Thierry Guiard-Marigny, Ali Adjoudani, and Christian Benoit. A 3-d model of the lips for visual speech synthesis. *Proceedings of the Second ESCA/IEEE Workshop*, September 1994.

[GMAB97] Thierry Guiard-Marigny, Ali Adjoudani, and Christian Benoit. 3d models of the lips and jaw for visual speech synthesis. In Jan P. Van Santen, Richard Sproat, Joseph P. Olive, and Julia Hirshberg, editors, *Progress in Speech Synthesis*, chapter 19, pages 247–258. Springer-Verlag, New York, 1997.

[GMOB95] Thiery Guiard-Marigny, David J. Ostry, and Christian Benoit. Speech intelligibility of synthetic lips and jaw. In *13th International Congress of Phonetic Sciences*, pages 222–225, Stockholm, Sweeden, August 1995.

[GMTA+96] Thierry Guiard-Marigny, Nicolas Tsingos, Ali Adjoudani, Christian Benoit, and Marie-Paule Gascuel. 3d models of the lips for realistic speech animation. In *Computer Graphic '96*, Geneve, 1996.

[Gol91] Ronald Goldman. Area of planar polygons and volume of polyhedra. *Graphics Gems II*, pages 170–171, 1991.

[Gou71] H. Gouraud. Computer display of curved surfaces. Technical Report UTEC-CSc-71-113, Dept of Computer Science, University of Utah, Salt Lake City, Utah, June 1971.

[Gou89] Roger L. Gould. Making 3-d computer character animation a great future of unlimited possibility or just tedious? *SIGGRAPH '89 Course Notes 4: 3-D Character Animation*, pages 31–63, 1989.

[Gra77] Henry Gray. *Anatomy, Descriptive and Surgical*. Gramercy Books, New York, edited by t. pikering pick and robert howden edition, 1977.

[Gro00] Moving Picture Experts Group. MPEG home page. http://www.cselt.it/mpeg/, 2000.

[Gue89a] Brian Guenter. *A computer system for simulating human facial expression*. PhD thesis, The Ohio State University, 1989.

[Gue89b] Brian Guenter. A system for simulating human facial expression. In Nadia Magnenat-Thalmann and Daniel Thalmann, editors, *State of the Art in Computer Animation*, pages 191–202. Springer-Verlag, 1989.

[HFG94] Michael Hoch, Georg Fleischmann, and Bernd Girod. Modeling and animation of facial expressions based on b-splines. *The Visual Computer*, 11(2):87–95, 1994.

[HK93] Pat Hanrahan and Wolfgang Krueger. Reflection from layered surfaces due to subsurface scattering. *Proceedings of SIGGRAPH 93*, pages 165–174, August 1993.

[HO79] Ernest Hinton and David Roger Jones Owen. *An introduction to finite element computations*. Pineridge Press, Swansea [Wales], 1979.

[Hof93] Elizabeth E. Hofer. A sculpting based solution for three-dimensional computer character facial animation. Master's thesis, The Ohio State University, 1993.

[HPW88] David R. Hill, Andrew Pearce, and Brian M. Wyvill. Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer*, 3(5):277–289, March 1988.

[HSGR83] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983.

[IK91] Satoshi Ishibashi and Fumio Kishino. Color/texture analysis and synthesis for model-based human image coding. *SPIE Visual Communications and Image Processing '91: Visual Communication*, 1605:242–252, 1991.

[ISO99a] ISO/IEC 14496-1:1999 Information technology – Coding of audio-visual objects – part 1: Systems. International Organization for Standardization/International Electrotechnical Commission, 1999.

[ISO99b] ISO/IEC 14496-2:1999 Information technology – Coding of audio-visual objects – part 2: Visual. International Organization for Standardization/International Electrotechnical Commission, 1999.

[JB71] Janet Jeffers and Margaret Barley. *Speechreading (lipreading)*. Thomas, Springfield, 1971.

[KAA83] Arthur I. Karshmer, Daniel Allan, and Dolores Anderson. 'Virtual' craniofacial surgery using interactive computer graphics: A pilot

study. In *Proceedings National Computer Graphics Association*, pages 125–129, Chicago, Illinois, June 1983.

[Kan77] Takeo Kanade. *Computer Recognition of human faces.* Birkhauser Verlag, Basel and Stuttgart, 1977.

[KBG85] Harriet Kaplan, Scott J. Bally, and Carol Garretson. *Speechreading : a way to improve understanding.* Gallaudet College Press, Washington, D.C., 1985.

[KGB98] Rolf M. Koch, Markus H. Gross, and Albert A. Bosshard. Emotion editing using finite elements. Technical Report 281, Computer Science Department, ETH Zrich, 1998.

[KGC+96] R. M. Koch, Markus H. Gross, F. R. Carls, D. F. von Buren, G. Fankhauser, and Y. I. Parish. Simulating facial surgery using finite element models. In Holly Rushmeier, editor, *Proceedings of SIGGRAPH '96 (New Orleans, Louisiana, August 4–9, 1996)*, Computer Graphics Proceedings, Annual Co, pages 421–428. ACM SIGGRAPH, Addison Wesley, August 1996.

[KGPG96] Erwin Keeve, Sabine Girod, Paula Pfeifle, and Bernd Girod. Anatomy-based facial tissue modeling using the finite element method. In *IEEE Visualization '96.* IEEE, October 1996.

[KI86] Wilton Marion Krogman and Mehmet Yasar Iscan. *The human skeleton in forensic medicine.* C.C. Thomas, Springfield, 1986.

[Kin93] William Joseph King. Holistic model of the face in human computer interaction. Technical report, Southwestern University, 1993.

[Kle89] Jeff Kleiser. A fast, efficient, accurate way to represent the human face. *SIGGRAPH '89 Course Notes 22: State of the Art in Facial Animation*, pages 36–40, July 1989.

[KM77] R. D. Kent and F. D. Minifie. Coarticulation in recent speech production models. *Journal of Phonetics*, 5:115–135, 1977.

[KMCT96] T. Konno, H. Mitani, H. Chiyokura, and I. Tanaka. Surgical simulation of facial paralysis. In Hans B. Sieburg, Suzanne J Weghorst, and Karen S Morgan, editors, *Medicine Meets Virtual Reality: Health Care in the Information Age*, pages 488–497. IOS Press, January 17–20 1996.

[KP00] Scott A. King and Richard E. Parent. Talkinghead: A text-to-audiovisual-speech system. Technical Report OSU-CISRC-2/80-TR05, Computer and Information Science, Ohio State University, 2000.

[Lar86a] Wayne F. Larrabee Jr. A finite element model of skin deformation. *Laryngoscope*, pages 399–419, April 1986.

[Lar86b] Wayne F. Larrabee Jr. A finite element model of skin deformation: Part i - biomechanics of skin. *Laryngoscope*, pages 399–419, April 1986.

[Lar86c] Wayne F. Larrabee Jr. A finite element model of skin deformation: Part ii - experimental model of skin deformation. *Laryngoscope*, pages 399–419, April 1986.

[Lar86d] Wayne F. Larrabee Jr. A finite element model of skin deformation: Part iii - the finite element model. *Laryngoscope*, pages 399–419, April 1986.

[Lee93] Yuencheng Victor Lee. The construction and animation of functional facial models from cylindrical range/reflectance data. Master's thesis, University of Toronto, 1993.

[Lew91] John Lewis. Automated lip-sync: Background and techniques. *The Journal of Visualization and Computer Animation*, 2(4):118–122, 1991.

[Löf90] A. Löfqvist. Speech as audible gestures. In W. J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modeling*, pages 289–322. Kluwer Academic Publishers, Dordrecth, 1990.

[LP87] J. P. Lewis and Frederic I. Parke. Automated lip-synch and speech synthesis for character animation. In J. M. Carroll and P. P. Tanner, editors, *Proceedings Human Factors in Computing Systems and Graphics Interface '87*, pages 143–147, April 1987.

[LTW93] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Constructing physics-based facial models of individuals. In *Proceedings of Graphics Interface '93, Toronto, Ontario, Canada*, pages 1–8. Canadian Information Processing Society, May 1993.

[LTW95] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Realistic modeling for facial animation. *Proceedings of SIGGRAPH 95*, pages 55–62, August 1995.

[MA96] I. R. Murray and J. L. Arnott. Synthesizing emotions in speech: is it time to get excited? In *Proceedings of ICSLP 96 the 4th International Conference on Spoken Language Processing, Octover 3-6*, pages pp 1816–1819, Philadelphia, PA, October 1996.

[Mad69] Roy P. Madsen. *Animated Film: Concepts, Methods, Uses*. Interland Pub., New York, 1969.

[Mae90] Shinji Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In William J. Hardcastle and Alain Marchal, editors, *Speech Production and Speech Modelling*, pages 131–149. Kluwer Academic Publishers, Dordrecht, 1990.

[MAH90] Shigeo Morishima, Kiyoharu Aizawa, and Hiroshi Harashima. A real-time facial action image synthesis system driven by speech and text. *SPIE Visual Communications and Image Processing*, 1360:1151–1157, October 1990.

[MBR99] The MBROLA project. www.tcts.fpms.ac.be/synthesis/mbrola/, 1999.

[MC83] Dominic W. Massaro and Michael M. Cohen. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology*, 9(5):753–771, 1983.

[McG85] Matthew McGrath. *An examination of cues for visual and audio-visual speech perceptions using natural and computer-generated faces*. PhD thesis, Univ of Nottingham, UK, November 1985.

[Mil88] Gavin S. Miller. From wire-frames to furry animals. In *Proceedings of Graphics Interface '88*, pages 138–145, June 1988.

[MM76] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[Mon80] A. A. Montgomery. Development of a model for generating synthetic animated lip shapes. *Journal of the Acoustical Society of America*, 68:S58(A), 1980.

[MP95] Baback Moghaddam and Alex P. Pentland. An automatic system for model-based coding of faces. In *Proceedings of Data Compression Conference '95*, pages 362–370, 1995.

[MTMdAT89] N. Magnenat-Thalmann, H. T. Minh, M. de Angelis, and D. Thalmann. Design, transformation and animation of human faces. *The Visual Computer*, 5(1/2):32–39, March 1989.

[MTPT88] N. Magnenat-Thalmann, E. Primeau, and D. Thalmann. Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297, March 1988.

[MTT87] Nadia Magnenat-Thalmann and Daniel Thalmann, editors. *Synthetic Actors in Computer-Generated 3D Films*. Springer-Verlag, Tokyo, NY Paris, 1987.

[Mur91] Shigeru Muraki. Volumetric shape description of range data using bby model". In Thomas W. Sederberg, editor, *Computer Graphics (SIGGRAPH '91 Proceedings)*, volume 25, pages 227–235, July 1991.

[MVBT95] Kevin G. Munhall, Eric Vatikoitis-Bateson, and Yoh'ichi Tohkura. X-ray film database for speech research. *Journal of the Acoustical Society of America.*, 98:1222–1224, 1995.

[NHRD90] Monique Nahas, Herve Huitric, Marc Rious, and Jacques Domey. Registered 3D-texture imaging. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '90 (Second workshop on Computer Animation)*, pages 81–91. Springer-Verlag, April 1990.

[NHS88] Monique Nahas, Herve Huitric, and Michel Saintourens. Animation of a B-spline figure. *The Visual Computer*, 3(5):272–276, March 1988.

[NLH99] Visible human project. National Library of Medicine, http://www.nlm.nih.gov/research/visible/visible_human.html, 1999.

[NNT30] Edward Bartlett Nitchie, Elizabeth Helm Nitchie, and Gertrude Torrey. *Lip-reading Principles and Practise*. Lippincott, Philadelphia, 1930.

[NS62] Ramond J. Nagle and Victor H. Sears. *Denture Prosthetics Complete Dentures*. The C. V Mosby Company, Saint Louis, 1962.

[OH77] David Roger Jones Owen and Ernest Hinton. *Finite Element Programming*. Academic press, London; New York, 1977.

[OH80a] David Roger Jones Owen and Ernest Hinton. *A simple guide to finite elements*. Pineridge Press, Swansea [Wales], 1980.

[OH80b] D.R.J. Owen and E. Hinton. *Finite elements in plasticity : theory and practice*. Pineridge Press, Swansea, U.K., 1980.

[Ost98] Jorn Ostermann. Animation of synthetic faces in mpeg-4. In *Computer Animation '98, Philadelphia, University of Pennsylvania*, pages 49–55. IEEE Computer Society, June 1998.

[OTK88] Kazunori Ohmura, Akira Tomono, and Yukio Kobayashi. Method of detecting face direction using image processing for human interface. In *Proceedings SPIE Visual Communication and Image Processing '88*, pages 625–632, November 1988.

[PALS97] Frederic Pighin, Joel Auslander, Dani Lischinski, and David Salesin. Realistic facial animation using image-based 3d morphing. Technical Report UW-CSE-97-01-03, University of Washington, Department of Computer Sceince & Engineering, 1997.

[Par72] Frederic I. Parke. Computer generated animation of faces. In *Proceedings ACM National Conference*, pages 451–457, 1972.

[Par74] Frederic I. Parke. *A parametric model for human faces*. PhD thesis, University of Utah, Salt Lake City, Utah, December 1974.

[Par75] Frederic I. Parke. A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, 1(1):1–4, 1975.

[Par82] Frederic I. Parke. Parameterized models for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–68, November 1982.

[Pat91] Majula Patel. *Making Faces: The Facial Animation, Construction and Editing System*. PhD thesis, University of Bath, Bath, UK, December 1991.

[PB95] J A Provine and L T Bruton. Lip synchronization in 3-d model based coding for video-conferencing. In *Proc. of the IEEE Int. Symposium on Circuits and Systems*, pages 453–456, Seattle, May 1995.

[PBS91] Catherine Pelachaud, Norman I. Badler, and Mark Steedman. Linguistic issues in facial animation. In Nadia Magnenat-Thalmann and Daniel Thalmann, editors, *Computer Animation '91*, pages 15–30. Springer-Verlag, Tokyo, 1991.

[PBV94] Catherine Pelachaud, Norman I. Badler, and Marie-Luce Viaud. Final report to nsf of the standards for facial animation workshop. Technical report, University of Pennsylvania, October 1994.

[Pel91] Catherine Pelachaud. *Communication and Coarticulation In Facial Animation*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 19104-6389, October 1991.

[Per69] Joseph S. Perkell. *Physiology of speech production: results and implications of a quantitative cineradiographic study*. M. I. T. Press, Cambridge, Massachusetts, 1969.

[PH87] J. Pierrehumbert and J. Hischberg. The meaning of intonational contours in the interpretation of discourse. Technical report, AT&T Bell Laboratories, 1987.

[PH89] K Perlin and E Hoffert. Hypertexture. In *Computer Graphics (SIGGRAPH '89 Proceedings)*, volume 23, pages 253–262, July 1989.

[PHL+98] Frederic Pighin, Jamie Hecker, Dani Lischinski, Rich ard Szeliski, and David H Salesin. Synthesizing realistic facial expressions from photographs. In *Computer Graphics Proccedings SIGGRAPH '98*, pages 75–84, 1998.

[Pie89] Steve Pieper. Physically-based animation of facial tissue for surgical simulation. *SIGGRAPH '89 Course Notes 22: State of the Art in Facial Animation*, pages 71–124, July 1989.

[Pla85] Stephen M. Platt. *A Structural Model of the Human Face*. PhD thesis, University of Pennsylvania, Philadelphia, Pennsylvania, 1985.

[PLG91] Elizabeth C. Patterson, Peter C. Litwinowicz, and Ned Greene. Facial animation by spatial mapping. In Nadia Magnenat-Thalmann and Daniel Thalmann, editors, *Computer Animation '91*. Springer-Verlag, Tokyo, 1991.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing (2nd ed.)*. Cambridge University Press, Cambridge, 1992.

[PvOS94] Catherine Pelachaud, C. van Overveld, and Chin Seah. Modeling and animating the human tongue during speech production. In *Proceedings of Computer Animation '94*, pages 40–49, May 1994.

[PVY93] Catherine Pelachaud, Marie-Luce Viaud, and H. Yahia. Rule-structured facial animation system. *IJCAI*, August 1993.

[PW91] Manjula Patel and Philip J. Willis. FACES — facial animation, construction and editing system. In Werner Purgathofer, editor, *Eurographics '91*, pages 33–45. North-Holland, September 1991.

[PW96] Frederic I. Parke and Keith Waters. *Computer Facial Animation*. A K Peters, Wellesley, Massachusetts, 1996.

[PWWH86] Andrew Pearce, Brian M. Wyvill, Geoff Wyvill, and David Hill. Speech and expression: A computer solution to face animation. In M. Green, editor, *Proceedings of Graphics Interface '86*, pages 136–140, May 1986.

[RC80] J. S. Rhine and H. R. Campbell. Thickness of facial tissues in american blacks. *Journal of Forensic Sciences*, 25(4):847–858, 1980.

[RCI91] R. E. Rosenblum, Wayne E. Carlson, and E. Tripp III. Simulating the structure and dynamics of human hair: modelling, rendering and animation. *Journal of Visualization and Computer Animation*, 2(4):141–148, 1991.

[Red91] Subhash Chandra Reddy. *Model based analysis of facial expressions*. PhD thesis, University of Illinois at Urbana-Champaign, November 1991.

[Ree90] William T. Reeves. Simple and complex facial animation: Case studies. *SIGGRAPH '90 Course Notes 26: State of the Art in Facial Animation*, pages 88–106, August 1990.

[RGTC98] S. H. Roth, Markus Hans Gross, Silvio Turello, and Friedrich Robert Carls. A bernstein-bzier based approach to soft tissue simulation. Technical Report 282, Computer Science Department, ETH Zrich, 1998.

[RI82] J. S. Rhine and C. E. Moore II. Facial reproduction: Tables of facial tissue thicknesses of american caucasoids in forensic anthropology. Technical Report 1, Maxwell Museum Technical Series, Albuquerque, NM, 1982.

[RvBSvdL95] M.J.T. Reinders, P. J.L. van Beek, B. Sankur, and J. C. A. van der Lubbe. Facial feature localization and adaptation of a generic face model for model-based coding. *Signal Processing: Image Communication*, 7:57–74, 1995.

[RvdGG96] Marcel J. Reinders, M. J. van der Goot, and J. J. Gerbrands. Estimating facial expressions by reasoning. In *Proceedings of the ASCI'96 Conference*, pages 259–264, 1996.

[Ryd87] Mikael Rydfalk. Candide: A parameterized face. Technical Report LiTH-ISY-I-0866, Linkoping University, Sweden, October 1987.

[SC90] A. D. Simons and S. J. Cox. Generation of mouthshapes for a synthetic talking head. *Proceedings of the Institute of Acoustics (Autumn Conference)*, 12(10):483–490, 1990.

[Seg84] Larry J. Segerlind. *Applied Finite Element Analysis 2nd Edition*. John Wiley & Sons, New York, 1984.

[SF93] Richard Sproat and Osamu Fujimura. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, 21, 1993.

[She96a] Jonathan Richard Shewchuk. Triangle: A two-dimensional quality mesh generator. http://www.cs.cmu.edu/∼quake/triangle.html, 1996. Accessed Sep 15, 1999.

[She96b] Jonathan Richard Shewchuk. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In Ming C. Lin and Dinesh Manocha, editors, *Applied Computational Geometry: Towards Geometric Engineering*, volume 1148 of *Lecture Notes in Computer Science*, pages 203–222. Springer-Verlag, May 1996.

[SK73] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces,. *Journal of the Optical Society of America*, 4(3):519–524, March 1973.

[SL96] Marueen Stone and Andrew Lundberg. Three-dimensional tongue surface shapes of english consonants and vowels. *J. Acoust. Soc. Am*, 99(6):3728–3737, June 1996.

[SMC00] J Paul Siebert, Roy L Middleton, and W Paul Cockshott. Michelangelo project. http://www.faraday.gla.ac.uk/michelangelo.htm, 2000.

[SMP81] Norman I. Badler Stephan M. Platt. Animating facial expressions. *Computer Graphics (Proceedings of SIGGRAPH 81)*, 15(3):245–252, August 1981.

[SNK72] T. Sakai, M. Nagao, and T. Kanade. Computer analysis and classification of photographs of human faces. In *Proceedings of the first USA-Japan Computer Conference*, pages 55–62, 1972.

[SPS96] Alexei I Sourin, Alexander A Pasko, and Vladimir V Savchenko. Using real functions with application to hair modelling. *Computers and Graphics*, 20(1):11–19, 1996.

[Sto91] Marueen Stone. Toward a model of three-dimensional tongue movement. *Journal of Phonetics*, 19:309–320, 1991.

[Stu98] David J Sturman. Computer puppetry. *Computer Graphcis and Applications*, 18(1):38–45, January/February 1998.

[SVG95] Agnes Saulnier, Marie-Luce Viaud, and David Geldreich. Real-time facial analysis and synthesis chain. In *International Workshop on Automatic Face - and Gesture - Recognition*, Zurich, Switzerland, June 1995.

[Tao97] Hai Tao. Facial modeling. Technical report, University of Illinois, March 1997.

[TN93] Akikazu Takeuchi and Katashi Nagao. Communicative facial displays as a new conversational modality. In *ACM/IFIP INTERCHI '93*, Amsterdam, 1993.

[TO99] A. Murat Tekalp and Jörn Ostermann. Face and 2-d mesh animation in mpeg-4. *Signal Processing: Image Communication*, 1999.

[TPBF87] Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. Elastically deformable models. *Computer Graphics (SIGGRAPH '87 Proceedings)*, 21(4):205–214, July 1987.

[TST87] Yosuke Takashima, Hideo Shimazu, and Masahiro Tomono. Story driven animation. In J. M. Carroll and P. P. Tanner, editors, *Proceedings of Human Factors in Computing Systems and Graphics Interface '87*, pages 149–153, April 1987.

[TW90] Demetri Terzopoulos and Keith Waters. Physically-based facial modelling, analysis, and animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, March 1990.

[UGO98] Baris Uz, Ugur Gudukbay, and Bulent Ozguc. Realistic speech animation of synthetic faces. In *Computer Animation '98, Philadelphia, University of Pennsylvania*, pages 111–118. IEEE Computer Society, June 1998.

[vB84] Renee van Bezooyen. *Characteristics and Recognizability of Vocal Expressions of Emotion*. Netherlands Phonetic Archives. Foris Publications, Dordrecht, Holland, 1984.

[VY92] Marie-Luce Viaud and Hussein Yahia. Facial animation with wrinkles. In *Eurographics 92*, Cambridge, United Kingdom, September 1992.

[Wai89] Clea Theresa Waite. The facial action control editor, face: A parametric facial expression editor for computer generated animation. Master's thesis, Massachusetts Institue of Technology, February 1989.

[Wal82] Edward F. Walther. *Lipreading*. Nelson-Hall, Chicago, 1982.

[Wat87] Keith Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics (SIGGRAPH '87 Proceedings)*, 21(4):17–24, July 1987.

[Wat89] Keith Waters. Modeling 3d facial expressions: Tutorial notes. *SIGGRAPH '89 Course Notes 22: State of the Art in Facial Animation*, pages 127–152, July 1989.

[Wat92] Keith Waters. A physical model of facial tissue and muscle articulation derived from computer tomography data. In *SPIE Visualization in Biomedical Computing*, volume 1808, pages 574–583, 1992.

[Wer94] Josie Wernecke. *The Inventor Mentor*. Addison-Wesley, 1994.

[WF94] Carol L. Wang and David R. Forsey. Langwidere: A new facial animation system. In *Proceedings of Computer Animation '94*, pages 59–68, 1994.

[WH89] Brian M. Wyvill and David R. Hill. Expression control using synthetic speech. *SIGGRAPH '89 Course Notes 22:State of the Art in Facial Animation*, pages 163–175, 1989.

[WH90] Brian M. Wyvill and David R. Hill. Expression control using synthetic speech. *SIGGRAPH '90 Course Notes 22:State of the Art in Facial Animation*, pages 185–212, 1990.

[Wil90a] L Williams. Performance-driven facial animation. In *Computer Graphics (SIGGRAPH '90 Proceedings)*, volume 24, pages 235–242, August 1990.

[Wil90b] Lance Williams. Electronic mask technology. *SIGGRAPH '90 Course Notes 26: State of the Art in Facial Anim ation*, pages 165–184, August 1990.

[WKMT96] Yin Wu, Prem Kalra, and Nadia Magnenat-Thalmann. Simulation of static and dynamic wrinkles of skin. In *Proceedings Computer Animation 96*, pages 90–97, 1996.

[WL93] Keith Waters and Thomas M. Levergood. Decface: An automatic lip-synchronization algorithm for synthetic faces. Technical Report CRL 93/4, Digital Equipment Corporation Cambridge Research Lab, September 1993.

[WMTT95] Yin Wu, Nadia Magnenat-Thalmann, and Daniel Thalmann. A dynamic wrinkle model in facial animation and skin ageing. *JVCA*, 6:195–205, 1995.

[WS89] Y. Watanabe and Y. Suenaga. Drawing human hair using wisp model. In *Proceedings Computer Graphics International '89*, pages 691–700, 1989.

[WT91] Keith Waters and Demetri Terzopoulos. Modeling and animating faces using scanned data. *The Journal of Visualization and Computer Animation*, 2(4):123–128, 1991.

[WT95] Reiner Wilhelms-Tricarico. Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *J. Acoust. Soc. Am.*, 97(5):3085–3098, May 1995.

[WTP97] Reiner F. Wilhelms-Tricarico and Joseph S. Perkell. Biomechanical and physiological based speech modeling. In Jon P.H. Von Santen et al, editor, *Progress in Speech Synthesis*, pages 221–233. Springer, 1997.

[XANK89] G. Xu, H. Agawa, Y. Nagashima, and Y. Kobayashi. A stereo based approach to face modeling for the ATR virtual space conferencing system. In *Proc. SPIE Conference on Visual Communication and Image Processing*, pages 365–379, 1989.

[YCH88] A. Yuille, D. Cohen, and P. Hallinan. Facial feature extraction by deformable templates. Technical Report 88-2, Harvard Robotics Laboratory, 1988.

[YD88a] John F. Yau and Neil D. Duffy. 3D facial animation using image samples. In N. Magnenat-Thalmann and D. Thalmann, editors, *New Trends in Computer Graphics (Proceedings of CG International '88)*, pages 64–73. Springer-Verlag, 1988.

[YD88b] John F. Yau and Neil D. Duffy. A texture mapping approach to 3-D facial image synthesis. *Computer Graphics Forum*, 7(2):129–134, June 1988.