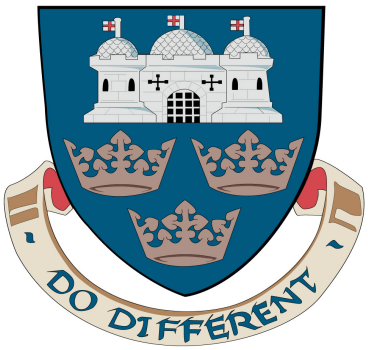
Expressive Modulation of Neutral Visual Speech

Felix Shaw

PhD Thesis

University of East Anglia
School of Computing Sciences



Tuesday 29th September, 2015

For Rosalie

Contents

List of Figures

"Give me a Leonard Cohen afterworld, so I can sigh eternally."

— Kurt Cobain, *Pennyroyal Tea*

"There is no landscape that we know as well as the human face. The twenty-five-odd square inches containing the features is the most intimately scrutinized piece of territory in existence, examined constantly, and carefully, with far more than an intellectual interest. Every detail of the nose, eyes, and mouth, every regularity in proportion, every variation from one individual to the next, are matters about which we are all authorities."

— Gary Faigin, *The Artist's Complete Guide to Facial Expression*

Chapter 1

Acknowledgements

I would like to thank my supervisor Dr. Barry Theobald for his wisdom, patience and guidance. I have yet to see Barry's intuition proved wrong. I would further like to thank the other members of faculty in the Speech group at UEA who helped me, namely Prof. Stephen Cox and Dr. Richard Harvey as well as Dr. Hui Yu for being my examiner. Thanks also to Dr. Geoff McKeown for keeping me gainfully employed teaching undergraduate students how to program, and to Dr. Tony Bagnall for thinking of me when the funding opportunity arose. Thanks are due to all the great friends I made during my time at UEA, in particular Dom Howell, Jason Lines, John Taylor, Jake Newman and Luke Davies. Thanks also to my friends outside of academic life who kept me sane, particularly to Phin and my second family the van Rijbroek/Carneys. The darkest day can be made bright if shared with friends. Special thanks of course to my parents Helen and Peter, my sister Genevieve, my aunt Barbara and uncle Tom, for their emotional support over the last few years. Finally thanks to my beautiful partner Jemma for her patience, care and humour. You looked after Rosa for so many evenings and weekends on your own whilst I was working and I couldn't have done this without you.

Chapter 2

Introduction

Non-verbal communication such as body language, paralanguage (the use of non-verbal voice communication) and facial expression are probably the oldest form of human interaction inherited from our animal ancestors. Indeed these are the only forms of communication understood by infants for the first few months of their lives, with even very young babies being able to encode and decode the fundamental emotional meaning of a smile or a raised voice. Therefore non-verbal communication represents an import aspect of language, the absence of which often leads to confusion. As an example in our modern digital lives, the so-called "emoticons" (textual representations of emotion) have become digital vernacular in an attempt to alleviate the problem of mis-interpreting written text in emails, text messages and status updates etc. We are all experts in non-verbal communication and are attuned to every nuance of facial expression and intonation of pitch. In Spanish, the intonation of the same sequence of words is used to indicate whether a sentence is declarative or interrogative. In Mandarin Chinese the intonation is used to indicate different meanings for the same words. We also use tone of voice to signify more subtle differences in meaning such as irony, sarcasm or simply

to provide emphasis to pertinent information. Likewise, facial expression can be used in a similar way and usually complements tonal intonation. The sentence "I can't wait for the World Cup" said with a groan and roll of the eyes, has a completely different meaning to the same sentence said with elevated pitch and raised eyebrows.

The production of facial expression on graphical models is useful for many applications. In the future, digital tutors may be able to recognise human emotion and adapt the way they communicate and instruct a student. Computerised customer service representatives could likewise adapt their behaviour to diffuse highly fraught situations, or simply ease the instinctive frustration and uneasiness many people have talking to computers. Psychologists could use such systems to design novel experiments into human expression, using the model to carefully control the expressive parameters or law enforcement agencies could use such a system to train operatives into how humans communicate and display expression when they are trying to deceive. Perhaps the area of greatest academic and commercial interest is animation and visual speech synthesis.

The need for animated graphical models of the human face is commonplace in the movies, video games and television, appearing in everything from low budget advertisements and free mobile apps, to Hollywood blockbusters costing hundreds of millions of dollars. Software for the production of such animation is mature with several software packages forming the backbone of most commercial animation projects. Industry standard techniques for character animation include motion capture (mocap) and handcrafted keyframe based animation. In mocap animation an actor has reflective markers placed on their body and/or face, the movement of which can be detected by an array of cameras. Using a triangulation of such cameras, the movement of the actor can be calculated in 3D space. These movements can subsequently be retargeted to a virtual model. This has the advantage

of being a relatively fast process. Models can be animated in real time which is an obvious advantage in a commercial enterprise with tight deadlines. While this technique works well for full body motion (so long as the target model is not too dissimilar in configuration to the source), it is less well suited to the capture of facial motion. The subtleties of facial movement such as those required to express emotion (sometimes a matter of a few millimetres), are often beneath the resolution of such systems, even with many hundreds of markers on the actor's face. When this relatively sparse configuration of markers *is* enough, post-processing such as the interpolation of occluded markers and smoothing, mean the performance can lack realistic dynamics and liveliness.

In hand crafted animation, an animator creates frames using an animation model for important points in a sentence, such as plosive lip shapes, or important expressive gestures like eyebrow raising. The intervening frames between these keyframes are then interpolated by the software package. In the pre-digital days of animation, this workflow was done by a lead artist drawing the key frames, and junior artists drawing the transitional frames in between. In an early attempt to computerise the process, Parke [1972] showed that it was possible to design keyframes of facial geometry, and then use a mixing coefficient to linearly interpolate between keyframes to create the intervening frames. Cosine functions were used to modulate the mixing coefficients to simulate the acceleration/deceleration dynamics observed in natural speech. Key frame techniques give the animator tight control of the fine movement of a model and can lead to spectacular results. There are however major drawbacks. Firstly the technique is hugely labour intensive. Even for simple animations, several keyframes must be produced each second for each character model. It can take a professional animator days to produce a few minutes of footage. It is easy therefore to see why it takes months or even years to animate a feature length picture. When the key frames are created

from "blend shapes" (linear combinations of pre-modelled facial poses), there is the phenomenon of blend shape interference Lewis et al. [2005a], where certain combinations of blend shapes have non-orthogonal areas of influence leading to unwanted artefacts in the linear combination and must be corrected with additional blend shapes and many iterations of fine tuning. Finally, the interpolation of intervening frames may not provide the dynamics or liveliness required therefore leading to a greater concentration of keyframes.

Generative statistical models of animation attempt to address some of these drawbacks. Using a corpus of training data, statistical models can be built and driven in various ways to produce high quality animation quickly, alleviating keyframe animation's major drawback of labour intensity. To address the problem of the capture of facial movement with mo-cap, markerless face tracking tecniques can be employed borrowing methods from computer vision, such as tracking with Active Appearance Models (AAMs), deformable surface models driven by laser scanned face data or dense stereo techniques. Some form of representational features are extracted from the corpus and used in a generative model e.g. as selection units in a concatenative synthesiser or emission features in an HMM based synthesiser. A major drawback in both these cases is the requirement of a large amount of training data which grows linearly as more expressive detail is added. The complexity of the synthesis technique itself also grows as a function of the amount of expressive detail.

Such statistical models have been the subject of much research over recent years in an attempt to alleviate some of the problems of traditional animation techniques, but as yet have found little widespread industrial acceptance. This is because the quality of lip movements produced by such systems is still sub-optimal and the emotional component is lagging behind. This work is an attempt to build a technique capable of modulating neutral visual output such as that produced

by modern visual speech synthesisers with emotional expression and therefore to create a compelling, convincing and complete visual speech synthesis solution.

Chapter 3

Statement of Originality

What follows is a list of the main features of this work which set it apart from preceding research. Taking Cao et al. [2003] as a starting point, this thesis shows the following:

- That Cao's original technique can be applied to unseen data which was never in the training set.
- That it is possible to create an ICA model from neutral visual speech and visual speech in a single expressive style. This reduces the complexity of the technique specified in Cao et al. [2003] (where a model was created for each pair of expressive styles) from n^2 models to n models (where n is the number of expressions in the training set).
- That it is possible to modulate an arbitrary amount of unseen neutral visual speech data with the expression found in the training set, thus making the technique of potential use to animators.
- That it is possible to train a mixed visual speech ICA model from neutral and multiple expressive styles. From this ICA is able to project these ex-

pressive styles into different independent components where it is possible to manipulate these components orthogonally to a speech signal, therefore taking neutral visual speech data and modulating it with any of the expressions found in the single ICA model. This further simplifies the Cao's original work from from n^2 models to 1 model.

- That the technique is capable to modulating neutral expressions with a blend of expressions found in the training set i.e. Anger mixed with Surprise. This is necessary for any emotion synthesiser since human emotion is generally ambivalent. Furthermore it allows the trained expression space to be interpolated producing novel expressions not present in the original training set.

- That the technique works for a variety of data types. Specifically, it works for point cloud data, Active Appearance Model features and rig controller activations.

Chapter 4

Data Description

The work described in this thesis uses two datasets which are discussed below. As explained in Cicconetti et al. [2009], there are three types of expressive dataset, known as posed, re-acted and interacted. Posed data is collected by filming or otherwise capturing the speech and movements of actors performing a pre-designed corpus of lines. This is useful as it gives tight control over exactly what is said, and in which expressions. However, the plausibility of the output can only be as good as the quality of the actor. Re-acted data involves showing the actor a video clip, and them trying to copy the delivery observed. This tends to lead to better output for untrained actors, but finding appropriate video clips for participants to copy can be time-consuming, and the technique limits sentence content. Interacted data is where participants are simply filmed having conversations with others. This leads to the most plausible and natural data, but is the most difficult to capture. The participant must be guided (by carefully choreographed conversations) into displaying different expressions. Therefore, only certain expressions may be ethically obtained. For example obtaining happiness and laughter is easy, however it would be unethical to elicit genuine fear in the participant. Additionally, the less

constrained nature of interacted data can lead to difficulties in processing later in the pipeline due to things like head movement, view distance and lighting.

4.1 Biwi 3D Audiovisual Corpus of Affective Communication

Some of our experiments use the Biwi 3D Audiovisual Corpus of Affective Communication, Fanelli et al. [2010a]. This re-acted corpus comprises of depth scanned data captured using a novel method allowing for 3D data capture using relatively inexpensive equipment (a digital projector and three cameras), by employing phase shifting and stereo unwrapping (see Weise et al. [2007], for a detailed description). The content of the dataset consists of 40 english sentences, each performed by eight female and six male subjects, all native english speakers. Their ages ranged from 21 to 53 (mean 33.5). Each sentence was performed in a neutral style and then in an expressive style. The expressive styles were elicited using clips from well known films e.g. Pulp Fiction and Pride & Prejudice, and were chosen so as to cover a large range of expressive styles. The performer was shown the text on a screen and asked to say it in as neutral way as possible. Then they were shown up to 30 seconds of the film leading up the sentence to be performed, so as to provide some emotional background. The movie clips could be seen multiple times. They were asked to repeat the sentence using the emotional tone they perceived from the clip. The average length of a sentence was 4.67 seconds and the data was captured at 25 fps. Since the captured 3D data was noisy, not in correspondence and contained occluded points, some cleanup was required. A generic mesh was warped to the rest position of an actor's scanned face, i.e. neutral expression, mouth closed, eyes open. Then displacements between the neutral mesh and each frame of the

Figure 4.1 An example frame rendered in Matlab with the "patch" function. A simple light provides illumination for the scene

performance were computed using geometric and texture constraints to optimise the solution. Therefore the final 3D output mesh is in geometric and temporal correspondence for every frame. The dataset itself contains the original noisy 3D scans, the cleaned up 3D data, corresponding audio as well as phonological labels and code (written in C and Matlab) for reading the cleaned up data files. An example frame from the dataset can be seen rendered using Matlab's patch function in Figure 4.1 with a standard "camlight" to show the full surface model, and as a simple 3D plot in Figure 4.2 to show the density of the mesh.

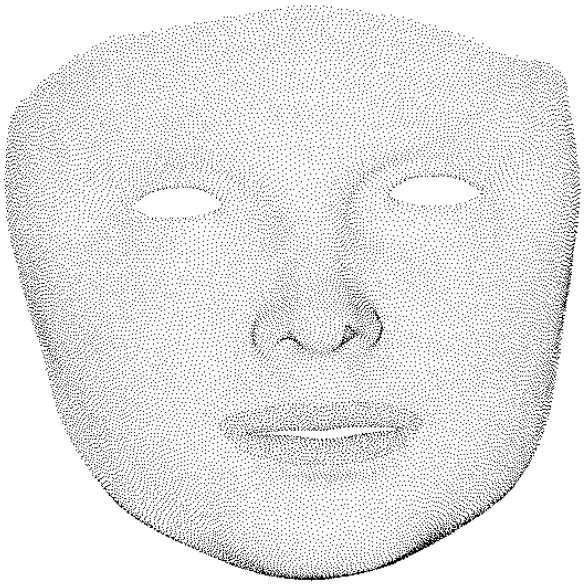


Figure 4.2 The same frame as in Figure 4.1, but this time rendered as a 3D plot to give an idea of the density of the data. The plot consists of 23370 vertices in three dimensions.

4.2 Our Custom Dataset

In addition to the dataset described in Section 4.1, we collected a data corpus. The primary purpose of this was to have the same utterance spoken in multiple styles, rather than neutral plus one additional expressive style as in the BiWi corpus. Therefore we opted to produce a posed dataset giving tight control over delivery and sentence content. A single male actor was recorded uttering 15 sentences each in Happy, Sad, Angry, Surprised and Neutral styles. Each sentence was recorded twice in each expressive style so that audio could be held out, meaning that during testing, no video was shown alongside its originally recorded audio, thus eliminating a source of bias.

The sentences where chosen so as to make as much sense in each of the required expressive styles. A back story was created for each to provide context for the

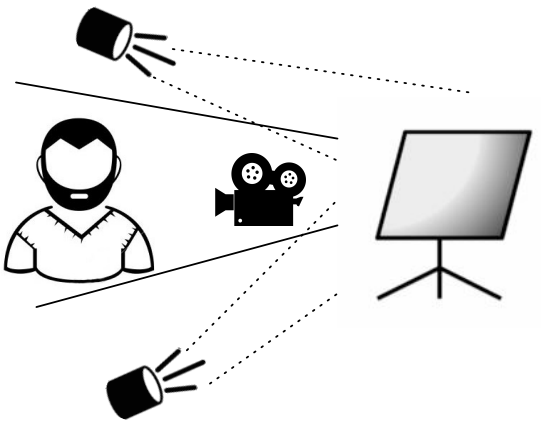


Figure 4.3 Setup of the capture session for our custom dataset.

actor to base their emotional response. The sentences and backstories are listed in appendix B. As was mentioned at the top of this chapter, a problem with posed datasets is that the output can only be as plausible as the actor(s) performing the lines. It should be noted that the actor used in this dataset was not professional therefore this should be taken into account when considering subjective opinions of system output later on in this work.

The video was recorded on a JVC GY-HM750E professional video camera with a 14x Canon lens. The frame rate was 25 fps, captured in 720 progressive scan mode. The camera captured the video in XDCAM EX format, wrapped into Quicktime mov files. Audio was collected via a clip microphone, attached to the camera's external XLR interface. Lighting was supplied by two 500W floodlights pointed at a white screen directly opposite the actor behind the camera as shown in figure 4.3. All sentences were captured in a single session to avoid any issues arising from differences in lighting, camera position, actor clothing etc. Figure 4.4 shows some example frames from the training set. Since the data capture session was

Figure 4.4 Example frames from our custom dataset.

continuous and lasted around an hour, each full video had to be manually edited into sentence long chunks. This was was done using Final Cut Pro Apple [2014]. Each sequence was then separated into its component frames using FFmpeg Bellard [2014] and then tracked with a custom AAM tracker written at the University of East Anglia Theobald [2014].

4.3 Conclusion

This section has introduced the two datasets used for all the experiments in this thesis. Initial work used the dataset described in Section 4.1 is a re-acted database with neutral and a single expressive version of each sentence. Each frame is presented as a point cloud leading to easy manipulation, transformation and visualisation. For some experiments this was found to be too limited since the same sentence was required in neutral and multiple expressive styles. Therefore a second corpus was recorded in a posed manner to allow tight control of the expressive style and sentence content. The frames from this dataset were projected to numerical features using Active Appearance Modelling. The next section of this thesis provides context for the work and introduces the current state of the art.

Chapter 5

Literature Review

This review gives an overview of the main families of techniques for producing expressive speech animation, their advantages and disadvantages, and finally a detailed description of Cao et al. [2003] which provides the foundation to the work presented in this thesis.

5.1 Blendshape Approaches

Blendshapes are the extreme facial posses within a training set e.g. mouth fully open and mouth tightly shut. Novel facial expressions and posses can be created as linear combinations of these blendshapes. Alternatively, delta blendshape models are used where the blendshapes represent an offset from a neutral reference mesh. To generate a mesh, combinations of delta blendshapes are added to the reference neutral. Figure 5.1 shows a simple two blendshape example.

In Parke [1972], a mesh was painted onto a model's face, and then front and side photographs were taken of a variety of different poses. The points were hand

(a) Blendshape A (b) A and B (c) Blendshape B

Figure 5.1 A simple blendshape example mixing two meshes

measured from the photographs and transferred into a coordinate system. During animation, an approximation of the target pose was found by linearly combining different meshes. One drawback was that usability for the animator reduced as these multi-combinatorial models increase in complexity. Blendshape interpolation was quickly adapted to a region based model with the lower portion of the face representing speech and the upper portion representing expression Kleiser [1989]. Arai et al. [1996] presented a bilinear model to implement a system where one mixing coefficient is used to blend between actor meshes, and the other is used to blend between expression meshes.

In Yu and Liu [2014], a regression based approach to blendshape modelling is taken. Based on an analogy of FACS (Facial Action Coding System) Action Units (AUs) and blenshapes, the system takes hand-sketched facial expressions, with important facial features labelled (thus giving the overall impression of the facial expression being depicted). A training set of AUs is preformed by professional FACs trained actors and is tracked and transferred onto an intermediate graphics model. The hand sketched facial expression is also transferred onto the model. Using a least squares optimisation, linear combinations of AUs are solved for, allowing for an approximation of the sketched face to be produced on the model. To avoid blendshape interference (where AUs/blendshapes have non-orthogonal areas of influence) a course optimisation is first performed, which determines and

fits the major AUs, then a finer grained fit is performed on the less influential AUs. The AU interpolated mesh can then be retargeted to a different model if required. Thus novel shapes can be produced using this regressive technique using a low dimensional representation. Since the technique was only reported to be working on single frames, it is unclear whether the system would model the timing dynamics which would be encountered in animated sequences of frames.

As mentioned above, blendshape interference can be a problem, where the region of influence of different blend shapes is non-orthogonal. The animator must go back and edit previous blendshapes to adjust for the interference. The problem has traditionally been solved using additional blend shapes known as correctives. In the "The Lord of the Rings" movies, the character "Gollum" was constructed with 64 blendshapes. Interference had to be corrected with 946 additional blendshapes Lewis et al. [2005b]. An alternative solution is to allow a user to draw onto the model where vertices are in the correct position. These vertices are then shielded from further movement due to additional blendshapes. Another technique is to divide the face into blendshape regions. Points across blendshape space are mapped into areas of similar deformation (i.e. points in the same area which move a similar amount across blendshapes). Segmenting this map gives a logical breakdown of where to divide the blendshapes into areas of separate control Joshi et al. [2005]. Blendshape techniques have also been combined with mocap data. Mocap frames describing the full range of motion are paired with video frames recorded simultaneously. For each of these pairs, blendshape weights are manually tuned so that the output subjectively matches the video frame. Using these correspondences, a radial basis function regression model is used to predict new blendshape weights for novel mocap input Deng et al. [2006].

5.2 Physically Based Approaches

Physically based systems describe how a geometric face mesh is deformed by "pseudo-muscles" which are geometric representations of the movements of the face under the influence of real facial muscles. Platt and Badler [1981] describe a system which uses "tension nets". These are networks of adjacent points in the face mesh. Each point is a three dimensional coordinate, and each has a spring constant which represents a type of tissue (muscle, fat etc.). A muscle vector is attached to a point in the network and to a static point which represents bone. When the muscle vector is contracted, the movement is propagated out along the tension net according to the spring constants at each edge, intuitively like a network of interconnected springs. Tension nets are directly correlated to Ekman's FACS Action Units Ekman and Friesen [1977], meaning that if a movement can be notated using FACS, then it should be representable with the corresponding tension nets. Tension nets give fine grained control over the face surface, but large numbers of these are required, leading to high computational complexity.

Waters [1987] attempts to simplify muscle based approaches. Three types of muscles are used; linear, radial and sheet muscles. A linear muscle is a vector, which is inserted into a geometric facial mesh and defines an area of influence. When activated, all points within this area are moved towards the origin of the muscle. Sheet muscles differ from linear muscles in that there is no single point source. Instead they act like a rectangle which contracts along one of its axes. Finally, radial muscles define an ellipsoid shape where points within the region of influence contract towards a centre point. In all cases, the activation applied to a point is scaled according to its location within the area of influence, to prevent mesh tearing at the edges of the area of influence. Although these simplified psuedo-muscles reduce computational complexity, they also reduce the anatomical

fidelity of the face mesh. For instance, wrinkling and buckling of the skin cannot be modelled.

Edge and Maddock [2001] present a system based on the work of Waters [1987]. Linear, radial and sheet muscles are inserted into a graphical face model. The activations are hand tuned to match exemplar lip shapes for each of 56 visemes. Ekman's six universal expressions Ekman [1992] are also approximated in muscle activation space by creating a single set of activations for each expression. During synthesis, a phoneme to viseme mapping is used to look up the required muscle activation parameters for key speech targets. These are then interpolated to produce the final concatenation of muscle activations. The key targets can also be linearly mixed with the expressive parameter sets to produce expressive looking output. The author acknowledges that during subjective evaluation, certain lip shapes scored poorly in preference testing. This is probably due to co-articulation not being modelled resulting in inappropriate viseme shapes being selected.

In Terzopoulos and Waters [1990] a tri-layered geometry of the face is implemented to produce a more anatomically realistic model. Different layers attempt to reproduce the characteristics of the epidermis, the sub-cutaneous fat layer, and the stiffer muscular layer just above the bones of the face. Each layer is a tetrahedral mesh of varying stiffness and deformability connected to the adjacent layer meshes. The lowest layer is deformed by muscle activations, which in turn deform the layer above it, and so on. Movements from an actors face were transferred onto the geometry by integrating "deformable contours" to match captured video The model is extended in Lee et al. [1995] by using laser scanned geometry and photometry to automatically build a model of the underlying bone structure, and insert muscles into the tri-layer mesh in correct positions.

Choe et al. [2001b] presents another system which uses modified versions of

Waters' muscles. Captured mocap video is transferred to a dense point mesh with a muscle rig using reference points and affine transforms. The muscle activations required to produce each frame (or pose) are then automatically found using a steepest descent algorithm. Fitted activation trajectories in time for each muscle can then be edited using spline interpolation, and optionally constrained using keyframe points. The transformations to the skin surface are then represented using a finite element model.

Albrecht et al. [2005] presents an enhanced muscle based approach which attempts to model the complex mix of expressions observed in expressive speech. A tri-layer geometric model is implemented into which muscle functions are inserted. A text-to-speech system is used to map between text and muscle activation space for speech. Then Cohen and Massarro's dominance function approach Cohen and Massaro [1993] is used to account for coarticulation (a phenomenon whereby the lip shapes produced for a given phoneme utterance can be different depending on the surrounding phonemes). The input was also provided as audio which was mapped to the "Activation/Evaluation/Power" space Cowie et al. [2000]. A mapping is defined relating this space to a large set of complex exressive types such as distress, gloating, gratitude and shame, as well as the more fundamental expressions. Static mappings between these expressive types and muscle activation space are defined. The articulatory features are then combined with linear combinations of the expressive activations to produce expressive output on the mesh based face model. Supplied panels in the paper show convincing expression although no subjective evaluation is provided. One drawback is that the input data needs to be supplied as both text and audio. Additionally, the audio-expression correlates used to predict expression are known to be unreliable.

In an attempt to extend the physical analogy still further, Sifakis et al. [2005] model the face as a complicated tetrahedral mesh on top of a rigid bone structure.

Magnetic Resonance Imaging data is used to estimate the structure of the face beneath the skin, and adjust the stiffness of the mesh according to whether an area is bone, muscle, fat or dermis matter. A high resolution outer mesh is constructed using point data from a depth scanner. Muscles are inserted into the tetrahedral mesh. Muscle activations which fit the model to reference mocap data points are automatically found using a Gauss-Newton algorithm and then clustered into groups which correspond to phonemes which they call "physemes". Once learned, these physemes can be concatenated to provide new speech. Additionally, activations for muscles associated with expression such as the Zygomatic Major muscles (responsible for smiling) can be added in to provide emotional context. The system they describe does require large amounts of post-processing to overcome errors in the initial transcription of audio into phonemes during synthesis and blending of physemes to overcome co-articulation problems, although this is a general problem not only encountered in this technique. The approach is extended to expressive synthesis in Sifakis et al. [2006], where static expressive poses are mapped from mocap points to muscle activations and blended with the activations generated for mouth articulation. Although this is a flexible approach that in theory should allow for blending of many different expressions, being a static linear blend of speech and expressive muscle activations, and does not attempt to dynamically model the interaction between expression and mouth articulation.

In Kahler et al. [2001], the face model is simulated as a tri-layer geometry representing skin, muscles and bone. Facial muscles are modelled as strings of piecewise linear segments with ellipsoids fitted to each segment to simulate volume. Contractions are achieved by performing affine transforms on the ellipsoid segments. Although this is a novel approach to modelling the physics of the face, it makes no attempt at automatic speech articulation or portrayal of expression.

In Rumman and Fratarcangeli [2014] a similar approach is used employing vol-

ume preserving position based dynamics, which impose non-linear constraints on a deformable geometric model. Pseudo-muscles are inserted into the mesh to mimic the major facial muscles in the human face. Each muscle has its own stiffness coefficient. The layered mesh is deformed using radial basis functions. The major contribution of the work is that the facial model is mesh independent. Different mesh configurations and numbers of layers can be used, and the response to the inserted pseudo-muscles is solved using the Gauss-Seidel (Liebmann) method. The authors note that this convergence can sometimes be unstable, the method makes not attempt at mouth articulation (only expressive poses) and takes upward of 1 second to process the the mesh (on an Intel Core 2 Duo 2.4 Ghz CPU) for a single frame therefore is not real-time.

Other techniques model muscle movements by abstracting away from an exact physical analogy. Magnenat-Thalmann et al. [1988] describe facial poses as combinations of "Abstract Muscle Action" procedures or AMAs. Each AMA controls a combination of muscles which may for example open the jaw, or pucker the lips. The overall movement of the face is viewed as a set of AMA trajectories through time. Keyframes are modelled by hand, and the trajectories between them are constructed using b-spline interpolation. Liu et al. [2001] present a technique for transferring geometric representations of expression between faces and also the consequent changes in lighting. Firstly, expressions are transferred between faces. Offsets are calculated between landmarks on neutral and expressive faces. Since all human faces are roughly similar in size and shape, it is assumed that applying the same offset (altered with an affine transformation), should be able to warp a novel neutral face into an expressive pose. Further image processing is applied to warp the texture map to the new shape of the landmarks. The novelty of this work lies in the transferring of lighting intensities between poses. They describe what they term the *"Expression Ratio Image"* (ERI). This is essentially the ratio of light in-

tensity across all points in the source neutral and expressive images. The ERI takes into account surface normals, number of light sources and their relative intensities and colour of light. Once the target image has been warped to its new pose, the corresponding ERI is then applied to each pixel in the image. This allows details like skin wrinkling (such as that caused by brow furrowing) to be reproduced in the output. Provided results look impressive although the authors acknowledge that the technique has difficulty in mapping features between scenes recorded with very different lighting conditions. Also the range of expressions is limited and all provided examples are shown with the mouth closed. More recently, Yu et al. [2014] describe a system for transferring facial expressions between models. A library of different facial poses is collected from human actors according to different FACs requirements. These poses are then transferred onto a template mesh. A correspondence between the template mesh and a target mesh is established. The template is aligned to the target mesh using a laplacian coordinate system and 1-ring vectors and is therefore invariant to pose and scale. Once the meshes are aligned, the target mesh is then warped to the exact shape of the template mesh using a system of vertex projection and barycentric coordinate calculation. They then create what they call "morph functions" which are strings of FACs units parameterised by things like acceleration, peak latency and peak amplitude. Morph functions output can then be used as key frames allowing for strings of animation showing changing face pose to be produced. Although supplied panels in the paper look convincing in terms of expressive realism, the technique does not attempt to model articulatory speech so cannot be considered a visual speech synthesis technique.

Mazzei et al. [2012] was an interesting hybrid attempt at producing realistic facial expressions on both an animatronic robot and a graphical model. A human-like face was created out of a patented rubberised compound into which 32 servo

actuators were inserted to mimic the movement of real facial muscles. A mapping was created between FACS parameters and actuator control, allowing for the continuous gamut of expressive space to be displayed on the face. Preset combinations of activations were created which mapped from Ekmann's fundamental expressions, and also from valance/arrousal space, to the servo inputs. A rule based conflict resolution module prevented incompatible movements from being sent to the servos (e.g. eyebrows raised and frown), and another module added nuances such as eye gaze and head movement. For the graphical output, a mesh based psuedo-muscle model was created with the pseudo-muslces inserted into the mesh at the same place as the servo activators were inserted in the rubber face. Mesh responses were hand-crafted and the graphical model was then driven by the same input parameters as the animatronic model. The technique and models were used in a therapeutic setting to treat children with Autism Spectrum Disorders (ASDs). The authors report half of the expressions produced by the model as being correctly identified (by both children with ASDs and those without). The technique made not attempt to model speech.

Bermano et al. [2014] present a novel approach to generalising the input to a geometry based facial synthesiser with functionality to alter the dynamics of delivery. A database of facial geometry is created capturing the full range of poses captured from an actor. The database is processed with a low pass filter to make low frequency versions of each frame. The differences between the low frequency and original versions of each frame are then calculated. Input to the system then takes the form of land marked tracked video frames, rigged blend shape models or Xbox Kinect captured face parameters. These undergo an affine transform to align them to the training data and are used to warp the low frequency mesh. Appropriate high frequency details are the added from the pre-calculated offsets and texture patches applied thus allowing high definition output on the model

in a computationally inexpensive way from different sources. The technique is untested on dissimilar input and output data however, since the same actor was used to create the training set and provide the input frames.

5.3 Concatenative Approaches

Concatenative speech synthesis works by concatenating individual video frames, visual features or some other representation of visual speech, from a corpus which ideally contains all shapes, appearances and expressions needed in a synthesised sequence. Typically some kind of mapping is established between text/audio and unit selection. Often a cost function is used to calculate a suitability metric for each concatenation unit based on the unit's initial distance from the target unit and any processing the unit will require after concatenation. The synthesiser chooses units by minimising the sum of these cost functions. One of the first problems one encounters with this approach is coarticulation. This is when the same sound produces multiple mouth shapes. To account for co-articulation Pelachaud et al. [1991] describe a system where each phoneme is given a deformability rank. Forward and backwards coarticulation rules are then applied and more deformable phonemes are morphed towards the shapes of less deformable phonemes. Bregler et al. [1997] use a triphone selection method where the cost function is a distance based on a combination of phoneme context distance and lip shape distance. Phoneme context difference is scored according to whether the candidate phoneme and the target phoneme are the same, are a different phoneme but the same viseme class (e.g. /b/ and /m/), or whether they are different phonemes. The lip shape distance is the distance between lip shapes in overlapping phonemes at the beginning and end of each triphone. During concatenation, triphones are overlapped to find the best match of lip shapes between the end of one and the beginning of the next. Thus

accounting for co-articulation artefacts occurring within one frame of the target.

Distances between candidate and target units can be calculated in a variety of ways such as with Principal Components Analysis (PCA) coefficients of feature shapes and pixel luminance Theobald [2007], geometric features such as mouth width and height Bevacqua and Pelachaud [2004], or phonetic/prosodic information. In Cosatto and Graf [2000a] the text-to-speech module Festival Clark et al. [2004] is used to tokenise a text input into phonemes and durations. For each phoneme, a number of candidate mouth shapes are selected from a corpus. This creates a lattice with several frames at each time point, each with a cost. The Viterbi algorithm Forney [1973] is used to find the path of minimum cost through the lattice. To minimise coarticulation errors, they also use the triphone selection method from Huang et al. [2002]. By considering phonemes at the triphone level, there is a better chance that the middle phoneme will have a phonetically appropriate context with the surrounding phonemes.

For flexibility in unit selection, a corpus of images may be decomposed into their component parts i.e. mouth, eyes, nose etc. and labelled. When an utterance is synthesised, the corpus is searched for the most appropriate features for the target face. The individual face parts are selected, stitched together and processed to produce a candidate frame Cosatto and Graf [2000a]. Because the synthesised frames are created from actual images of the original speaker, they are photorealistic. However in common with other concatenative synthesisers, the output is dependant on a large corpus of labelled image samples from which to generate novel output. All generated content will look like the actor in the training database. Approximations are made where no matching frames can be found. Adding additional emotional contexts will increase data storage requirements. There is also the issue of changing between emotional states. A smile doesn't instantly appear on a person's face, so intermediate frames between neutral and emotional expressions

would also be needed or the categorical expressions blended.

Deng and Neumann [2006] use a novel projection to a 2D manifold which they term "isomaps". They claim that isomaps are a more intuitive and useful dimensionality reduction than other comparable techniques such as PCA. The training corpus of expressive visual speech is subjected to a forced alignment in audio space, and the corresponding phoneme isomaps are clustered into phoneme groups and expressive groups. A novel phoneme transcript can then be synthesised by concatenating isomaps using a dynamic programming algorithm to minimise a cost function. Additionally, soft constraints in the form of expressive groups and hard constraints in the form of isomap phoneme groups can be specified by the user, thus allowing the addition of expression into the output. However, only categorical expressions can be achieved, and there is no control of expression intensity. The authors also acknowledge their lack of subjective evaluation.

Kshirsagar et al. [2001] attempt to produce expressive visual animation using PCA. They capture expressive visual speech from an actor using optical facial markers placed to imitate the MPEG-4 FAPs standard Ostermann [2002]. PCA is then applied to the raw captured co-ordinates of these markers. They term the resulting features "expression/viseme" space. To synthesise novel speech, phonemes in the training data are mapped to visemes in expression/viseme space. The visemes are then concatenated in this space, projected into FAPs space and then applied to a model which can be driven with FAPs parameters. Single frames of Anger, Fear, Happines, Surprise, Sadness and Disgust are also projected into the viseme/expression space. When a novel utterance is required in an expressive style, the visemes are linearly combined with the expressive features in expression/viseme space. The novel sequence is then subjected to cubic spline smoothing. Given that there is no consideration of surrounding phonemic context, there is presumably a problem with coarticulation which is not addressed in the paper. Cubic spline

smoothing will attenuate output somewhat, and it is not clear how the linear addition of expression is managed to give dynamic expressive output.

Beskow and Nordenberg [2005] use a similar approach where data is captured using reflective markers placed on the face to simulate MPEG-4 FAPs feature points. Sentences are recorded in the six universal expressions Ekman [1992] and neutral. A FAPs drivable graphics model was developed. PCA was applied to the raw coordinate values of the collected feature points. A separate PCA model was trained for each expression. Forced alignment was performed on the audio and a mapping of phonemes to visemes was established in PCA space. Visemes were concatenated, and the blending technique described in Cohen and Massaro [1993] was used, employing minimisation of error using dominance functions to account for coarticulation. The PCA features were projected back into FAPs space by multiplying the principal components by one of the expressive PCA models, thus allowing novel speech to be projected into an expressive modality. The resulting FAPs parameters could be applied to the face model to produce visual speech. A subjective evaluation showed that expression identification was significantly better than chance. The model was however not able to create blends of expression and could only output categorical expressive speech. Complete phonemic training sets were required for each expression.

In Bevacqua and Pelachaud [2004], seven phonetically relevant parameters define lip shapes. During synthesis, a phonetic transcript is used to select viseme control points for each parameter and b-spline curves are used to interpolate between such points. Coarticulation error is avoided by taking each consonant shape in the context of each vowel, and then carefully selecting which consonant shape to use for each control point. Additionally the dominance based coarticulation rules described in Cohen and Massaro [1993] are used. A database of consonantal targets is created for each expressive type and can be specified to allow for expressive

synthesis. Weighted combinations of these database units can be specified to allow for mixes of expressions and modulation of expressive intensity. This approach relies on having a full corpus of training speech in each expressive style. B-spline interpolation may lead to over smoothed output. Furthermore although the plots of ground truth and synthesised trajectories look convincing, no subjective evaluation is offered. In Bevacqua et al. [2007] the system is extended. The face is divided into 8 separate regions. A mapping between FAPs parameters and different expressive styles is created allowing for an expressive style to be controlled as a single parameter. Keyframes for speech and expressions are defined during synthesis where the expressive parameter is subject to an attack, decay, sustain and release window. B-splines are again used to interpolate between the keyrames. Output is provided by a FAPs drivable graphics model. A framework is presented which allows for the combining of two expressions, where expressive dominance functions deal with the relative application of the two expressions and conflict resolution in the case where control points are manipulated non-orthogonally. A technique for synthesising head position and eye gaze is also presented in the paper. Objective evaluation was offered via the "copy-synthesis" method Buisine et al. [2006], where synthesised movies are subjectively compared with hand crafted animation. Results indicate that the automatic approach is "satisfactory" and that participants were able to identify at least some of the expressions displayed.

More recently Liu and Ostermann [2011] built an image based concatenative synthesiser heavily influenced by Cosatto and Graf [2000b]. A training corpus of neutral and happy speech was recorded. The mouth part of each frame was located, extracted and then subjected to PCA. Each mouth patch was stored as a PCA feature vector, along with its corresponding phoneme and phoneme context. Natural expressive speech was then analysed to discover certain rules governing the interaction of smiling and speech. Synthesis is achieved by using a text-to-

speech program to generate a phoneme stream from audio which is used to select frames taking into account phonemic context (as in Cosatto and Graf [2000b]). The discovered rules are used to decided when to switch between smiling frames and neutral frames. The string of selected mouth images is stitched back into the face image, and head movement is applied. A subjective evaluation showed that participants were not able to reliably tell the synthsised happy faces from ground truth. Mean opinion scores were lower for synthesised sequences that ground truth. Although this technique appears to produce high quality output, it is only able to produce happy and neutral sequences. Since it only models the mouth region, it is very limited in its ability to produce other expressions such as surprise and anger, which rely on the upper part of the face.

In Serra et al. [2012], a concatenative approach is described which uses a straight forward text-to-speech audio synthesiser or audio analysis module to create a phonemic transcript from either textual or audio input. A one-to-many mapping is created between phonemes and visemes and used to translated between phoneme transcription and concatenations of viseme selections. The viseme selections are then mapped to animation curves which have been manually crafted by an animation artist. Finally the concatenation is passed into a Maya rig for rendering. While this system is extremely simple and easy to implement, it makes no attempt at co-articulation resolution. Output movies hosted on YouTube also show that inappropriate plosive lip shapes are produced.

Although concatenative based synthesis has produced some impressive results, there are several reasons why it is not suitable for emotional synthesis. For each new emotional context to be added to the corpus, an actor is required to repeat the same sentences in a different emotional context. Much of the data being recorded will be redundant. For the so-called six basic emotions (anger, joy, surprise, disgust, sadness and fear) each sentence must be recorded seven times (one extra for

neutral). The speech signal of each of these emotionally expressed sentences is approximately the same. Only the emotion portion of the signal varies between them. Therefore separating the speech and emotion into separate modalities, would reduce storage requirements. Of course, the expression of human emotion must be considered as a complex mix of different feelings. Reducing it to six basic emotions is an unacceptable simplification if realistic synthesis is to be achieved. Having discrete emotional contexts doesn't allow for subtle mixes of emotional context which make up the rich texture of human facial expression.

5.4 Statistcal Approaches

Statistical models provide a framework where an input signal (typically text or audio) can be parameterised and used to drive a model which in turn produces visual output. Ideally the model would be flexible enough to account for co-articulation, mixes of expression and changes of actor or output model.

5.4.1 Decompositional Models

Deng et al. [2004] describe a system in which a low dimensional representational of expression is used, which they term PIEES *Phoneme Independent Expression Eigen-space*. Matching sets of expressive and neutral sequences are recorded using motion capture. They are time aligned and a simple subtraction is performed, leaving a set of expressive residuals which are phoneme independent. PCA is performed on these residuals which yields a low dimensional expressive surface. Neutral visual speech is then synthesised (how is not described) and blended with expressive styles derived from the PIEES. Intensity of expression is controllable. This could be a promising approach providing a mixed space which might be capable of pro-

ducing subtle mixes of expression. However it is an over simplification to simply subtract neutral speech from expressive speech, in the hope of completely removing phoneme contribution. For instance speech articulation affects the production of a smile, so subtracting articulation contributions from smile contributions will lead to an irregular signal. Also as the authors accept, a large amount of training data is required. No subjective evaluation for the technique is offered. Chan and Tsai [2010] extend this technique by using PCA with expectation maximsation to create their PIEES space. The PCA algorithm then predicts the most likely position for missing data in a sample, thus making the technique robust to occluded landmarks.

Du and Lin [2003] describe a system for synthesising expressive face images given neutral input images. A training set of 213 images of 10 female participants make up a training set. Each frame is shown in neutral and 6 basic emotions and is given an "emotional parameter" score which is how much of the 6 component expressions in the training set is observed in the image and is therefore a 6D vector. The emotional parameters are the mean opinions of 60 students. The images are labelled using a landmarking technique and PCA is applied. Each expressive image is then subtracted from its neutral equivalent to produce a set of expressive deltas. To add expression to a novel neutral image, the corresponding landmarks are found and the deltas for the required expression are applied. This provides an expressive shape. The appearance patch is then subjected to the same process described in Liu et al. [2001] and is warped to the new shape. This creates a patch which contains the correct lighting and appearance for the new expression. A polynomial function is derived which maps between the emotional parameters and the PCA coefficients used to encode the shape component of the model. Therefore the technique allows expression to be added according to a given emotional parameter and is interesting in that it allows the expressive gamut of training images to be

modelled. However it makes no attempt to produce time varying strings of frames so cannot be considered an animation or visual speech synthesis technique.

Hong et al. [2002] offer a Neural Network (NN) based approach. Their NN maps between audio blocks and expressive visual parameters they term "Motion Units" (MUs). A training set is acquired from video by tracking points on an actor's face. The training set consists of all the English phonemes spoken in neutral, happy and sad and amounts to around 1000 frames. Each frame is tracked and the landmarks on the face are aligned to and subtracted from a neutral reference frame. PCA is then applied. These residual deformations in PCA space are the MUs. The audio in the training set is then clustered into groups defined by Gaussian Mixture Models. A separate audio to MU Neural Network is trained for each cluster using context for the three preceding and following frames to account for output discontinuities and some level of coarticulation. During synthesis, audio input is designated to one of the training clusters, and the corresponding Neural Net is selected and used to predict the MU output given the audio input. A further set of Neural Nets is trained to map between neutral MUs to Happy and Sad MUs. The synthesised neutral MUs are then passed into one of the expressive NNs in order to produce expressive output. Objective numerical testing shows that the technique produces output comparable to ground truth. In a subjective test, participants were able to identify the expression produced, although no test statistic was provided. The participants were shown the sequences with expressive audio, therefore the experiment must be considered biased since the audio would have given cues as to the expressive style. Additionally, they were only choosing between happy and sad, so it is difficult to attach much significance to the result. Since the training set only contained examples of English phonemes and no bi-phones, tri-phones or any other higher order representation of speech, it is unclear how coarticulation was modelled although the use of recurrent neural nets could

potentially solve such problems. Since the expression prediction Neural Nets are discrete, the system is unable to produce blends of expression.

Chuang [2002] describes an expression retargeting system, where video is tracked and keyframes (in coordinate space) are picked which describe extremes of variation. A set of linear weights describing the correct combination of keyframes to match each frame of a video is then obtained (it is assumed this interpolation works since the keyframes describe extremes). To retarget the expression, corresponding keyframes in a target model are crafted by hand, and the weights discovered in the previous step are applied to recreate the source video on the target mesh. In later work Chuang and Bregler [2005] describe a technique based on Active Appearance Models, where a training set models variation in face shape under the influence of speech and also expression. Input parameters are factorised using a bilinear model, into those representing speech articulation and those representing expression. Since expression affects the way the lips move when producing speech, a weighting function is used to modulate the contribution of the speech and expression parameters to the final frame. The bi-linear factorisation however is a difficult problem which is simplified by either holding the expression or speech component constant, and solving for the other component. This is obviously an oversimplification since signals by their nature vary and should not be considered static for any window larger than a few milliseconds.

Vlasic et al. [2006] describe another technique based on statistical factorisation using higher order multilinear algebra. A tensor (a matrix of dimension higher than 2), is used to arrange a dataset of dense 3D face scans, displaying 16 actors articulating 5 visemes in 5 expressions. A Singular Value Decomposition is used to reduce the dimensionality of the data and provide controls for manipulating identity, expression and mouth shape. Face movements can then be tracked in video, and the resulting features mapped to tensor coefficients. These coefficients

may be manipulated to alter either the identity or expression of the actor and is therefore a movement transfer method. Only modelling five visemes seems unlikely to produce enough variation in lip shape to create convincing articulation and no subjective or objective evaluation is supplied.

5.4.2 Hidden Markov Models

One application of Hidden Markov Models (HMMs) is to predict hidden states given a sequence of observations. The model is an arrangement of state sequences each with an emission probability and transition probabilities. Emission probabilities govern the likelihood of seeing a particular observation given a particular state. Transition probabilities govern the likelihood of moving to the another state given the current state Stamp [2004]. As a basic example applied to speech recognition, some parameterised form of audio speech is obtained (such as Mel Frequency Cepstral Coefficients), and grouped into phonemes (or some other type of unit). Each group is used to train an HMM model to recognise the optimum parameter set for that group as a Gaussian distribution. During recognition, observations for a unit are passed through each trained model, to calculate the probability of the unit corresponding to that model. The unit is classified as that whose model probability is highest Rabiner and Juang [1986]. Figure 5.2 shows a simple three state HMM with transition and emission probabilities.

More recently Tokuda et al. [2000], HMMs have been used for synthesis where observations are predicted from a sequence of known states as opposed to predicting hidden states from observations. The HMM sequences and gaussian distributions are trained in the same way as for recognition. During synthesis, a sequence of phoneme level HMM models is concatenated according to an input sequence. As the HMM concatenation is traversed, observation parameters are output accord-

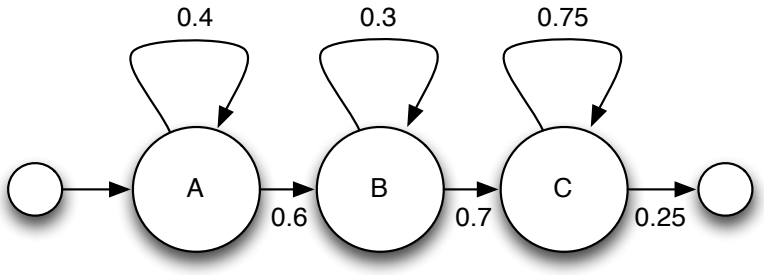


Figure 5.2 A simple three state HMM with transition probabilities. At each state, the model either stays in the same state or transitions right to the next state.

ing to each state's Gaussian distribution. The deltas and delta-deltas are usually computed in addition to the current feature which allows for the dynamics (acceleration, attack, decay etc.) of speech to be modelled. The generated features and deltas are then passed into some visual output generator for rendering.

Cox and Simons [1990] present an early HMM based speech synthesiser, designed for transmission of Facial Action Codes (FACs) across a wire. In training, speech audio is quantised into speech vectors and labelled. Corresponding video frames are labelled using image codes which encode only the mouth shape. A fully connected Markov Model is trained in which each state represents one of the image codes and has associated emission probabilities (the likelihood of observing each of the quantised speech vectors), and transition probabilities. During synthesis, the model produces the most likely sequence of mouth shape states, given the input speech vectors using the Viterbi algorithm. The mouth shape states are converted to FACs to be transmitted or to drive a graphics model.

Brooke and Scott [1994] present another early HMM synthesiser. Greyscale video footage of a talking mouth is used as input to the system. The images are projected onto a PCA model to reduce dimensionality. The PCA parameters are then clustered into triphone groups and used to train HMM models. The

number of states in each model is set to be the average number of non coincident peaks and troughs across plotted PCA coefficients for the corresponding triphone, thus allowing for the different durations of triphones. The triphone models are trained using the Baum-Welch algorithm. During synthesis, triphone models are concatenated according to an input triphone string. The models are traversed left to right, with no skips. At each time point, a feature vector of PCA coefficients is generated according to the current state's Gaussian distribution. These feature vectors are constrained using quadratic curves and then smoothed using a five point window. The smoothed PCA coefficients are then projected back onto the PCA model to create output greyscale images. A problem with this approach is the fact that each state produces feature vectors independently of the surrounding states. Although the output is smoothed in order to reduce jerkiness, it doesn't allow for coarticulation corrections, and the smoothing will attenuate the motion of the output. There is also no timing model meaning that dynamics cannot be realistically produced.

Masuko et al. [1998] designed an HMM based synthesiser which attempts to model the dynamics of speech and coarticulation. A training set of images is phonetically labelled at syllable level. Features for each training image are automatically extracted. These are the height and width of the mouth, and discrete cosine transforms of some other lip measurements. Additionally, deltas are calculated for each feature vector simply as $\Delta c_t = c_t - c_{t-1}$. These feature vectors and their deltas are used to train three state, syllable level HMMs. Transition probabilities are then calculated by aligning the training data to the models with the Viterbi algorithm. This ensures that the fixed length HMM models produce the correct number of observations for a given syllable. Since the delta coefficients provide each observation with some context of the surrounding observations, the output of the HMMs takes some consideration of coarticulation and does not re-

quire smoothing. Tamura et al. [1998] extend the technique by driving it from audio. Another HMM is trained using MFCCs to recognise input speech audio and output a text syllable string. This syllable string is used as the input to an HMM as described above.

Brand [1999] developed another HMM based approach to visual synthesis. An HMM is trained where the hidden states represent different face configurations. This model is combined with synchronised audio to give a combined HMM capable of observing audio features and representing the hidden states as visual face configurations. During synthesis, novel audio is analysed and the most likely HMM face configuration state sequence is calculated. From this, face configuration output probabilities are generated which can be used to select frames of video or drive a parameterised model. The model is constrained such that the output features require no smoothing and coarticulation is handled by virtue of the contextual nature of HMMs.

Tao et al. [2009] describe an HMM based expressive visual synthesiser which uses Gaussian Mixture Models to add expression to the outputs. A large corpora of 700 neutral sentences and 300 sentences each in neutral, happy, angry, sad and surprised states was captured using mocap markers placed on the actors face to replicate MPEG-4 FAPs parameters. The data was then considered in 150ms blocks of audio, and the corresponding 7 frames of visual data, meaning that for each audio chunk, a degree of context was modelled in the visual domain thus accounting for coarticulation. A fused HMM was trained which modelled the interaction of the audio and visual domains. During synthesis, novel audio was passed into the fused HMM which produced costs for the candidate visual sub-sequences. The Viterbi algorithm was used to find the minimum path through the sub-sequences and the best ones were concatenated to produce neutral output. To introduce expression, GMMs were trained on time aligned neutral and expressive

data to model a joint probability distribution between speech and expression. One GMM was trained per expression. The neutral synthesised output was passed into one of the expression predicting GMMs which then produced the most likely expressive output. Objective trajectory comparisons with ground truth and other methods show this technique appears to work well. No subjective evaluations were reported. Some drawbacks of the method are that a large amount of training data is required, and although the GMMs will model the variation within an expressive class, the technique is unable to produce blends of expression. Furthermore, the neutral synthesis component is relatively complicated.

In Anderson et al. [2013] an Active Appearance Model is trained, and modes encoding head rotation and blink are identified and removed. A static model and texture are used to represent the teeth. Quinphone HMM states are trained for all the data in the training set composed of 6925 sentences split between six expressions. Cluster Adaptive Training is used and search trees are created containing the mean and covariance for each quinphone HMM state. A separate tree is used for each expression. During synthesis quinphone HMM models are selected (based on the output of a text-to-speech system) and concatenated. Their emission parameters are a matrix containing the quinphone in all the different expressive styles in AAM space which is then multiplied by an expressive modulation vector allowing for a synthesis of the quinphone across the continuous gamut of expression. The AAM parameters are then projected back onto the AAM model to produce visual output. A large crowd-sourced evaluation was executed. The technique scores an expressive realism score of 3.7 (out of 5) surprisingly low, considering that the output is ostensibly photorealistic. Expressive recognition rates are significantly better than chance but interestingly are also better than scores for ground truth footage. A drawback of the approach is the very large training corpus required (over 1000 sentences in each of the expressive styles).

5.4.3 Other Techniques

Jia et al. [2011] use a technique based on PAD (Pleasure/Displeasure, Arrousal/Non-Arrousal, Dominance/Submissiveness). These can be seen as a high level description of emotion. The system takes text and PAD parameters as input. Neutral audio is synthesised, and then expression is added by a system of "boosting GMMs" which model the difference between neutral and expressive speech. An intermediate coding of expressive features called Partial Expression Parameters (PEPs) are presented. These are a middle state between PAD parameters and the MPEG-4 FAPS parameters which are the output of the system. PEP features are converted into FAP parameters by a linear interpolation. The mapping between PAD and PEP is done subjectively by asking participants to mark various training images according to how they perceive them. Once the neutral audio is synthesised, a static mapping between phonemes and visemes is used to produce visual output FAPs. For non-mouth portions of the face, these are replaced by the PEP mapped FAP parameters, the magnitude of which is modulated by the F0 pitch of the speech. For mouth FAPs, a simple linear mix is applied between PEP mapped FAP parameters and FAP parameters for articulation. The mixing coefficient is static at 0.8 in favour of articulation viseme. This approach has several drawbacks. Firstly the problem of static visemes and coarticulation does not appear to be addressed. Secondly, it relies on subjective opinions of expression in the PAD to PEP mapping. Thirdly modulating expression based on the F0 pitch of speech is invalid, as the frequency of certain emotions (such as sadness) are invariant to voice pitch. Forthly a static blend of expression and speech is invalid. For instance, when one puckers the lips in the articulation of words such as *"why"* and *"fortune"*, the stretching of the lips for a smile must be relaxed somewhat in order for the articulation to take place. Therefore the blend coefficient should model the dominance of articulatory movement which it does not.

Yu et al. [2012] present a system for synthesising facial expressions and evaluating the correlations between FACS action units (AUs) and subjective inference of expression. 4 FACs trained actors were filmed performing various AUs. Their performances were tracked using optical flow techniques and their movements transferred onto a graphics model. Random selections from the transferred AUs were then played to untrained participants. The participants were asked to categorise the facial expression they were seeing in terms of Ekmann's fundamental expressions, and in terms of expression intensity. By performing this type of noisy sampling (producing random AUs to be performed), an unbiased correlation between AU and perceived expression was learned. A statistical evaluation showed significant correlations between AU and perceived expression across participants.

An interesting rule based system for expressive pose creation was presented in Seif El-Nasr et al. [1999]. An intelligent agent is modelled, which is able to react to external events and produce appropriate expressive responses. The agent has a series of predetermined goals. External events when perceived by the system are given a desirability rating based on their impact upon the agent's goals. The desirability for an event is then combined with the expectation of that event based on recent history. In this manner, the agent is able to learn from its experience, and will respond less to frequently occurring events. A linear mapping is created between the expressive scoring given to an event and the controls of a 2D cartoon baby face on which the expressive output is modelled. Although not a technique for synthesis of expressive visual speech animation, it is nevertheless interesting in that it attempts to intelligently react to external stimulus. However the model's rule based system which is used to decide on output expressions is an unacceptable over-simplification of human emotion and can only be considered an early prototype for a more sophisticated model of expressive intelligence.

Mlakar and Rojc [2011] is an attempt to create another rule based approach. A

unit based synthesiser relying on FAPs units, the system takes such units as inputs to create speech articulation, expression and other more subtle movements such as blinking and head movement due to breathing. A sophisticated set of rules governs temporal, spatial and power components allowing for changes in the speed and attack of delivery. A further set of "fluidity" rules govern the transition between action units to ensure that movements are smoothly contiguous and that non-linear dynamics can be modelled. Another set of rules governs conflict resolution of units with overlapping areas of influence and their relative dominance. Output is rendered via a graphical model. No evaluation is provided and it is unclear how co-articulation is handled. However, this is a full-body synthesiser, so accurate speech articulation may not be the author's first priority.

5.5 Conclusion

So far in this chapter the major techniques for visual speech animation have been explored. Although many of these techniques are capable of producing convincing results, there are themes in the literature representing major drawbacks which will now be summarised.

It is commonly accepted that there are fewer visemes than phonemes. This is because there is simply not the same level of information contained in the visual signal. This can be intuitively proved by considering that it is possible to completely understand speech when one can hear the speech but cannot see the speaker's face. However, when one can see the speaker's face but cannot hear the speech the same cannot be said. This is partly because some articulators cannot be seen such as the teeth and tongue which are responsible for some of the plosive consonants such as /k/, /t/, /d/ and /g/. It is also impossible to see the production of nasal phonemes such as /n/ and /ng/, as well as unvoiced sounds such as

/s/ and /sh/ and affricates such as /ch/. Certain phoneme groups look identical on the mouth such as the (/p/, /b/, /m/) group and the (/f/, /v/) group. Software such as Nuance's Dragon, and Apple's Siri show that audio recognition, although not perfect, is now a mature technology achieving very acceptable accuracy. If there was a one-to-one mapping between audio phonemes and visual visemes, then the accuracy of visual recognition would be high and visual speech synthesis would be trivial. There is no such one-to-one mapping which makes both problems very difficult. Since there are multiple phonemes for each viseme, one might think that it makes the task of selecting visemes to match a string of phonemes easier. Then one encounters the major issue in visual speech synthesis, coarticulation. This is where the lip shape used to articulate a phoneme changes depending on the surrounding phonemes. Taylor et al. [2012] show that there are at least 6 lip shapes corresponding to the phoneme /t/ depending on what word the phoneme appears in meaning the assumption of a static one-to-one relationship between phonemes and visemes is invalid. Since many of the techniques described above make this assumption, they must be considered to have issues of coarticulation.

Although some of the techniques described above attempted to produce expressive visual speech, most are only capable of producing categorical expressive styles. The human face is a window onto the complex gamut of feelings and thoughts that run through all our minds. It is rare to feel categorically happy or sad or surprised etc. More commonly we feel ambivalence e.g. a musician might feel happy, excited and nervous before a concert; or a soldier might feel fear, anticipation and boredom before a battle. In order to simulate these subtle complexities, a synthesis technique must either have examples of all these emotions in its training set or must be able to interpolate between the more limited expressive examples it *does* have to create new expressions. Furthermore, it must be able to modulate the magnitude of the expression displayed in a realistic manner. Very few existing techniques are

able to account for both these things.

Other techniques (particularly concatenative methods) require very large training sets. These are time consuming to collect and label for neutral speech. This problem is multiplied when considering a collection of expressive speech. For concatenative methods, either complete corpora must be collected and labelled for each expression, or some kind of transform must be used on the neutral speech which is likely to lead to noise and/or degradation of realism. Furthermore, the problem of categorical expressions described above still apples.

Some of the techniques described above are computationally complex such as the partial derivative based gradient descent training described in Sifakis et al. [2006], the tuning of dominance functions described in Cohen and Massaro [1993], or the resolution of mass spring weighted tetrahedral meshes under the influence of multiple activation inputs described in Terzopoulos and Waters [1993]. These kind of computational complexities will cause bottlenecks in an industrial workflow. Although computers have and should (at least for the foreseeable future) continue to grow in speed and capacity, so grows the complexity, realism and visual quality demanded by the animation industry and the consumers of digital media. Therefore, the argument that as computers grow in power, the techniques become more feasible does not necessarily hold. In comparison, it will be seen in this work that the proposed solution, although not an entire synthesis pipeline, is at least computationally cheap and should work in real-time, this on consumer level hardware using Matlab with its relatively slow performance (when compared to compiled low-level languages such as C / C++).

Another issue with many of the existing techniques is their lack of a robust and consistent evaluation. If any evaluation is offered at all, it is usually of an objective type e.g. comparing feature trajectories between synthesised and ground

truth data, or calculating RMS error between synthesised and ground truth meshes etc. However, Theobald and Matthews [2012] show that objective measures are not necessarily a good indicator of the subjective perception of naturalness in a synthesis technique. RMS error seems a particularly poor choice since it averages across an entire sequence, whereas the authors show that an artefact in a single frame of a sequence can lead to the entire sequence being perceived as bad. They report that using the DTW distance between a synthesised and a ground truth sequence provides the highest correlation with subjective opinions, a technique which is not used in any of the literature. When subjective evaluations are provided, they tend to be mean opinion score based, sometimes do not compare to ground truth or ask the participant a trivially simple question (such as "is this sequence happy or sad?"), unlike the forced choice turing test presented later in this thesis.

5.6 Desiderata

Having investigated the current state of expressive visual speech synthesis and outlined the major problems, it is possible to produce a list of aims for a proposed system.

1. Coarticulation: Since coarticulation is the major issue in this area, any expressive visual speech technique must produce coarticulation free output, that is the lip shapes produced must plausibly match the audio phoneme to which it corresponds.

2. Training Data: Since large corpora are difficult and time consuming to collect, a technique which requires only a comparably moderate amount of training data is highly desirable.

3. Expressive Controls: The system should have controls providing an animator with a means to control the expression produced.

4. Expression Blending: Since facial expression is a physical manifestation of the complex emotional gamut of human expression, our feelings are usually a mix of more "fundamental" emotions. Being able to reproduce this blending of emotion is necessary to realistically model facial expression.

5. Time Complexity: Geometric modelling of facial geometry is highly processor intensive. A system which can produce frames of animation in real-time is desirable as it lends itself to useful functions such as real-time actor expression retargeting.

6. Flexibility: Ideally any proposed system should not be tied to a particular dataset, output model or actor. This increases the usefulness allowing the technique to be used by more people for a greater variety of tasks.

7. Simplicity: A system which is easy to understand and implement is more likely to be useful to people. Busy professionals don't have the time or inclination to learn complicated function calls or statistical techniques. Therefore the proposed system should be simple for the end user to use and should obfuscate complexity.

8. Subjective Testing: As has been stated, most existing techniques have not been subjectively evaluated by the viewing public. Since people are the consumers of this kind of media, it is their opinion which ultimately counts. Therefore the proposed system must stand up to subjective testing.

What follows is a detailed description of the techniques on which the rest of this work is based.

Chapter 6

Technical Background

6.1 Introduction

The task of Expressive Visual Speech Synthesis is to create a smoothly varying visual representation of speech, with convincing lip movements accurately synchronised to an audio track, coupled with recognisable facial expressions to convey to the observer an emotional context. Since expressive speech as it is perceived by humans is a mix of signals some conveying speech and some conveying expression, it can be seen as a two part problem. If this mixed signal can be factored into its component parts of speech and expression, then separate models can be built to synthesise each signal, and the factorisation can be inverted to recombine these separate signals into novel expressive visual speech. Our research aims are therefore to discover a robust and reliable factorisation which is applicable to multiple data types and produces reproducible results, and use it to build generative models of expression.

In order to accomplish this, we rely on several well known statistical, algorithmic and machine learning techniques which will be covered in the rest of this chapter.

6.2 Principal Component Analysis

Central to this work is the technique of Principal Component Analysis (PCA) Pearson [1901]. Here PCA is used to reduce the dimensionality of complex point cloud data representations of facial geometry, and to build Active Appearance Models Cootes et al. [2001]. PCA allows the dimensionality of strongly correlated data to be reduced whilst retaining some predetermined amount of variation. Any example in a training set can be approximated by:

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}, \tag{6.1}$$

where $\mathbf{x}$ is the approximation, $\bar{\mathbf{x}}$ is the mean, $\mathbf{P} = (\mathbf{p}_1 \mid \mathbf{p}_2 \mid \cdots \mid \mathbf{p}_t)$ is the set of t orthogonal eigenvectors describing some predetermined proportion of the original variance, and $\mathbf{b}$ is a t dimensional vector given by:

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}). \tag{6.2}$$

Therefore it can be seen that $\mathbf{x}$ and $\mathbf{b}$ are equivalent. The process is carried out as follows Cootes [2000]:

- Compute the mean:

$$\bar{\mathbf{x}} = \frac{\mathbf{1}}{\mathbf{s}} \sum_{\mathbf{i=1}}^{\mathbf{s}} \mathbf{x_i}. \tag{6.3}$$

- Compute the covariance:

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{s-1}} \sum_{\mathbf{i=1}}^{\mathbf{s}} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^{\mathbf{T}}. \tag{6.4}$$

- Compute the eigenvectors $\mathbf{p_i}$ and the corresponding eigenvalues $\lambda_\mathbf{i}$, (sorted so that $\lambda_i \leq \lambda_{i+1}$).

- Each eigenvalue describes the variance about the mean of its corresponding eigenvector. Therefore calculate the total variance in the eigenvalues thus:

$$V_T = \sum_i \lambda_i. \tag{6.5}$$

- Now choose the t largest eigenvectors such that:

$$\sum_{i=1}^{t} \lambda_i \geq f_v V_T \tag{6.6}$$

  where f_v is the proportion of the original variance to be kept (typically 95-98%).

For certain types of data e.g. speech signals, it can be shown that a large proportion of the variance in a signal can be captured in a very few eigenvectors. In one of our PCA models, each example frame consisted of stacked coordinate data $\mathbf{x} = \{x_0, y_0, x_1, y_1, \ldots, x_{n-1}, y_{n-1}\}^T$, where $n = 51$. 150 training images of a male subject's face were labelled with these 51 landmarks marking prominent areas such as around the lips, eyes and nose (see Figure 6.4). After PCA was applied, plotting the values of the resulting eigenvalues yielded Figure 6.1. Note how the variance drops to near 0 after the 20^{th} eigenvalue. Therefore nearly all the variance was captured by the first 20 eigenvectors. By reducing the retained proportion of variance to 0.95, only 12 eigenvectors were required as shown in Figure 6.2. Since vector $\mathbf{b}$ is a t dimensional approximate encoding of vector $\mathbf{x}$ retaining nearly all the variance from the mean in only a fraction of the dimensionality, PCA can be

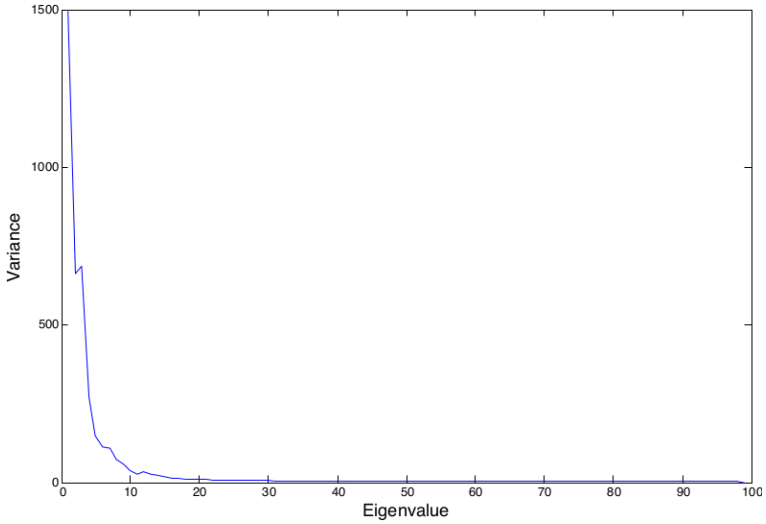


Figure 6.1 Plot of eigenvalues after PCA with 99.999% of variance remaining.

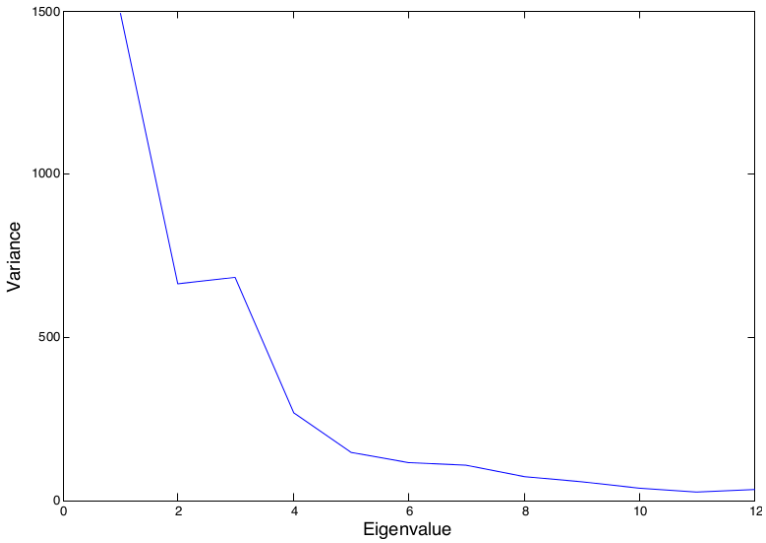


Figure 6.2 Plot of eigenvalues after PCA with 95% of variance remaining.

viewed in this case as a form of lossy data compression.

Figure 6.3 Example training images for an AAM displaying extreme facial poses

6.3 Active Appearance Models

Active Appearance Models (AAMs) are a standard way of encoding and compressing images in computer vision. As described in Cootes et al. [2001], AAMs are an amalgamation of two different PCA based models. One is a model describing the distribution of pixel intensities across a set of training images. The other is a point distribution model, encoding the variation of shapes appearing in the training images. An AAM must be trained by marking points (or landmarks) on a set of training images. The training images should be selected in order to capture the extremes of variation encountered across the entire set. See Figure 6.3 for some example training frames. In the case of facial images (see Figure 6.4), one must mark prominent feature outlines such as the outline of the face, and those of the lips (inner and outer), the nose, eyes, eyebrows etc. During the training phase,

Figure 6.4 Labelling of an image with landmarks, prior to training an AAM. Green markers show primary points which should always be in correspondence i.e. should always mark the same part of the face. Red markers are those whose position is simply interpolated by the position of the green markers.

the landmarks for each training image are aligned using procrustes analysis and then stacked such that $\mathbf{s} = \mathbf{s}(x_1, y_1, x_2, y_2 \ldots, x_n, y_n)$. A compact representation (as descrbed in Section 6.2) of the distribution of landmarks across the training set can be described by:

$$\mathbf{s} \approx \bar{\mathbf{s}} + \mathbf{S}\mathbf{b}_s, \tag{6.7}$$

where $\mathbf{s}$ is the concatenated vector of landmarks, $\bar{\mathbf{s}}$ is the mean vector of concatenated landmarks, and $\mathbf{S}$ is the set of n orthonormal basis vectors describing some given amount of variation. $\mathbf{b_s}$ is the low dimensional vector describing each basis vector's contribution to $\mathbf{s}$ and is defined as:

$$\mathbf{b}_s = \mathbf{S}^T(\mathbf{s} - \bar{\mathbf{s}}). \tag{6.8}$$

Figure 6.5 shows the result of taking the mean lip shape and adjusting various modes by $\pm$ 3 standard deviations from the mean of samples seen in the training set. The appearance component of the AAM is modelled in a similar way. Each training image is warped from its landmarked shape $\mathbf{s}$ to the mean shape $\bar{\mathbf{s}}$, thus creating a shape normalised image. The pixel intensities for the colour planes are then concatenated and a PCA model is trained such that:

$$\mathbf{a} \approx \bar{\mathbf{a}} + \mathbf{A}\mathbf{b}_a, \tag{6.9}$$

where $\mathbf{a}$ is the shape normalised image, $\bar{\mathbf{a}}$ is the mean normalised image, $\mathbf{A}$ is the set of n orthogonal basis vectors describing a predetermined proportion of variance and $\mathbf{b}_a$ is a k dimensional vector describing the contribution of each basis vector to the encoded appearance described by:

$$\mathbf{b}_a = \mathbf{A}^T(\mathbf{a} - \bar{\mathbf{a}}). \tag{6.10}$$

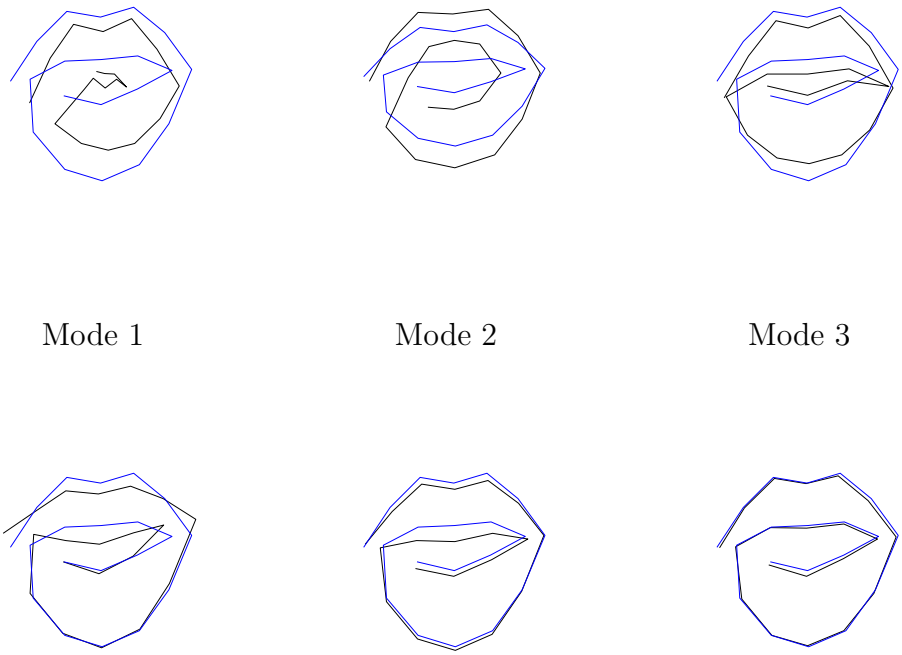


Figure 6.5 The first six shape modes showing the range of shapes captured in ± 3 standard deviations from the mean. Mode 1 appears to capture puckering of the lips as in a kiss. Mode 2 appears to capture an /oo/ shape as in "book". Mode 3 appears to shows a general parting of the lips, whereas mode 4 clearly shows a smile as does mode 5. Mode 6 shows little variation at all and could arguably have been left out of the model.

Figure 6.6 The first 2 appearance modes show the range of appearance images (warped to the mean shape) captured in ± 3 standard deviations from the mean appearance of the training data.

Figure 6.6 shows the result of adjusting the first two modes of appearance by ± 3 standard deviations from the mean. Concatenating the results of Equations 6.8 and 6.10 gives:

$$\mathbf{b} = \begin{bmatrix} \mathbf{b_s} \\ \mathbf{b_a} \end{bmatrix}, \tag{6.11}$$

a compact representation of shape and appearance in PCA space. The landmarks for unseen images, which are required for computing the parameters b, can be obtained using one of many AAM fitting algorithms. In this work we use the inverse compositional fitting algorithm Matthews and Baker [2004]. For convenience we project the shape and appearance components of the AAM into a combined space by first concatenating the shape and appearance parameters:

$$\mathbf{b}_{sa} = [w\mathbf{b}_s' \ \mathbf{b}_a']', \tag{6.12}$$

where w is used to weight the shape parameters so the overall energy in the shape and the appearance are equal. The weight is computed using:

$$w = \sqrt{\frac{t_a}{t_s}}, \tag{6.13}$$

where t_a is the trace of the covariance matrix for the appearance and t_s is the trace of the covariance matrix for the shape. Applying PCA to the concatenated vectors provides the combined model of shape and appearance:

$$\mathbf{b} = \mathbf{P}'\mathbf{b}_{sa}. \tag{6.14}$$

Once trained, the AAM can be used as a generative model. Given the parameters in Equation 6.11, the shape component can be projected from PCA space into the equivalent landmarks. The appearance component can likewise be projected into the equivalent mean normalised appearance image. This appearance image can then be warped from the mean shape to the shape of the landmarks. In this manner it is possible to create any combination of shape and appearance observed in the training data. Synthesis is achieved by a generative model creating appropriate parameters in the form of Equation 6.11.

6.4 Independent Component Analysis

As described in Hyvarinen [1997], the goal of Independent Component Analysis (ICA) is to express a set of random variables as a linear combination of statistically independent variables. Independence in this case means that the value of one variable tells us nothing about another variable. Whereas the related technique of PCA tries to find linear combinations where the data co-vary as much as possible, ICA tries to find linear combinations of the data which specifically do not co-vary. Consider the following Equation:

$$\mathbf{x} = \mathbf{Q}\gamma, \tag{6.15}$$

where $\mathbf{x}$ is a random multivariate normally distributed variable, and γ is a set of component signals which combine to make $\mathbf{x}$. $\mathbf{Q}$ then is an unknown $(m \times n)$ matrix, called the mixing matrix. ICA provides a framework for estimating the mixing matrix $\mathbf{Q}$ using only the data observed in $\mathbf{x}$ such that the independent components in a mixed signal can be calculated using:

$$\gamma = \mathbf{W}\mathbf{x}, \tag{6.16}$$

where γ are the independent components and $\mathbf{W}$ is the pseudo inverse of $\mathbf{Q}$. To accomplish this, ICA makes some assumptions about the natural world. Firstly that signals measured from discreet physical entities should be statistically independent. Secondly, such signals should have non-Gaussian distributions. This leads to three interesting observations:

- Source signals from different physical processes are independent, however, mixes of such signals are not independent.

- According to the central limit theory, the distribution of the sum of independent random variables should tend towards Gaussian. Therefore if independent non-Guassian signals are mixed, their sum should tend towards a Gaussian distribution.

- The temporal complexity of any mixed signal should be greater than the complexity of its simplest source signal.

These observations form the basis of ICA theory. Independent sources are found in a mix of data in the following ways:

- Maximisation of non-guassianity in the histograms of outputs.

- Minimisation of mutual information between outputs.

- Finding of the least complex signals in a mix.

ICA has a few notable drawbacks. The assumption made (that signals from different physical entities are independent) is arguable. There are certainly cases where this may not be entirely true. Let us examine for example foetal heart monitoring. The heart beats of the foetus and the mother are clearly from separate physical processes and are largely independent, but cannot be considered entirely independent since the physical condition of the mother will inevitably effect the heart beat of the foetus. Similarly, in the case of our work where ICA is used to estimate some signals representing expression and some representing speech, the signals are largely independent, but not completely. They are independent enough however for ICA to produce some useful estimates. PCA returns components based on the magnitude of the eigenvector's associated eigenvalues. That is, the first returned mode describes the axis of greatest variance, the second describes the axis of second greatest variance orthogonal to the first etc. Since ICA finds

the independent modes by maximising kurtosis and therefore non-Gaussianity, it cannot order these independent components by ranked variance. Therefore ICA makes no guarantees of the returned ordering or magnitudes of independent signals and indeed is unable to deduce the real number of independent signals in a mix instead relying on being instructed how many signals to find.

In this work we use the open source implementation of FastICA Gavert et al. [2005]. Figures 6.7, 6.8, 6.9 and 6.10 show a simple example of its operation. Three observations of a mix of three sine waves at different frequencies is shown in Figure 6.7. FastICA is able to separate this wave into its components seen Figures 6.8, 6.9 and 6.10.

Figure 6.11 shows how the FastICA algorithm transformed the first four PCA modes representing real expressive visual speech into independent modes by kurtosis maximisation. As previously said, signals in nature tend to a non-Gaussian distribution. Therefore by finding transforms of the data which yield non-Gaussian distributions, we tend to find independent signals. The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3 Mathworks [2015]. The actual kurtosis values for these modes are shown in Table 6.1. The mean kurtosis over all the modes for the PCA and ICA features is 3.8946 and 6.5290 respectively. The Matlab kurtosis function was used and is defined by:

$$k = \frac{E(x-\mu)^4}{\sigma^4}, \tag{6.17}$$

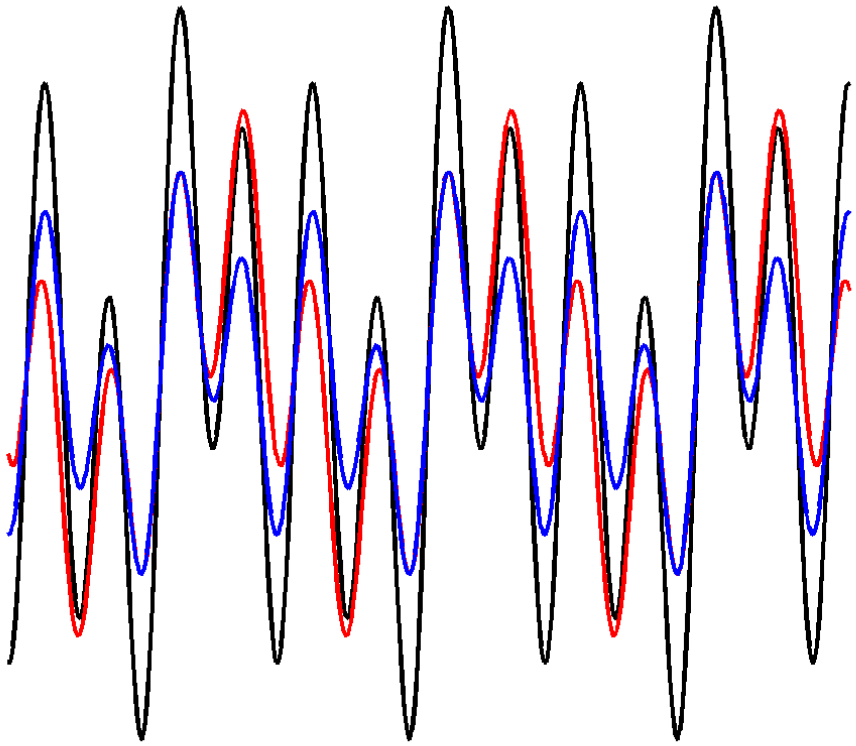


Figure 6.7 ICA Demonstration: Three mixes of the same three sine waves each with a different frequency. Each mix has a different contribution from each wave.

Mode / Component	Kurtosis in PCA	Kurtosis in ICA
1	2.9916	1.5625
2	2.7284	18.3728
3	2.7223	6.3689
4	3.8124	11.8785

Table 6.1 Kurtosis for principal components and independent components.

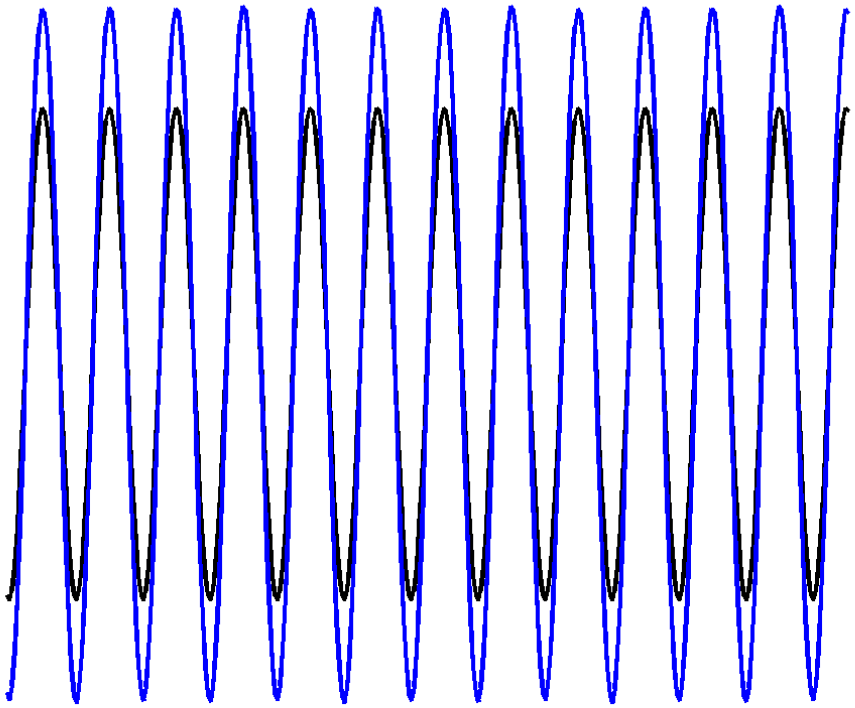


Figure 6.8 ICA Demonstration: ICA recovered sine wave (blue) against ground truth (black).

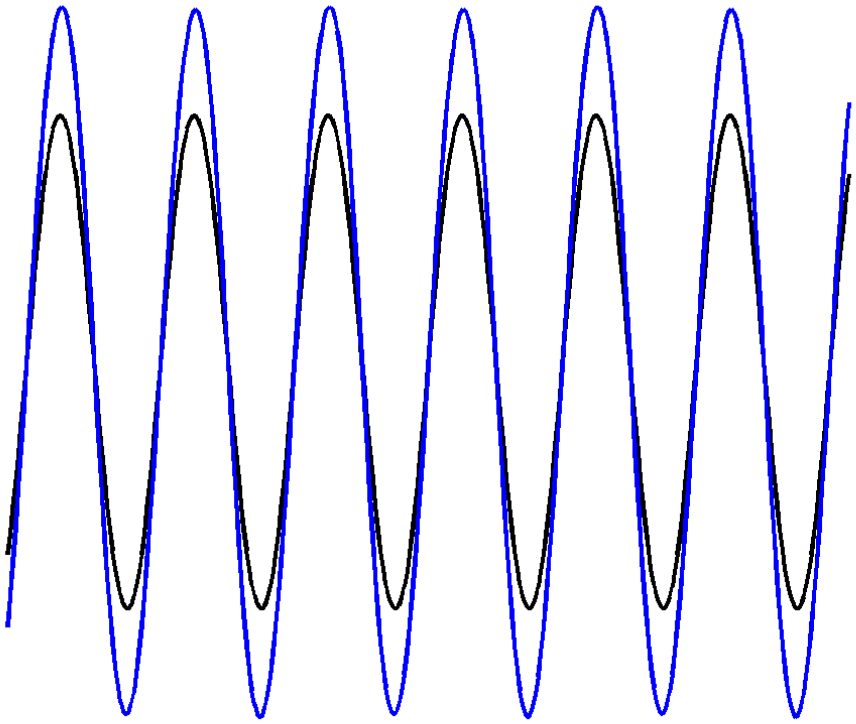


Figure 6.9 ICA Demonstration: ICA recovered sine wave (blue) against ground truth (black).

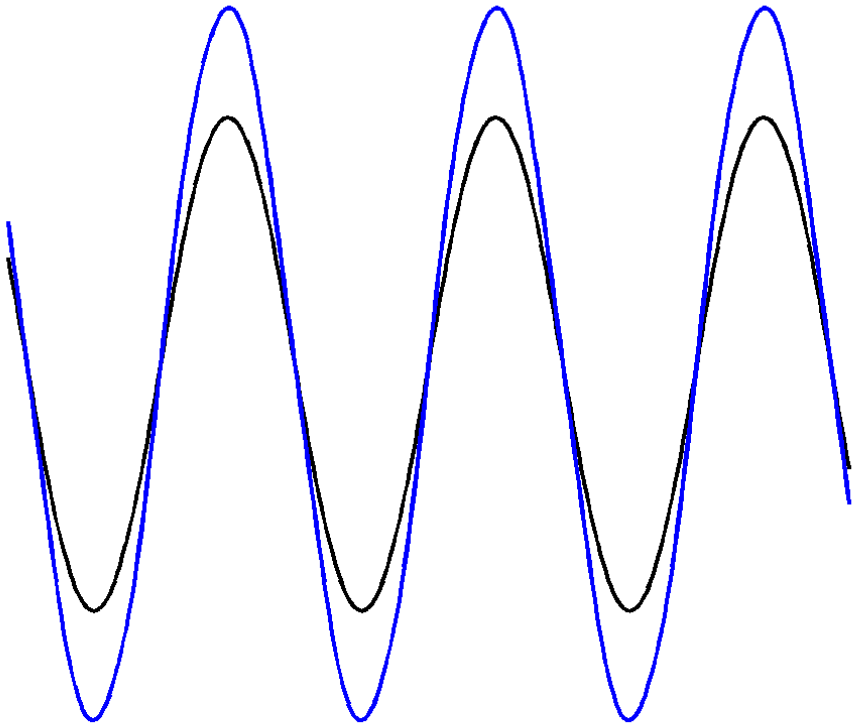


Figure 6.10 ICA Demonstration: ICA recovered sine wave (blue) against ground truth (black).

where μ is the mean of x, σ is the standard deviation of x and $E(t)$ is the expected value of the quantity t.

6.5 Retargeting of Animation using Scattered Data Interpolation

Scattered Data Interpolation (SDI) is the process of constructing new data points given the observations of a set of known data points, where the known points have no particular structure or pattern. Since the process needs no knowledge of how the known points are organised, it provides a convenient method of interpolation for irregular geometry such as that seen in computer models of the face. As presented in Pighin et al. [2006], the movement of a set of landmarks as described in Section 6.3 can be transferred to another model, so long as the configuration of both model and landmarks is not too dissimilar. It is entirely possible to transfer the movements of 2D landmarks to a 3D model. Firstly either the landmarks or the model must be aligned using affine parameters to eliminate differences in scale, translation and rotation. Then pairs of corresponding vertices between the landmarks and model are identified such as those defining the corners of the mouth, line of the lips, corners of the eyes etc. The movement of the landmarks from frame to frame is directly transferred to the corresponding (constrained) model vertices and the movement of the remaining (unconstrained) vertices is interpolated. More formally, a set of known displacements is described thus:

$$\mathbf{u}_i = \tau_i - \tau_i^0 \tag{6.18}$$

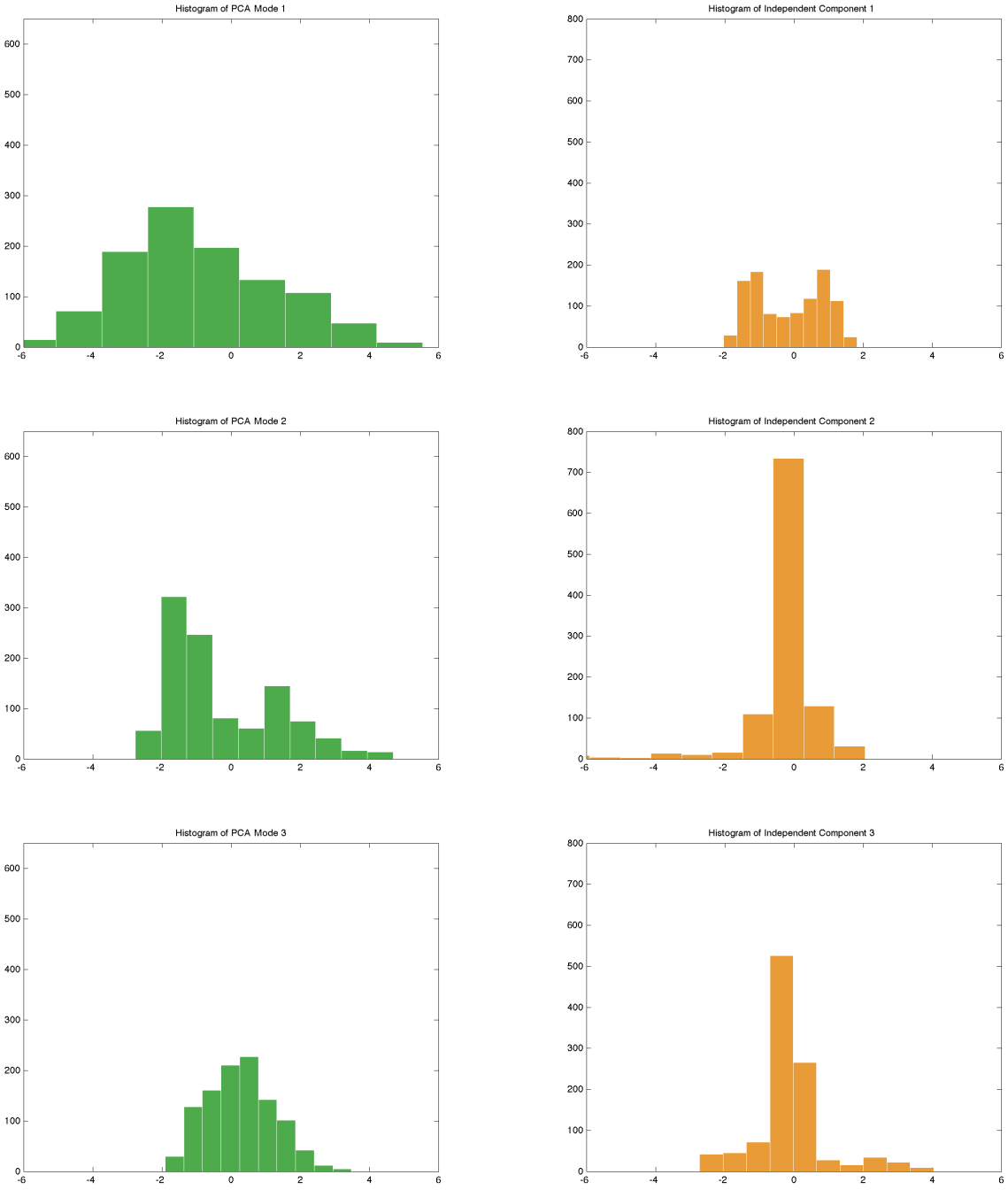


Figure 6.11 Histograms of first four PCA modes and independent components of expressive speech data before and after ICA transformation.

where τ_i is the new position and τ_i^0 is the original position. The task then is to construct a smooth function fitted to the known data:

$$\mathbf{u}_i = \mathbf{f}(\tau_\mathbf{i}) \tag{6.19}$$

which gives $\mathbf{u}_j$ the displacement for every unconstrained vertex thus:

$$\mathbf{u}_j = \mathbf{f}(\tau_\mathbf{j}) \tag{6.20}$$

In this work, the interpolant is a set of radial basis functions (RBF) of the form:

$$\mathbf{f}(\tau) = \sum_i \mathbf{c}_i \phi(||\tau - \tau_i||) \tag{6.21}$$

where $\mathbf{c}$ is a vector of weights describing each RBF's contribution to the interpolation, τ is the current unconstrained point to be interpolated and $\tau_\mathbf{i}$ is the current contained point. After experimentation, we define $\phi(\mathbf{r}) = e^{-r/(3/32)}$ as this gives a smooth interpolation, particularly across areas where the constrained points are sparse.

6.6 Nelder-Mead Downhill Simplex Optimisation

The Nelder-Mead downhill simplex optimisation Nelder and Mead [1965a], is a commonly used optimisation technique which is useful for solving for functions for which the derivatives are unknown. Since it is heuristic in nature, it is not guaranteed to always converge to a stable/optimal solution, but works well when a close approximation of the optimal solution is good enough and is a fair compromise in cases where an exhaustive method would be computationally infeasible. In the Nelder-Mead algorithm, a multivariate function with n inputs is represented by a

simplex in k dimensional space whose $k+1$ vertices represent different inputs to the function. At each iteration, the function is sampled at each of the simplex's vertices and the vertex with the worst error is replaced with an improved position. More precisely, at each iteration one of four actions takes place:

- Start of iteration:

  The simplex is made of $k+1$ vertices where n is the dimensionality of the function to be optimised. In Figure 6.12 the points of the triangle represent the inputs to a two dimensional function and the red dot represents the optimal input.

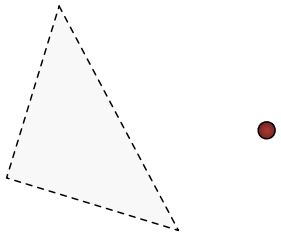


Figure 6.12 Nelder-Mead: Initial State

- Reflection:

  The vertex with the worst error is reflected through the centroid of the other vertices and the error at the new position is sampled, show in Figure 6.13.

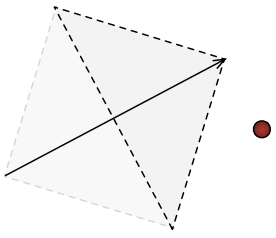


Figure 6.13 Nelder-Mead: Reflection

- Expansion:

  If the error of the function sampled at the reflected vertex is smaller than the error at the original worst point, then it is likely that interesting values lie along this reflected axis, so the simplex is expanded further as shown in Figure 6.14. If the expanded vertex is better than the reflected vertex, the current worst vertex is replaced by the expanded vertex. Otherwise, if the reflected vertex is better than the expanded vertex, the worst vertex is replaced by the reflected vertex. In either case, this completes the current iteration.

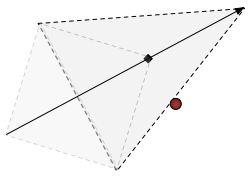


Figure 6.14 Nelder-Mead: Expansion

- Contraction: If the reflected vertex is worse than the current worst vertex, then the minimised solution must lie either within the simplex, or within its mirror image. Therefore two new vertices are calculated, one in the middle of the simplex and one in the middle of the simplex's mirror image as shown in Figure 6.15. If the errors at either one of these new contracted vertices is less than the worst vertex, the worst vertex is replaced by that with the smaller error. This ends the iteration.

- Reduction:

  If neither of the errors at the vertices calculated in the contraction stage are better than the error at the worst vertex, then all vertices other than the

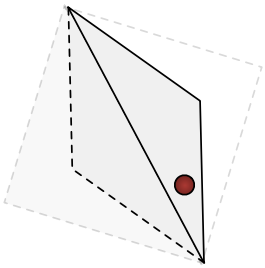


Figure 6.15 Nelder-Mead: Contraction

best are divided in half, thus collapsing the simplex in on itself and towards the optimised solution as shown in Figure 6.16. This ends the iteration.

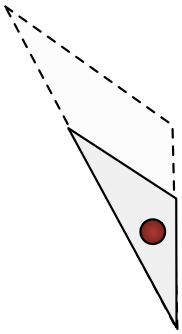


Figure 6.16 Nelder-Mead: Reduction

This process continues until either a pre-specified number of iterations has occurred, or some minimum error threshold is achieved, therefore finding an approximation of an optimised solution to the multivariate function.

6.7 Dynamic Time Warping

Dynamic Time warping Rabiner et al. [1978] has long been used in the speech recognition community for the tasks of alignment and recognition. More recently the technique has been used in the information retrieval community for tasks such as music classification (the basis for technologies such as Shazam) Müller [2007] Barton and Inghelbrecht [1999], and in the machine learning community for various classification tasks Ratanamahatana and Keogh [2004]; Keogh and Ratanamahatana [2005]; Lines et al. [2011]. In this work, we use the algorithm for aligning two sequences of different lengths, namely two sequences which have the same phonetic content but spoken in different styles e.g. happy and neutral. All expressive sequences were warped to their neutral equivalent. The path was calculated between Mel Frequency Cepstral Coefficient features of the sequences audio tracks, which were produced by HTK's HCOPY function Young and Woodland [2009]. The output from the DTW algorithm was then quantised from the audio frame rate of 44,100 Hz to the video frame rate of 25 Hz to produce frame sequence numbers required to align the two video sequences corresponding to the warped audio tracks.

6.7.1 Overview of DTW

Suppose there are two sequences S_x and S_y of lengths χ and ψ respectively such that:

$$S_x = (S_{x1}, S_{x2}, S_{x3}, \ldots S_{x\chi}) \tag{6.22}$$

$$S_y = (S_{y1}, S_{y2}, S_{y3}, \ldots S_{y\psi}) \tag{6.23}$$

The sequences may be aligned by constructing $\mathbf{M}$, a $(\chi \times \psi)$ matrix, where $\mathbf{M}(i, j)$ contains the euclidean distance between the two points $d(S_{xi}, S_{yj}) = (S_{xi} - S_{yj})^2$.

A warping path is defined which maps S_x and S_y by the minimising the distance through the matrix $\mathbf{M}$ such that:

$$DTW(S_x, S_y) = min\left\{ \sqrt{\sum_{g=1}^{G} \phi_g} \right. \tag{6.24}$$

where ϕ_g is the g_{th} element of the warping path. The warping path is calculated at each step as:

$$\phi(i,j) = d(S_x, S_y) + min\{M(i-1, j-1), M(i-1, j), M(i, j-1)\} \tag{6.25}$$

Figure 6.17 shows an example DTW warping path through a matrix of euclidean

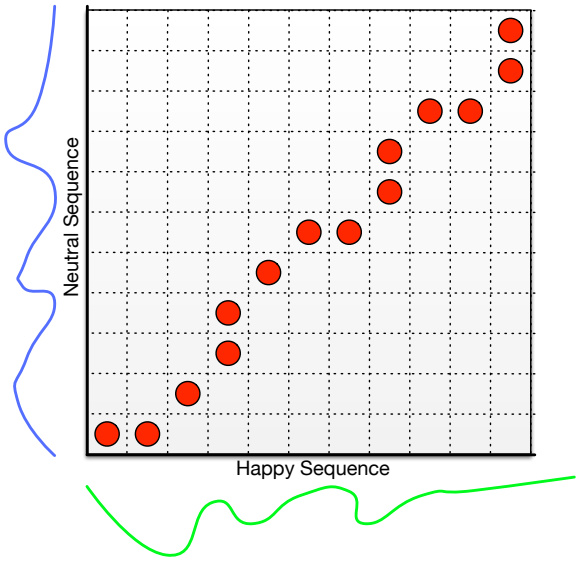


Figure 6.17 A DTW warping path through the matrix of euclidean distances. Outputting the frames for each sequence indicated by the warping path yields two new sequences which are time aligned.

distances between a happy and a neutral sequence. By concatenating the frames for each sequence as indicated by the warping path (in red), two new sequences

are obtained which are time synchronised and the same length.

6.8 Unsupervised Learning for Speech Motion Editing

The research described in this thesis is heavily influenced by the work of Cao et al. [2003, 2005]. In their work they describe an initial attempt to factorise speech and emotion, theorising that expressive speech is a linear combination of these independent components. What follows is a review of their initial findings.

Firstly, they captured a corpus of training data. A male actor was recorded issuing an unknown number of sentences, each in four different emotional styles (happy, angry, sad and frustrated). The movement of the actor's face was tracked using a Vicon8 optical motion capture system. Using 109 facial markers produced a fairly sparse mesh representation of the face, but retained enough detail to project the lip movement and emotional content of each sentence. The sample rate was 120 fps.

The coordinates of the facial markers were stacked such that:

$$\mathbf{x} = \mathbf{x}(x_1, y_1, x_2, y_2 \ldots x_n, y_n)^T \tag{6.26}$$

They then applied PCA to the stacked coordinates before projecting into ICA space thus:

$$\mathbf{x} = E\{\mathbf{x}\} + \mathbf{PAu} \tag{6.27}$$

where $E\{\mathbf{x}\}$ is the expectation of $\mathbf{x}$, $\mathbf{P}$ is the set of orthonormal basis vectors obtained by applying PCA, $\mathbf{A}$ is the mixing matrix obtained during ICA training and $\mathbf{u}$ are the independent components. Since ICA is unable to determine the

actual number of independent components, it is reliant upon the user to instruct the FastICA implementation how many independent components to return. Cao chose this to be equal to the number of PCA components and was therefore dependent on the remaining proportion of variance retained by the PCA process (typically between 95% and 98%).

ICA makes the assumption that a mixed signal is a linear combination of independent components and is able to estimate an un-mixing matrix to decompose such a signal into these components. Expressive speech is such a mixed signal, since the same utterance can be said with different emotions. The emotional content and the speech content can be viewed as approximately independent from each other. Therefore it is reasonable to hypothesise that ICA should be able to decompose an expressive speech signal into its independent components, speech and expression. To test this hypothesis they chose two utterances of the same sentence with two different emotions (happy and frustrated) and aligned them using a DTW algorithm. These were then used to train an ICA model and using this model, were projected into ICA space giving a pair of corresponding independent components $(\mathbf{u}, \mathbf{v})$. They computed the root mean squared (RMS) error between these components using:

$$d_{emotion,j} = \frac{1}{\sum q^i}(\sum_{k=1}^{q^i}(\mathbf{u}_j^i - \mathbf{v}_j^i)^2)^{\frac{1}{2}} \tag{6.28}$$

where q^i the number of samples in each independent component. Since the two sentences share the same speech content, and have been time aligned, if a mode represents speech, the error in equation 6.28 ought to be small. However if the mode represents emotion then the error should be large since this represents the difference between the sentences. Their results can be seen in figure 6.18.

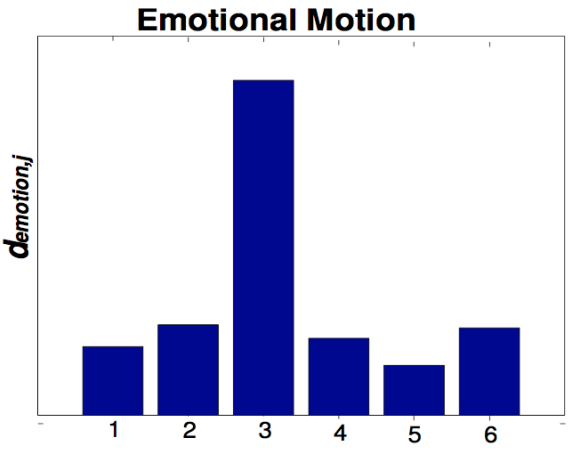


Figure 6.18 Cao's results from taking the RMS error between two time aligned expressive sentences in ICA space (copied from Cao et al. [2003]).

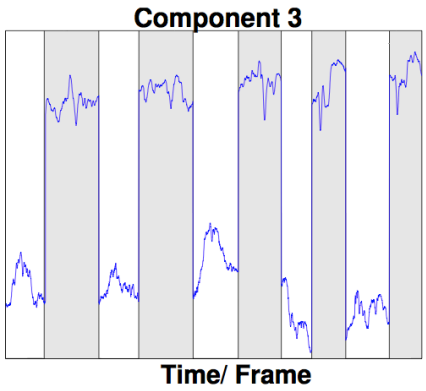


Figure 6.19 Swapping every other ICA element on the suspected emotional mode with the corresponding element from the other sentence (copied from Cao et al. [2003]).

Note how mode three appears to contain the largest error. In order to confirm that this mode contains emotional content, they took one of the sentences in ICA space swapped every other element of this mode with the corresponding element in the same mode from the other sentence in ICA space as seen in figure 6.19. Upon inverting the ICA transform, projecting from PCA space into coordinate space and

playing the frames back in sequential order, it was found that the configuration of points did indeed seem to show the face oscillating between the two expressions in the training set every other frame.

6.9 Summary

This chapter has broadly outlined the main families of techniques used in the rest of the work presented in this thesis. Principal Component Analysis is used extensively to project point cloud data into a low dimensional representation. This is essential to make computation feasible and timely. The related statistical technique of Active Appearance Modelling utilises PCA to perform an analogous task of projecting image data into a low dimensional representation. Independent Component Analysis, a form of blind-source separation, is used to split mixed data into independent modalities (expressive visual speech into speech and expression). Scattered Data Interpolation is used to warp template meshes onto novel, unseen geometric representations of the face. The Nelder-Mead downhill simplex optimisation is used to solve for multi-variate functions whose the derivatives are unknown, and Dynamic Time Warping is used to align sequences of different lengths by similarity. The next two chapters demonstrate how by combining various use cases of the techniques described in this section, we are able to modulate an existing neutral speech signal with a learned expression signal to produce realistic and smoothly varying expressive visual speech.

Chapter 7

Simple Model Modulation

7.1 Introduction

As was talked about previously, Cao et al. [2005, 2003] reported promising initial results detailing the factorisation of a statistical representation of expressive visual speech into linearly independent components (speech and expression). In Cao et al. [2003], the expressive style from one sentence was copied into another expressive sentence by training an ICA model on the data in the two sentences, allowing for the identification of modes representing expression, and then copying the values from these modes to the corresponding modes of another sentence with different expressive style. Whilst interesting, it is not clear how this is useful to an animator in the movie or computer games industries. Firstly the sentences manipulated were the same as those used in ICA training, meaning that the technique is not shown to work on unseen data and cannot therefore be said to be a generalised method. Secondly, simply changing the expressive style from one to another is not a task an animator is likely to want to perform. Thirdly, since one ICA model was necessary for each pair of sentences, the number of ICA models required would

grow as the square of the number of expressions one wishes to approximate. It was also reported that the expressive components of two different ICA models could be added together in order to create a mix of expressive styles. However, no results were provided, and there is no guarantee that such an addition would create plausible facial poses.

The rest of this chapter is structured as follows: firstly we discuss why ICA was chosen as a factorisation method and then describe our replication of results reported in Cao et al. [2003], followed by a description of work to generalise the method by applying it to unseen data. We then demonstrate a technique which would make the method of practical use to animators. We show how the original method can be improved so the model complexity goes from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ as a function of the number of expressions. Finally we present the results of an evaluative turing test and discuss the significance and limitations of this result.

7.2 Justification for ICA

There are several techniques which can be used for multivariate data factorisation such as Non-negative Matrix Factorisation, Multi-linear subspace learning or Tensor model learning. Some of these will be discussed. However, our first attempt was a novel Principal Component Analysis based approach. A PCA model was trained on neutral only visual speech features. Then frames of expressive speech were projected onto this neutral only model, the idea being that it would be able roughly approximate the expressive speech but using only the neutral variance on which it was trained. This neutral approximation could then be subtracted from the expressive, to create an expressive residual. The expressive residual would then be used as features to train some sort of generative model. The technique worked well for certain expressive styles. Sadness and Anger were particularly well fac-

torised. However when it came to Happiness, the technique failed. Whereas with the Sadness and Anger, the neutral approximations had simply looked like normal neutral speech, the neutral approximations of Happiness had a strange grimace type expression. This is probably because Happiness is an expression which requires mouth movement thus the influence of speech (moving of the lips and jaw), and movements associated with Happiness (curling of the lips) are non-orthogonal. In contrast, the movements associated with Sadness and Anger tend to be raising of the eyes and eyebrows, and furrowing of the brow and will therefore not effect the mouth movements produced in speech.

Non-negative Matrix Factorisation (NNMF) is a statistical technique which has been used amongst other things to separate mixed audio of multiple speakers Schmidt and Olsson [2006]. This technique seeks to represent data as sparse linear combinations of basis vectors. These basis vectors must be orthogonal and non-negative. Additionally there must be discrete classes from which to learn these basis vectors (or dictionaries) e.g. a large training corpus of phonetically discrete audio segments. In the case of our work, we have zero mean PCA coefficients as features meaning that there is lots of negative data. As has just been mentioned, speech and expression are not orthogonal, so finding orthogonal basis vectors will at best be sub-optimal. There is also not necasarilly a clean separation between expressions meaning that training the dictionaries will be problematic as discrete expression classes will probably not exist with real data.

Chuang et al. [2002] use bilinear subspace learning. By using a Singular Value Decomposition (SVD) a number of weighting matrices are estimated and can be used to combine speech and expression. However, it is stated that the estimation is difficult and therefore as a computational simplification, during training either the expression or the speech is held constant. Although this makes a good approximation of expressive speech, it fails to account for the dynamic interactions

occurring between speech and expression. There are multiple weight matrices required meaning that calculation is further complicated. Additionally a model is trained for every frame, meaning that the technique is unlikely to work in real-time. Vlasic et al. [2006] describes a similar technique based on multilinear algebra and N-mode SVD. The technique looks promising, but as it models identity as well as expression and viseme, it is overly complicated for our purposes.

ICA then, was tried next as it is completely unsupervised, requiring no a priori knowledge of the data. Therefore the discrete classes required in NNMF to train dictionaries of basis vectors are not required. Being unsupervised, ICA simply returns components based on the underlying structure of the data as discovered by kurtosis. ICA also handles negative input matrices unlike NNMF. ICA returns a single model which is general enough for matrix factorisation and multiplication of unseen visual speech data, therefore computationally simplifying the synthesis stage of an animation technique by not requiring a new model for each frame. Additionally, linear transformations by the same matrix are guaranteed to produce smooth concatenations of output frames given that the input features themselves are smooth. This cannot be said if the transformation matrix is retrained for each frame.

7.3 Reproduction of Cao's work

To reproduce Cao's original work we use our own dataset (see Section 4.2). This is because the BiWi corpus described in 4.1 does not contain the same sentence spoken in different styles, rather, each sentence only appears in its own style and neutral. As mentioned in Section 6.8, the purpose of this reproduction is to show that the expressive style from one sentence can be transferred to another sentence, where the style is different but the speech content is the same.

Firstly, a PCA model was trained using the features output by our AAM tracker using Equations 6.12, 6.13 and 6.14. Therefore the AAM shape and appearance features were projected into a single space representing combined shape and appearance. If this step is not preformed, the ICA process tends to separate the shape and the appearance into different modes (since these are largely independent from one another). The shape features were first normalised in order that they did not dominate the appearance features (since the shape co-ordinate values output by the tracker vary from 100-850 whereas the pixel intensities vary from 0-255). Then two sequences of the same utterance, but in different styles (happy and angry, happy and surprised or angry and surprised etc.) were projected onto the principal components of the mixed PCA model thus:

$$\mathbf{b} = \mathbf{P}^T([w\mathbf{b_s}\mathbf{b_a}] - \overline{[w\mathbf{b_s}\mathbf{b_a}]}) \tag{7.1}$$

where $\mathbf{P}$ is the set of orthonormal basis vectors describing 90% of the variance in the scaled AAM feature vectors, $\mathbf{b}_s$ are the shape features as described in Equation 6.8 and are chosen from a random selection of held out training frames, and $\mathbf{b}_a$ are the appearance features described in equation 6.10 from the corresponding held out frames. w is the square root of the ratio between the traces of the covariance matrices of $\mathbf{b}_s$ and $\mathbf{b}_a$ as shown in equation 6.13. The resultant features for the happy and angry sentences were then time aligned using dynamic time warping. At this stage an ICA model is trained. As mentioned previously, ICA is simply a linear transform the general form of which is:

$$\mathbf{x} = \mathbf{Q}\gamma, \tag{7.2}$$

where $\mathbf{x}$ is a linear combination of independent signals, $\mathbf{Q}$ is the mixing matrix and γ is the independent signals.

To perform this task, we used the open source FastICA algorithm Gavert et al. [2005]. The AAM features corresponding to the two expressive time aligned styles were arranged into a matrix thus:

$$\mathbf{b} = \begin{vmatrix} b_a^{11} & \dots & b_a^{1n} & b_h^{11} & \dots & b_h^{1n} \\ b_a^{m1} & \dots & b_a^{mn} & b_h^{m1} & \dots & b_h^{mn} \end{vmatrix} \tag{7.3}$$

where b_a are the angry features, b_h are the happy features, n is the number of frames (or samples) and m is the number of principal components retained after the PCA transformation. FastICA uses these stacked features as input and returns the estimated mixing and unmixing matrices $\mathbf{A}$ and $\mathbf{W}$.

The time aligned PCA features were then projected into ICA space using the inverse of equation 7.2 thus:

$$\mathbf{s}_h = \mathbf{W}\mathbf{b}_h \tag{7.4}$$

$$\mathbf{s}_a = \mathbf{W}\mathbf{b}_a \tag{7.5}$$

where $\mathbf{W} = \mathbf{A}^{-1}$ is the pseudo inverse of $\mathbf{A}$, $\mathbf{s}_h$ and $\mathbf{b}_h$ are the independent components and AAM features respectively for the happy style sentence, and $\mathbf{s}_a$ and $\mathbf{b}_a$ are the independent components and AAM features respectively for the angry style sentence. Since the speech component of the two sentences is very similar (the phonemic content of the sentence is the same and they have been time aligned based on audio features), there should be ICA modes which contain very similar information. However, the modes which contain the expression information should be less similar. Since FastICA provides no ordering to the modes which are returned, to identify the potential expressive mode/s, the root mean squared (RMS) error is calculated between corresponding modes of the happy and angry sentence. If the RMS error is small between two modes, it is reasonable to assume

this is a speech mode since the speech content is largely the same. However if the RMS error is large between two such modes, then we assume that this is an expressive mode. The RMS error is calculated as:

$$\epsilon_i = \sqrt{\sum_{k=1}^{n} (\mathbf{s_h}(i,k) - \mathbf{s_a}(i,k))^2} \tag{7.6}$$

where ϵ_i is the RMS error between i^{th} independent components, $\mathbf{s_h}$ are the independent components of the happy sentence, and $\mathbf{s_a}$ are the independent components of the angry sentence. Figure 7.1 shows a plot of the RMS error between the two sentences. Plots of the trajectories of the independent components are shown in Figures 7.2 through 7.6. From Figure 7.1 it is clear that component one contains the largest RMS error. Examination of the trajectory of this mode for both sentences (Figure 7.2) indeed shows that the mode is quite different in the two sentences.

To reproduce the experiment in Cao et al. [2003], we take mode one of the happy independent components and copy over values from mode one from the angry independent components. This is done every other frame, therefore the new expressive component contains alternating happy and angry values. Figure 7.7 shows the output from the AAM model for a test sentence after projecting the ICA components back into AAM space and projecting these AAM features onto the AAM.

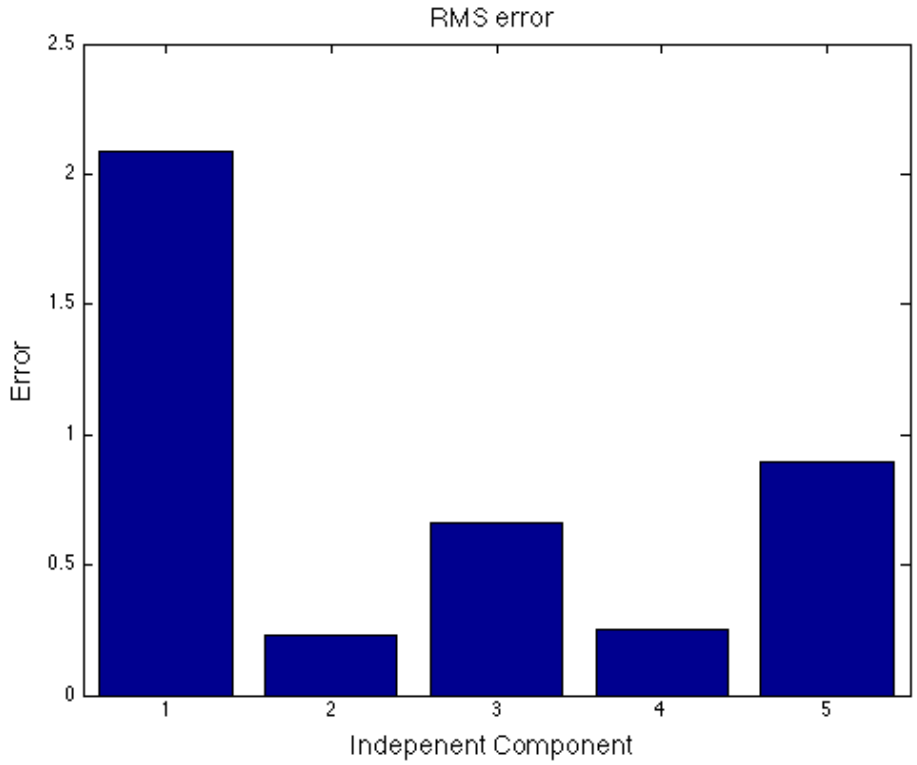


Figure 7.1 A plot showing the absolute RMS errors between independent components from two time aligned sentences, one happy and one angry.

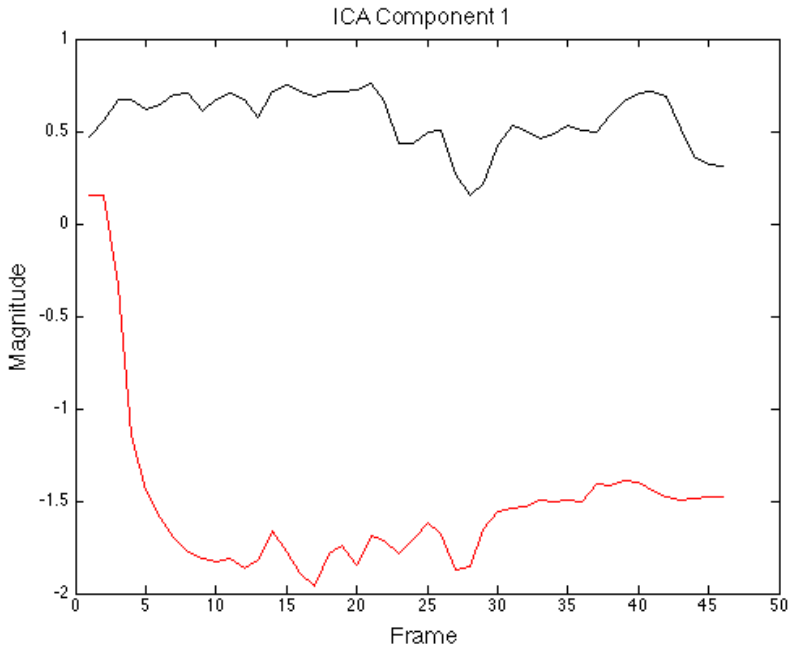


Figure 7.2 IC mode 1 for a particular ICA model: red is happy, black is angry.

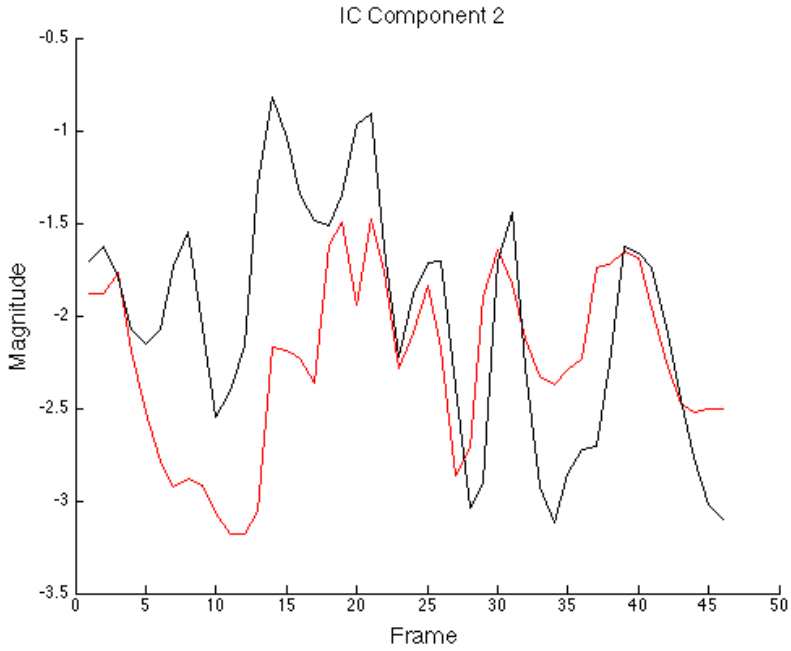


Figure 7.3 IC mode 2 for the same model: red is happy, black is angry.

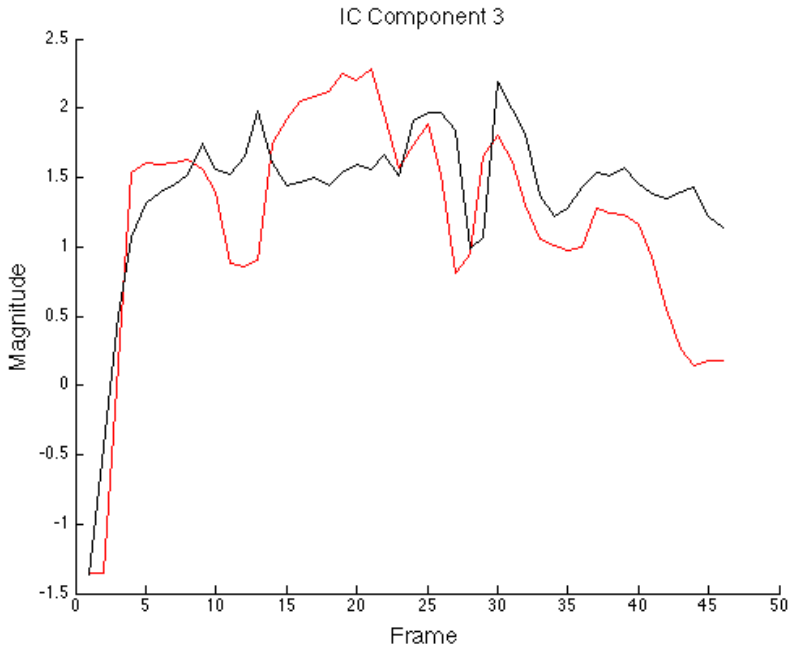


Figure 7.4 IC mode 3: red is happy, black is angry.

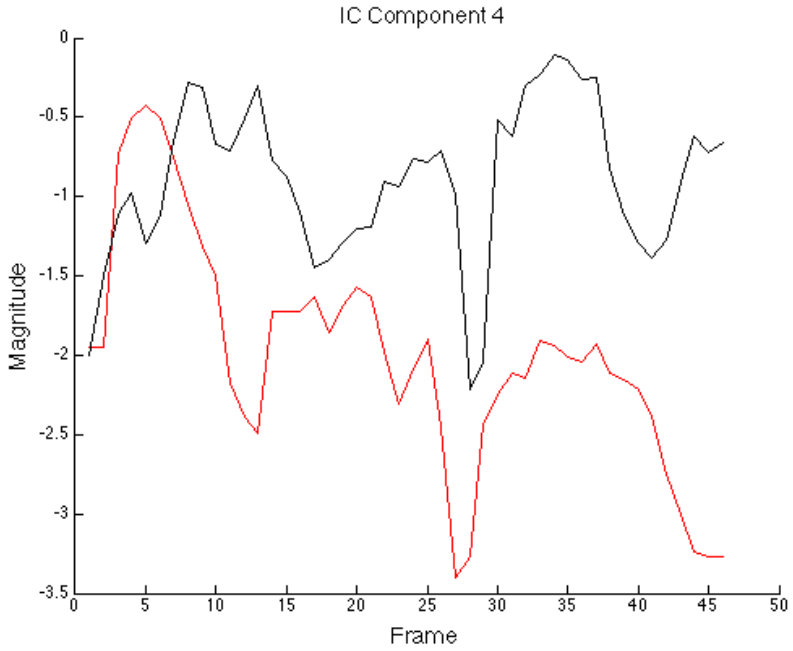


Figure 7.5 IC mode 4: red is happy, black is angry.

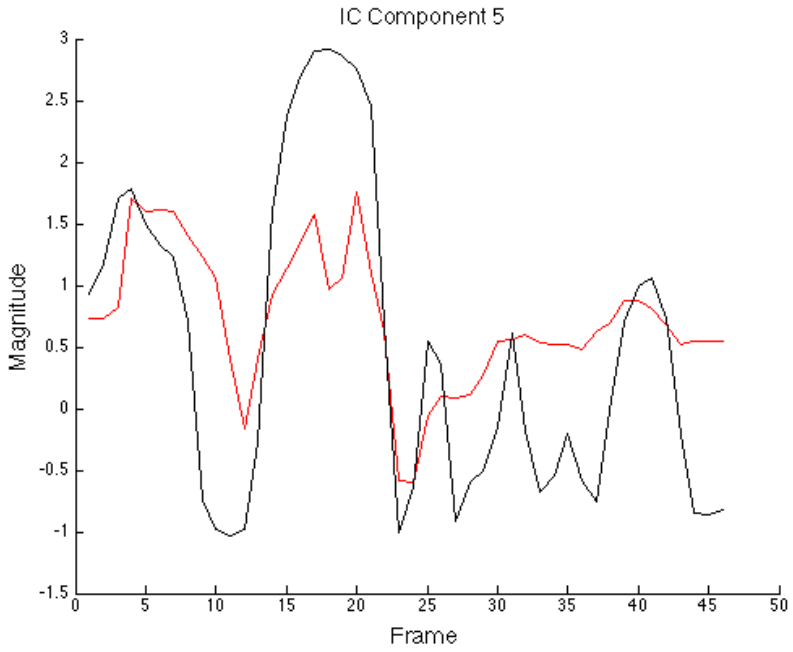


Figure 7.6 IC mode 5: red is happy, black is angry.

Figure 7.7 Frames from a happy sentence (top) and the same frames after replacing the expressive ICA mode values with those from an angry sentence (bottom), using the technique described in Cao et al. [2003].

7.4 Modulation of Neutral Speech

The steps in this section are summarised in Appendix A. What follows is a description of a novel method which attempts to generalise the contribution of Cao et al. [2003], and in so doing to form the basis of new tools which could be of potential interest to the animation industry. We show how it is possible to train an ICA model using only a few training examples. The model is able to separate expressive speech into its component parts of speech and expression. We then apply this model to unseen expressively neutral data, showing that it then becomes possible to manipulate this neutral data in ICA space in order to effectively modulate the mouth movements in the test sentence with an expressive style present in the training set.

When transforming parameters that encode neutral visual speech into those which encode expressive visual speech, the dynamics of the expression must appear natural and the mouth movements corresponding to speech must remain valid. If all of the assumptions of ICA held, some of the independent components would correspond exactly to speech and some exactly to expression. However, we have found that a 'clean' separation of the signals does not occur, and each component tends to represent both speech and expression to varying degrees. An additional problem is that in general ICA cannot guarantee an ordering to the returned independent components. Whereas PCA orders principal components based on eigenvalue, no such easy metric exits in ICA. Ordering components by change in kurtosis is one solution but is imperfect since certain distributions have a kurtosis value close to zero but are far from gaussian. Therefor it is usually left up to the user to make sense of the returned components. One must resort to empirical methods (such as RMS calculations or frequency domain analysis), to discriminate between those components which predominantly represent speech and those which predominantly

represent expression.

Figure 7.8 shows what we refer to as the *energy signatures* for neutral and expressive speech. The black bars in the figure represent the mean absolute deviation (MAD) in the components, computed using:

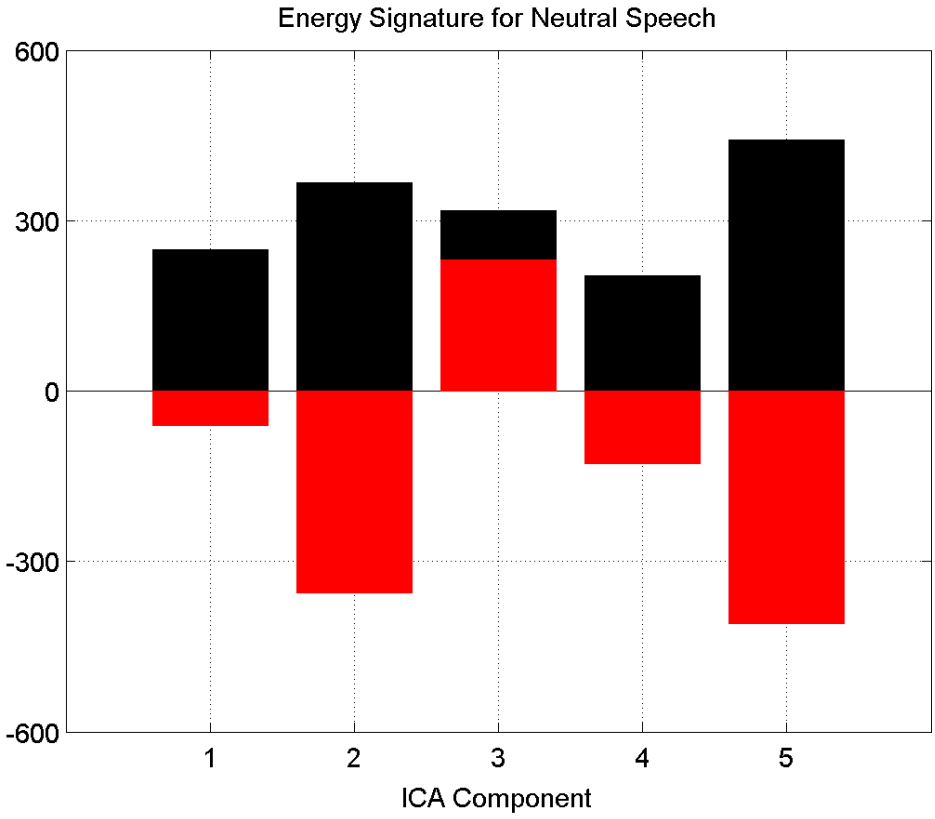
$$e^j = \sum_{t=1}^{k} |u^j(t)|^2, \tag{7.7}$$

where e^j represents the MAD across the j^{th} independent component and $u^j(t)$ represents the value of the j^{th} independent component at time t. The independent components themselves are derived from the combined PCA model of shape and appearance. The red bars in Figure 7.8 show the amplitude of the components, computed using:

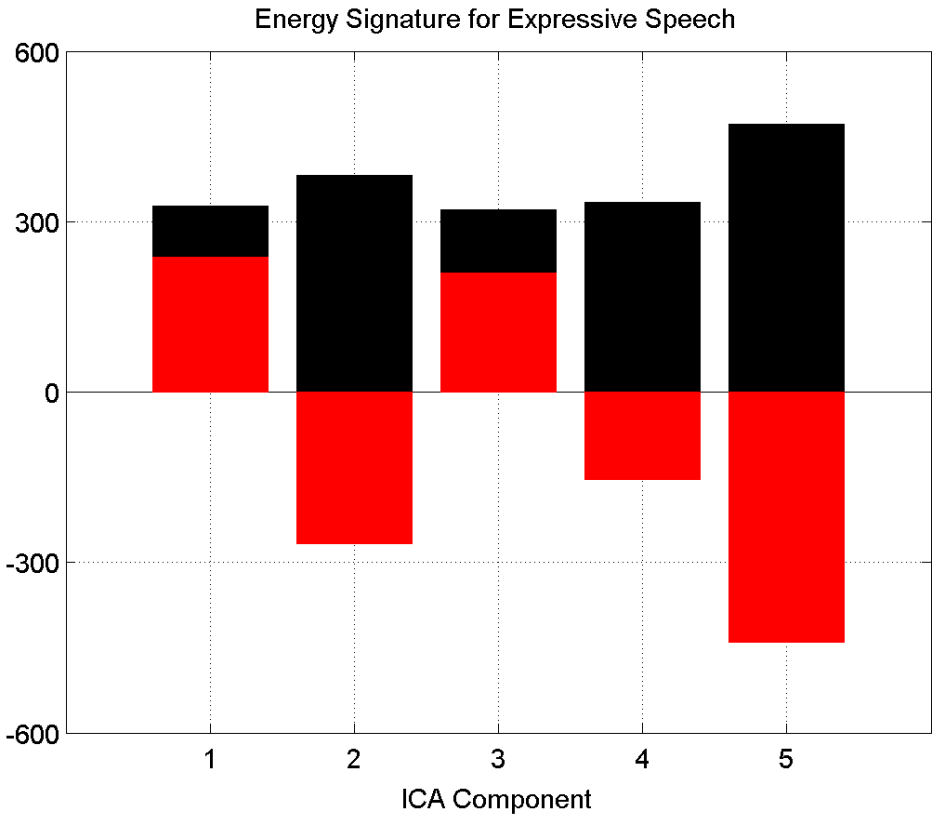
$$a^j = \sum_{t=1}^{k} u^j(t). \tag{7.8}$$

Although no component is fully responsible for expression, the distribution of the energy in the components is different for neutral speech and expressive speech. This is illustrated in Figure 7.9. Here we see how two independent components, one representing speech, the other expression might move through time. e_n^1 is the mean energy of the expressive component in a neutral sentence, e_e^1 is the mean energy of the expressive component in an expressive sentence. e_n^2 and e_e^2 then are the mean energies of the speech component for the neutral and expressive sentences respectively.

The main differences between the neutral and expressive energy signatures in this particular ICA model, is that one component for neutral speech tends to be positive, whereas it tends to be negative for happy speech (in this case component 1), and that another component tends to have more overall energy in happy speech



(A)



(B)

Figure 7.8 The energy (black) and amplitude (red) in the ICA components of (A) four neutral speech sequences and (B) four expressive speech sequences.

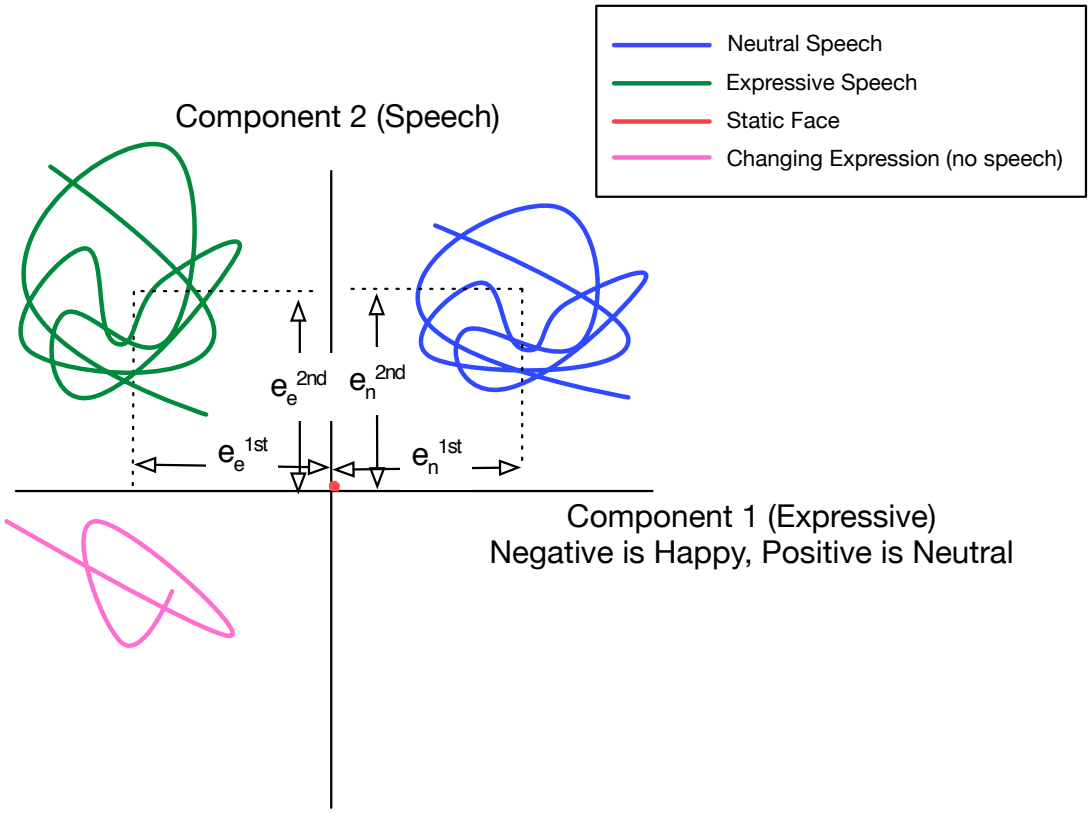


Figure 7.9 Illustration of independent component time series for neutral speech, expressive speech, changing facial expression and a static face.

(in this case component 4). The task then is to redistribute the energy in the ICA components computed from novel neutral speech so that they better match those observed in expressive speech. This involves rescaling the values of the components and (possibly) changing the sign. The weights used to transform the neutral speech components are computed using:

$$w^j = \frac{e_e^j}{e_n^j} \tag{7.9}$$

where w^j is the scaling for the j^{th} component, e_e^j is the energy in the j^{th} component of expressive speech and e_n^j the energy in the j^{th} component of neutral speech. Referring back to Figure 7.9 as an example, it can be seen how the scaling factor would be the ratio of $\frac{e_e^1}{e_n^1}$, giving a large scaling factor, and $\frac{e_e^2}{e_n^2}$ giving a scaling factor close to one.

Thus, given a sequence of novel neutral speech projected into ICA space, the parameter values are adjusted according to:

$$s^j(t) = \begin{cases} w^j u^j(t) & \text{sgn}(a_e^j) = \text{sgn}(a_n^j) \\ (-w^j(u^j(t) - \mu^j)) + \mu^j & \text{otherwise} \end{cases} \tag{7.10}$$

where μ^j is the mean value of the j^{th} component over the novel utterance, and sgn is $+1$ if the amplitude is positive and -1 if the amplitude is negative. The value $s^j(t)$ represents the new value of the j^{th} independent component at time t. Figure 7.10 illustrates this transform, where the blue line represents a time series on a speech mode and an expressive mode for input neutral speech, the green line represents the neutral speech after scaling, and the red dashed line indicates the time series on the same two modes of a ground truth equivalent expressive sentence.

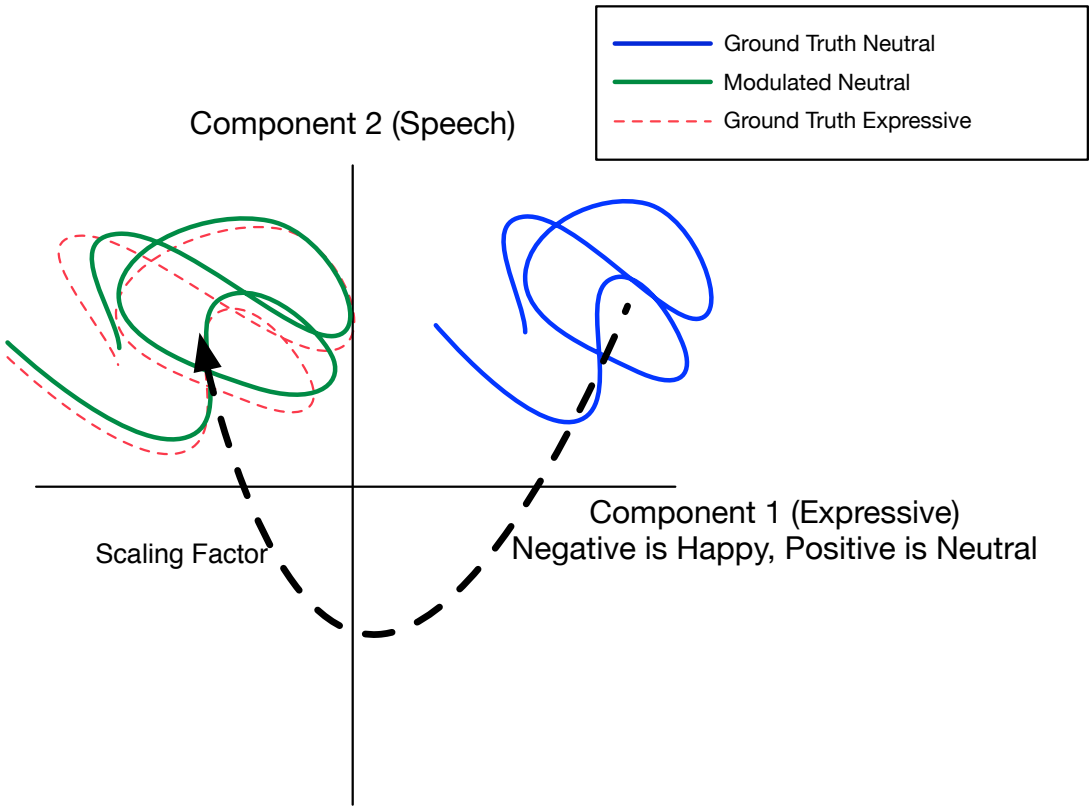


Figure 7.10 An illustration of the scaling process showing time series plotted on a speech and an expressive independent component. Blue is the input neutral, green is the scaled neutral, and red is the ground truth expressive equivalent.

7.5 Results

Five sequences from the B3D(AC)^2 corpus were chosen for each expressive style and paired with their equivalent neutral speech. To maximize the limited data available cross-validation was used, where five ICA models for an expressive style were trained using four of the sequence pairs, with a fifth sequence held-out for testing. Figures 7.11 and 7.12 show example time-varying trajectories in two of the five independent components for ground truth expressive, ground-truth neutral and the corresponding transformed neutral visual speech. Note that the transform is not attempting to recreate the expressive sequence exactly, rather the *style* of the expressive speech is being imposed onto the *content* of the neutral speech.

Sequences transformed from neutral to expressive styles using the process described in Section 7.4, not only show the correct change in facial expression, but also display the dynamics which are seen in the training set because real ICA data is being scaled rather than a style being statically imposed. Sample frames from video sequences containing real neutral speech, the same speech after transforming to an expressive style, and the corresponding real expressive sequences time-aligned to the neutral sequence are shown in Figure 7.13. Further examples of output created using this technique can be seen in Appendix C.

7.6 Evaluation

To assess the feasibility of this approach as a method of synthesising expression, small subjective evaluation involved a forced choice Turing test, where 14 viewers were each shown 8 sequence pairs (n=112 samples). The viewers were a mixture

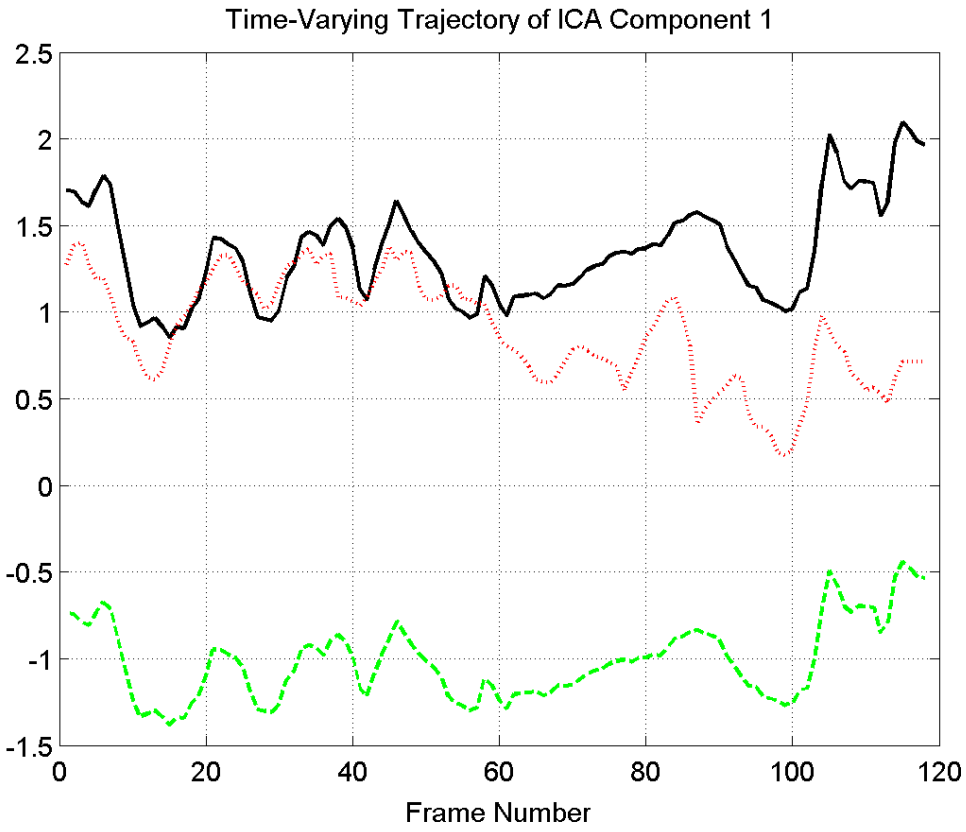


Figure 7.11 Time-varying independent components from an expressive like mode for a ground-truth neutral sequence (green dashed curve), the time aligned expressive equivalent sequence spoken in a happy style (red dotted curve), and the neutral sequence transformed into a happy style (black solid curve).

Figure 7.12 Time-varying independent components from a speech like mode for a ground-truth neutral sequence (green dashed curve), the time aligned expressive equivalent sequence spoken in a happy style (red dotted curve), and the neutral sequence transformed into a happy style (black solid curve).

Figure 7.13 Each row corresponds to an equivalent video frame for (left) real expressive speech time-aligned to (right) real neutral speech. The neutral versions transformed to expressive (center) display the *style* of the expressive sequences, but with intact visual speech gestures from the neutral sequences.

of graduate Computer Science students and faculty staff, some with experience of animation and visual synthesis techniques and some with little previous exposure. Sequences of time aligned real and transformed expressive speech were shown as visual only to ensure that acoustic artefacts due to time aligning the sequences had no influence on the results. The left-right ordering of the pair was randomised and viewers were asked to identify the real sequence in the pair. Of the 112 samples, 43 of the responses were correct. Using a binomial significance test we find that viewers cannot reliably identify the real sequences from the transformed sequences ($p>0.3$). In several cases, viewers stated that they found it difficult to choose between sequences in terms of realism, and so therefore chose their favourite. Responses tended to be biased in favour of transformed sequences being identified as real so we observe more false positives than false negatives. This is perhaps explained by the fact that transformed sequences tend to be slightly attenuated and thus smoother than the corresponding ground truth data.

7.7 Discussion

We have described a method for transforming sequences of neutral visual speech into expressive visual speech. Independent component analysis is used to decompose time aligned neutral and expressive visual speech, and weights are learned to distribute the energy in the independent components of (novel) neutral speech to better match the energy observed in expressive visual speech. This transformation results in expressive utterances that appear to display the same kinds of expression as seen in the expressive training set, and importantly the integrity of mouth shapes remains intact. Our approach uses ICA to separate neutral from expressive speech, unlike previous attempts which are trained to separate expressive data of different types (e.g. happy from sad). The advantage over the original technique

described in Section 7.3 of this is that the number of models grows linearly as we train for new expressions, whereas separating different expression types requires a model for each pair of expressions. This technique is flexible in that it allows any arbitrary neutral visual speech to be transformed into an expressive style using only a small training set of expressive and neutral speech.

The focus of the next chapter is to show that the technique described in this chapter can be further generalised. We show that a single model can be used to represent multiple expressions, therefore reducing the original data complexity of $\mathcal{O}(n^2)$ to $\mathcal{O}(1)$. We then demonstrate how the generalised technique is versatile enough to work on a different representation of facial data (a rigged animation model with movements defined in terms of the rig's controllers), and finally show how it is possible to represent expressive data as mixes of the ICA training data in order to produce subtle expressions which were not in the original training set.

Chapter 8

Mixed Model Modulation

8.1 Motivation

The work in this Chapter is an attempt to further generalise and utilise the ideas presented in Chapter 7. As was mentioned in Section 7.3, Cao et al. [2003] proposed that it is possible to project two time aligned sentences of the same speech content but different expressive styles onto an Independent Component Analysis (ICA) model which has been trained on the same two sentences. It is then possible to manipulate the expressive style of one of the sentences by identifying the independent component responsible for expression and transferring the expressive values from the same component in the other therefore allowing expressive style to be changed independently of the speech data. While interesting, Cao et al. [2003] does little to exploit this useful characteristic of ICA.

The work in Chapter 7 demonstrated generalisations to the original technique which not only make it of potential usefulness in the animation industry, but also offer significant improvements in terms of the technique's complexity. Whilst the approach in Cao et al. [2003] requires the same sentence in multiple expressive

styles and an ICA model was trained for each pair of expressions, the generalised approach only requires a neutral and an expressive example of the speech. Therefore for n styles, our approach requires n trained ICA models. Its real utility lies in the fact that the relationship between the speech and the expression is modelled. Using Independent Component Analysis, a projection space is learned in which it is possible to manipulate speech and expression independently of one another. We introduced a method where once an ICA model was trained using a very limited amount of training data (less than 100 frames), an arbitrary amount of expressively neutral test data could be projected onto the ICA model, and the independent components of the neutral test data could be manipulated in isolation. By redistributing the energy in the modes to match the energy distribution of expressive training data, it is possible to *remix* or modulate the neutral test data with the expression contained in the ICA model's training data. An evaluation showed favourable subjective viewer preferences.

Human expression and interpretation of emotion is highly complex and rarely contains "categorical" expressions (e.g. happiness, anger). Instead, our outward displays of emotion are the combined mix of many complicated internal factors. For any expressive modulation to be considered plausible, an attempt to model this complex mix of expression must be made. A limitation of the method described previously is that it still requires a different model to be trained for each expressive style and that neutral data can only be modulated with the categorical and discrete expressive modalities contained within the training set. Attempting to combine the outputs of multiple ICA models trained on different expressive styles is not guaranteed to produce plausible output.

This chapter therefore demonstrates how a combined ICA model trained on multiple expressive styles can be produced. We show its use in modulating neutral test data into the various styles on which the mixed ICA model was trained. A non-

trivial data transform is presented to show how the technique works on different representations of visual speech data. Different visual outputs are shown along with their subjective evaluations. Finally we show how the technique can be used to reasonably approximate the complex mixes of expressions which we as humans use to communicate not only our reasoning but also our emotions.

8.2 Modulating Neutral Speech in many styles using ICA

To further generalize the approaches in Chapter 7 we construct a single ICA model from several different expressive styles such that:

$$\mathbf{s}_{Mixed} = \mathbf{W}\mathbf{b}_{Mixed} \tag{8.1}$$

where $\mathbf{s}_{Mixed}$ are the independent components from a set of expressive sequences of various styles (happy, angry, sad and surprised) together with their time aligned neutral equivalents, and $\mathbf{W}$ is the estimated un-mixing matrix. Only the mixing and un-mixing matrices $\mathbf{W}$ and $\mathbf{W}^{-1} \Leftrightarrow \mathbf{Q}$ (returned in the FastICA implementation) are of importance, and therefore $\mathbf{s}_{Mixed}$, the calculated independent components are discarded. $\mathbf{b}_{Mixed}$ is a set of training sequences in AAM space representing all the training data in various styles and their time aligned neutral equivalents (except the neutral sequence to be modulated and its expressive equivalents) such that.

$$\mathbf{b}_{Mixed} = \begin{vmatrix} b_{angry} & b_{angry/neutral} & b_{happy} & b_{happy/neutral} & \cdots \\ b_{sad} & b_{sad/neutral} & b_{surprised} & b_{surprised/neutral} & \end{vmatrix} \tag{8.2}$$

The neutral equivalents for each expressive style are required in the training set. Including only the neutral corresponding to one of the expressive styles leads to FastICA converging to a non-optimal separation of speech and expression. It is unclear if this is because the exact timing between expressive and neutral versions of a sentence is required in training or whether there simply isn't enough training data without including the larger amount of neutral training data. Experimenting with the amount of data in the training set shows that the convergence does indeed improve as the amount of training data is increased. Training on between 10 to 14 sentences of four expressive styles and neutral (around 10,000 samples) seems to give the best convergence. Therefore it is likely to be a question of providing the ICA training process with enough training data. Since the test expressive and neutral sentence were never included in the training set, it is unlikely to be a cause of training bias. This linear projection is able to capture the range of expressive styles in different positive/negative regions of the independent components (modes) of the model. This is shown in Figures 8.1, 8.2, 8.3, 8.4 and 8.5 where the energy in each of the independent components is computed as the summed amplitude over time thus:

$$e_j = \sum_{t=1}^{k} s_{jt}, \tag{8.3}$$

where e_j represents the energy in the j^{th} component and s_{jt} represents the value of the j^{th} component at time t.

Note how mode three is very positive for angry, very negative for happy and has relatively little energy for neutral speech. This is the mode which the ICA model uses to discriminate between these two expressive styles. Mode five also shows

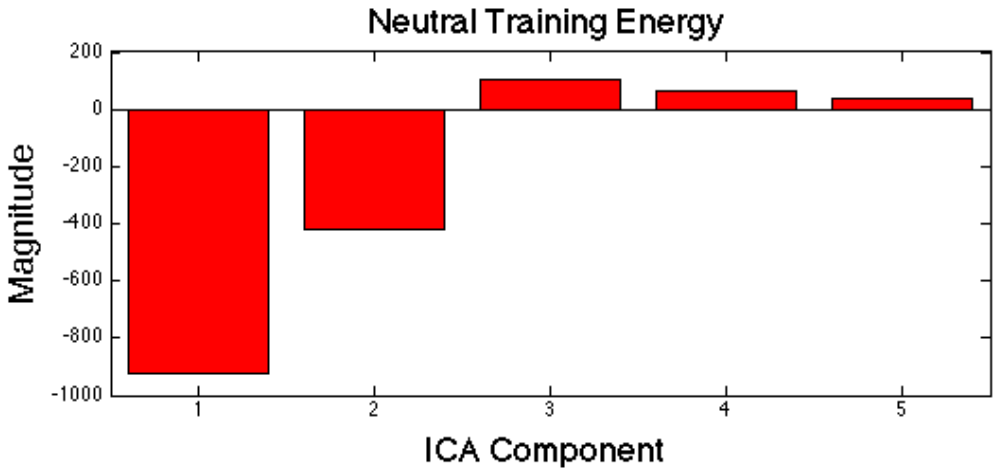


Figure 8.1 The distribution of the energy across the independent components for visual speech spoken in a neutral style.

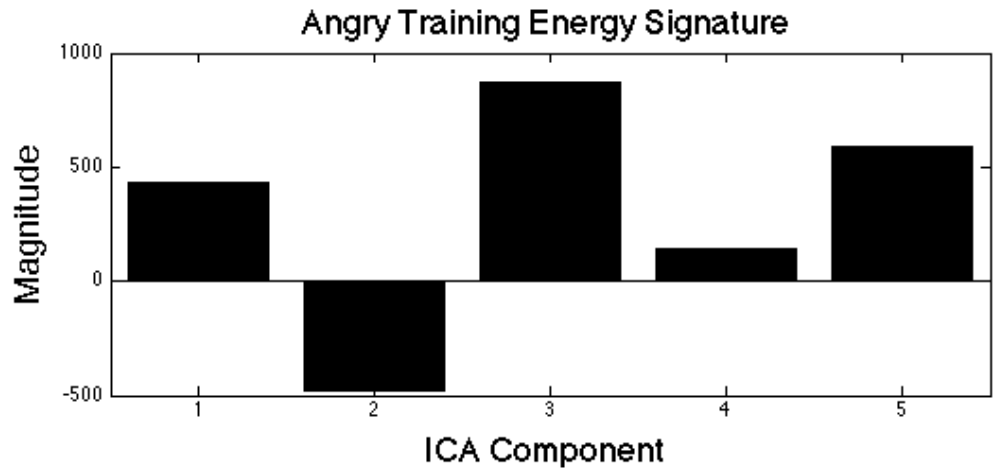


Figure 8.2 The distribution of the energy across the independent components for visual speech spoken in an angry style.

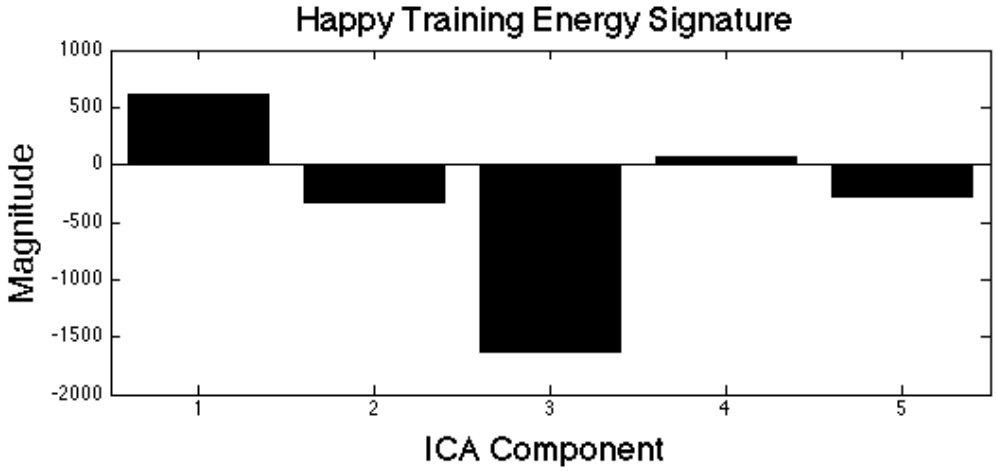


Figure 8.3 The distribution of the energy across the independent components for visual speech spoken in a happy style.

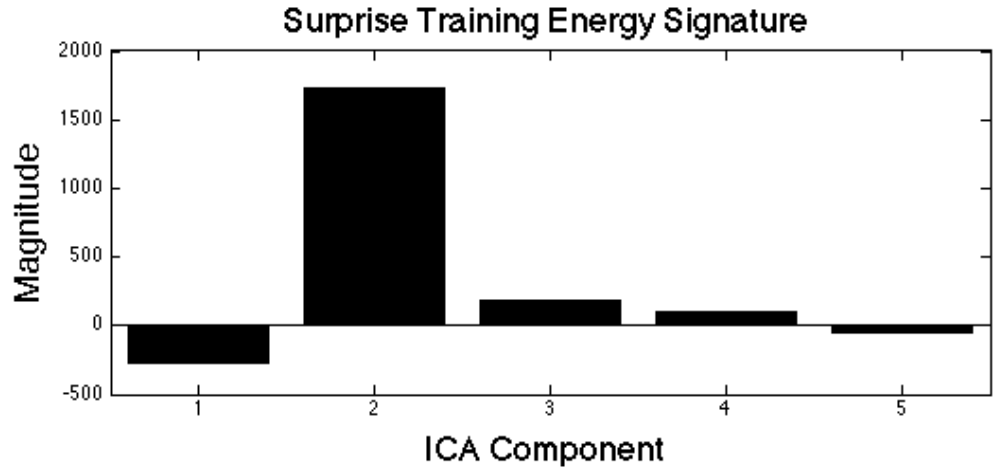


Figure 8.4 The distribution of the energy across the independent components for visual speech spoken in a surprised style.

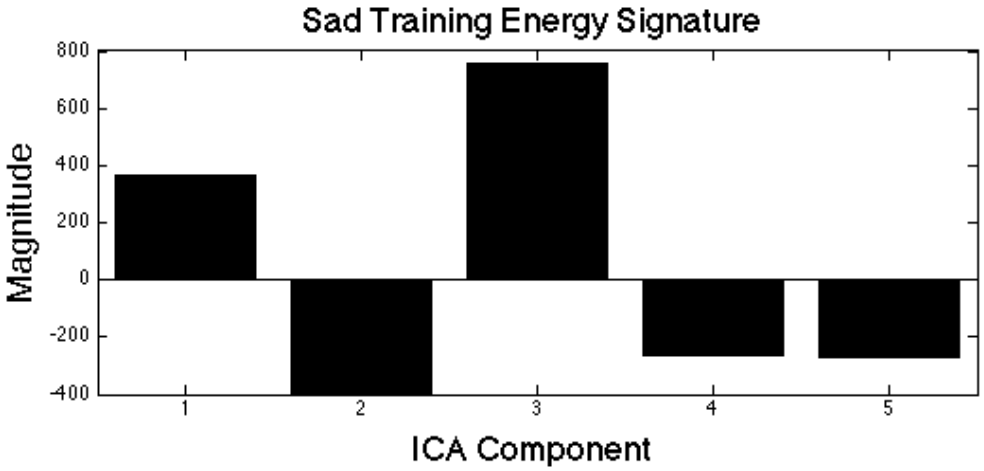


Figure 8.5 The distribution of the energy across the independent components for visual speech spoken in a sad style.

a similar pattern but to a lesser extent indicating that mode five contains more speech-related information than mode three. On the other hand, mode four stays much the same for all examples, indicating that it contains information related to speech. Mode two is highly positive for surprise and negative elsewhere, indicating that the ICA model uses this mode to discriminate surprise. Sad and angry show a similar energy distribution (the actor in our dataset does indeed look similar portraying these styles), except that mode five is positive for anger and negative for sad, indicating the ICA model discriminates between these subtly different expressions on this mode.

To transform synthesized neutral visual speech into an expressive style, the neutral speech can be projected on to the ICA model, and then the appropriate expressive modes manipulated to adjust the expression. When transforming parameters that encode neutral visual speech into those which encode expressive visual speech, the dynamics of the expression must appear natural and the mouth movements corresponding to speech must remain valid.

Figures 8.1, 8.2, 8.3, 8.4 and 8.5 show that no component is fully responsible for expression, but the distribution of the energy in the components is different for neutral speech and various expressive styles. So, the problem of manipulating the expressive style of speech involves redistributing the energy across the respective independent components for the unseen speech. To do this we first compute the ICA components for a novel neutral speech utterance then offset each ICA mode into the respective range, and then scale the energy in the speech-like components to account for the higher velocity movements observed in expressive visual speech.

To compute the offset we first calculate the mean values for the components using:

$$\mu_{jx} = \frac{1}{n}\sum_{i=1}^{n} s_{x(ji)} \tag{8.4}$$

$$\mu_{jn} = \frac{1}{n}\sum_{i=1}^{n} s_{n(ji)} \tag{8.5}$$

where μ_{jx} is the mean value for the j_{th} expressive component and μ_{jn} is the mean value for the j_{th} neutral component, and finally:

$$o_j = \mu_{jx} - \mu_{jn}. \tag{8.6}$$

This can then be applied to the independent components for a new sequence using:

$$\mathbf{s}^{jn}_{Modulated} = s_{jn} + o_j, \tag{8.7}$$

where $\mathbf{s}^{jn}_{Modulated}$ is the resulting expressive speech, composed of neutral speech modulated with expression.

The original variations of the neutral ICA time series are maintained by virtue of this constant offset which ensures that there is some variation in the modulated expression and that it is not simply a fixed grin or frown etc. However, since

neutral speech is naturally less animated than expressive speech, the output of the modulated neutral speech appears attenuated when compared with the original expressive ground truth, particularly with regard to mouth articulation. To address this problem, the speech-related components are scaled by computing the power in each component using 8.8 and 8.9:

$$p_{jx} = \frac{1}{n}\sum_{i=1}^{n} {s_{x(ji)}}^2 \tag{8.8}$$

$$p_{jn} = \frac{1}{n}\sum_{i=1}^{n} {s_{n(ji)}}^2, \tag{8.9}$$

where p_{jx} is the power in the j^{th} mode of the expressive training sequences and p_{jn} is the power in the j^{th} mode of the neutral test sequence. The ratio between the two powers is computed using:

$$r_j = \frac{p_{jx}}{p_{jn}}. \tag{8.10}$$

This ratio tends to be very large when the corresponding mode is an expressive mode, and much smaller when the corresponding mode is a speech mode. It makes intuitive sense that a mode encoding expression would have high energy in an expressive sequence and low energy in a neutral sequence, therefore the ratio between the two should be large. Conversely a speech mode should have high energy in both a neutral and an expressive sequence (although slightly higher in the expressive sequence due to the more animated nature of expressive speech), therefore the ratio is small. It can be seen that this is the case by studying modes three and four in Figures 8.1, 8.2, 8.3, 8.4 and 8.5, for examples of the energy in a speech mode and an expressive mode respectively. A scaling vector was calculated from this ratio, where if the ratio exceeded some threshold, we assume it corresponds to an expressive mode, so set it to 1 so it will have no effect on the scaling. The

scaling is calculated using:

$$scale_j = \begin{cases} 1 & \text{if } r_j >= threshold \\ r_j & \text{if } r_j < threshold, \end{cases} \tag{8.11}$$

where $threshold = 2$, because it was observed that, for our data, the ratio between two speech modes never exceeded this value. Since this is an empirical choice it would probably not generalise to all datasets. We then scaled the transposed neutral sequence as shown in Equation 8.12

$$\mathbf{s}^{jn}_{Modulated} = \mathbf{s}^{jn}_{Modulated} \cdot scale_j, \tag{8.12}$$

which has the effect of amplifying the speech like articulatory movements whilst leaving the expression like movements unchanged. The resulting ICA modes are then inverted to combined shape and appearance parameters thus:

$$\mathbf{b}_{Modulated} = \mathbf{W}^{-1}\mathbf{s}_{Modulated}, \tag{8.13}$$

where $\mathbf{W}^{-1}$ is the pseudo-inverse of the un-mixing matrix calculated in equation 8.1. The AAM parameters can then be applied to the respective components of the model and the video frames rendered by warping the resulting appearance image from the mean shape to the generated shape. Finally, the video frames are compiled into a movie file at 25fps, and audio is added.

8.3 Results for Mixed Model with AAM output

To test this new approach 15-fold cross validation from 14 sentences of all four expressive types and their time aligned neutral equivalents was conducted by holding

out all the data for the test sentence (i.e. the training set consisted of everything except the test sentence in all its forms of neutral, happy, angry, sad and surprised). Therefore each cross validation training set contained 112 sequences with an average length of around 60 frames, giving an total of around 6720 frames of training data. The mixing and unmixing matrices, $\mathbf{Q}$ and $\mathbf{W}$, were re-calculated for each training fold, and the time aligned neutral and expressive sequences were projected onto the independent components. The neutral test sentence was modulated with each of the four expressive styles. The output was rendered, along with the original ground truth neutral and expressive sequences.

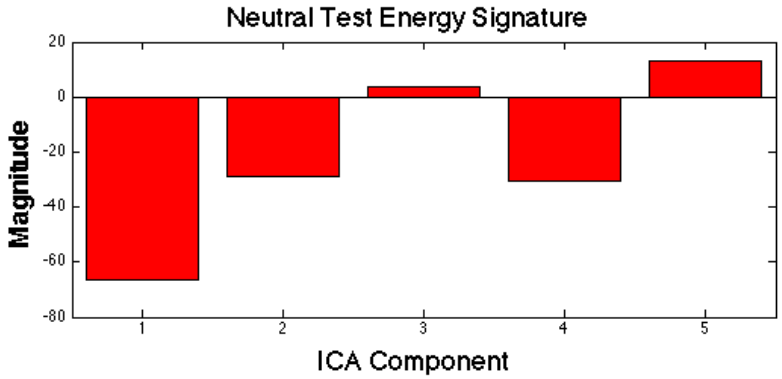


Figure 8.6 Energy distribution of a neutral test sequence before modulation.

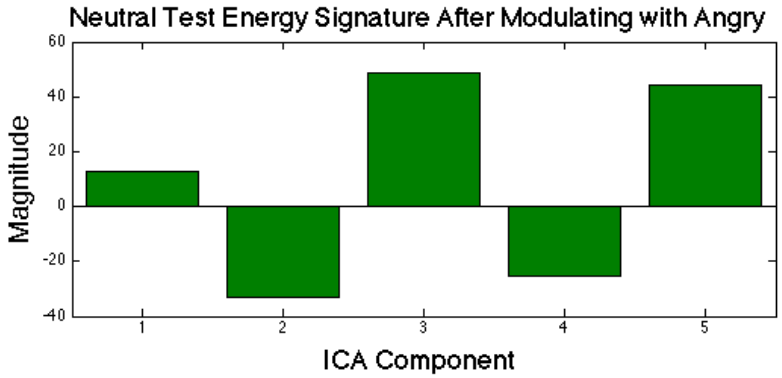


Figure 8.7 Energy distribution the neutral test sequence after modulation with anger.

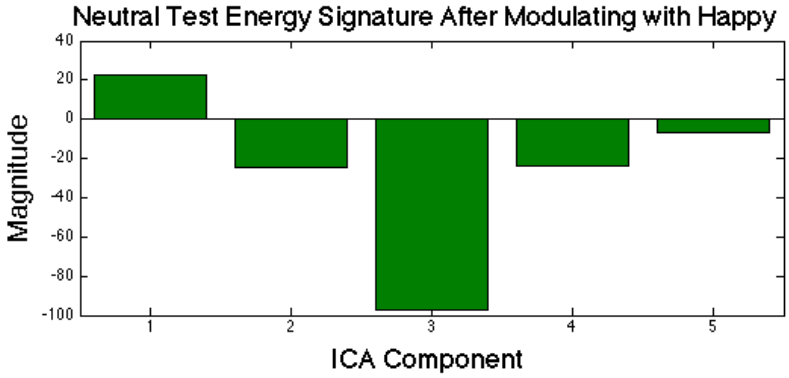


Figure 8.8 Energy distribution the neutral test sequence after modulation with happiness.

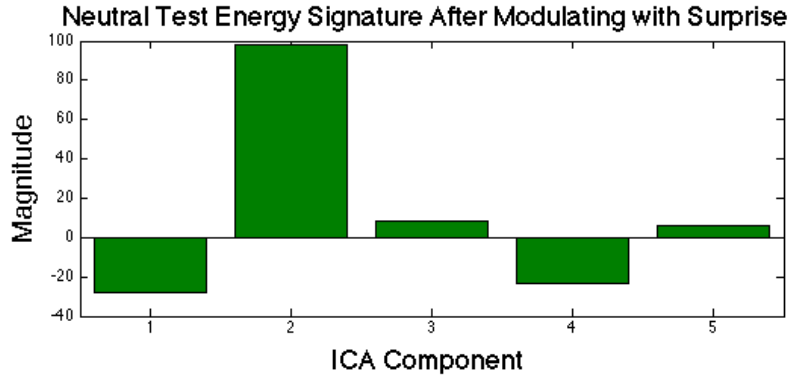


Figure 8.9 Energy distribution the neutral test sequence after modulation with surprise.

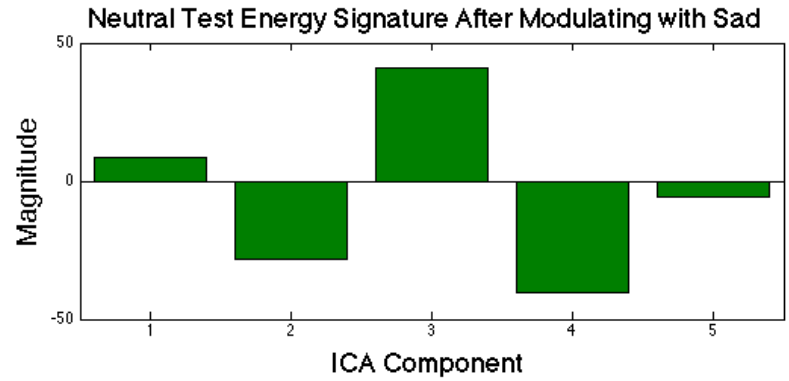


Figure 8.10 Energy distribution the neutral test sequence after modulation with sadness.

This shows the correlation learned by the ICA decomposition between the returned modes and the different expressive training types. By learning a single ICA model across a range of expressions, neutral speech can be transformed into different styles by varying the style from which we calculate the offset. A single ICA model is capable of separating independent components of speech and expression for multiple expressive styles. Figures 8.1, 8.2, 8.3, 8.4 and 8.5 show how an ICA model trained on all four expressive styles and neutral discriminates between happy and angry movements on mode three, between neutral and surprised movements on mode two and between neutral and sad movements on mode one.

The expressive and neutral video had been time aligned to another set of held out expressive sequences. Therefore both expressive and neutral video had undergone some time alignment and were aligned to the left out expressive audio. This left out expressive audio was then added into the output movies. An example of neutral and expressive ICA modes before and after modulation are shown in Figure 8.15, where the top plot shows how component three (an expressive mode) has been shifted into the correct range for angry speech, but still displays a similar shape to the ground truth neutral speech. The second plot shows the same data but transposed to the correct range for happy speech. The third plot shows the model distributing surprise onto mode two. The forth plot shows shows sad speech distributed onto mode one, and the last plot shows how component four (a speech mode) maintains its original shape, but has higher amplitude after modulating (with angry) and scaling. It can be seen that all these shifts correspond with the energy distributions in Figures 8.1, 8.2, 8.3, 8.4, 8.5 and 8.6, 8.7, 8.8, 8.9 and 8.10 . These shifts were all made with the same mixed emotion ICA model.

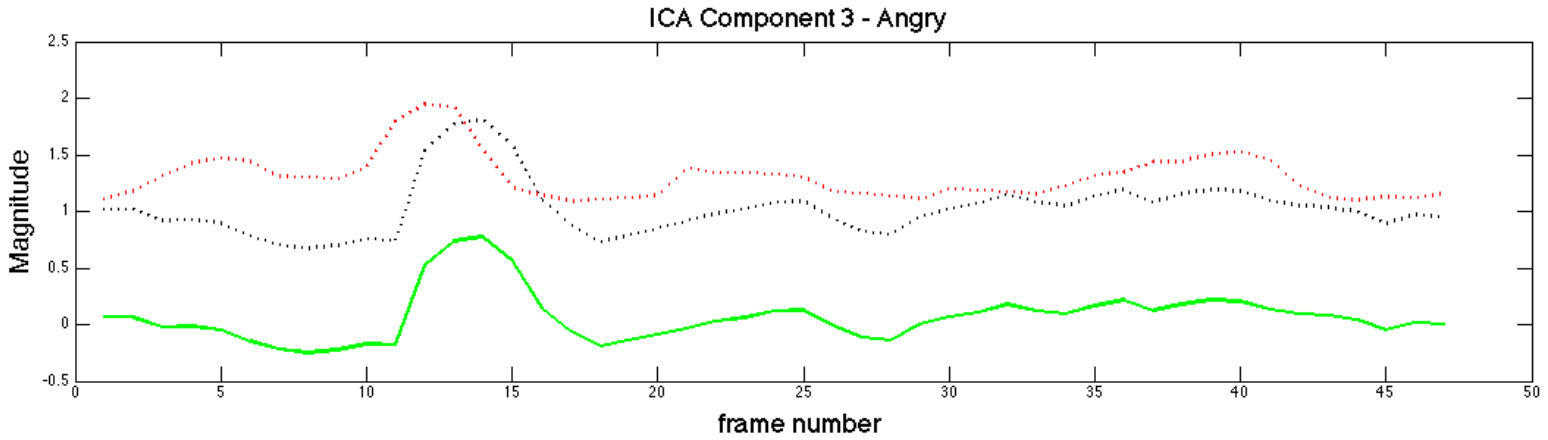


Figure 8.11 Time-varying independent components for a ground-truth neutral sequence (green solid curves), expressive equivalent sequences (red dashed curves), and neutral after expressive modulation (black dashed curves).

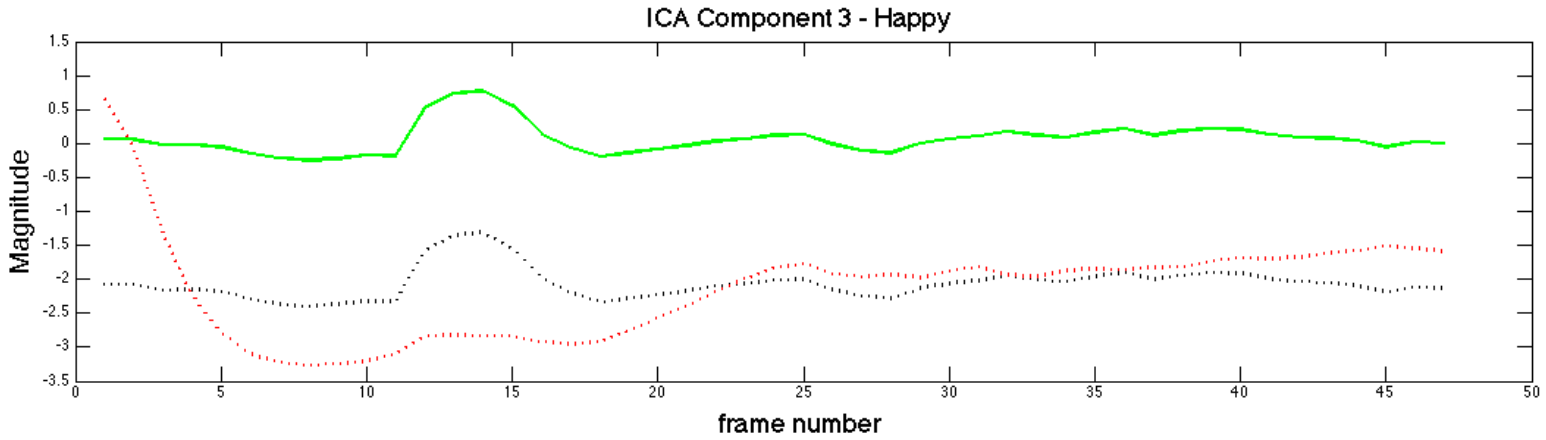


Figure 8.12 Time-varying independent components for a ground-truth neutral sequence (green solid curves), expressive equivalent sequences (red dashed curves), and neutral after expressive modulation (black dashed curves).

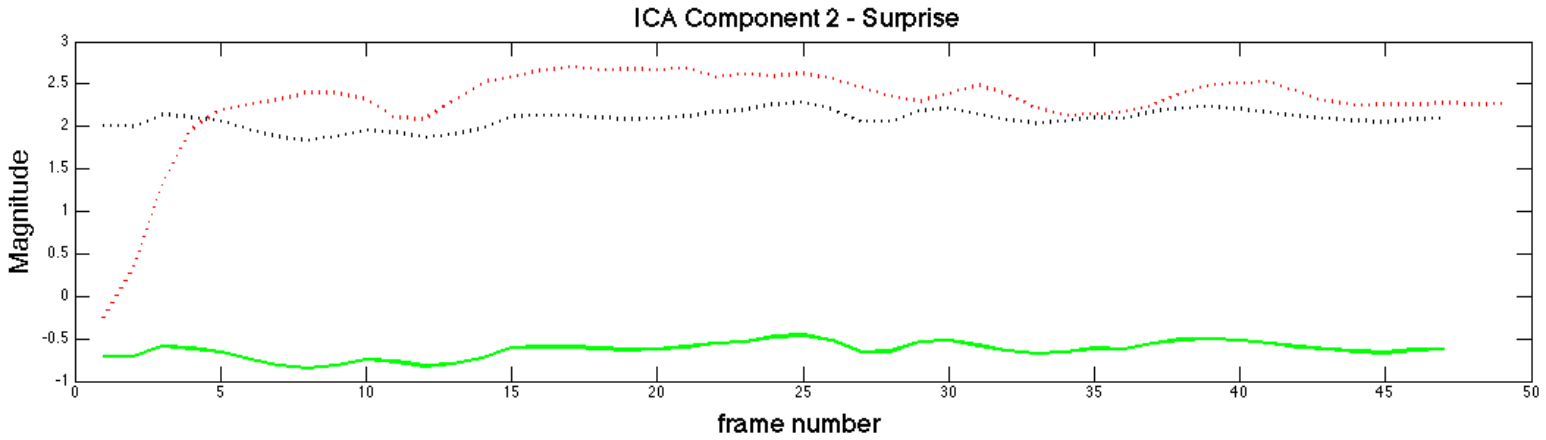


Figure 8.13 Time-varying independent components for a ground-truth neutral sequence (green solid curves), expressive equivalent sequences (red dashed curves), and neutral after expressive modulation (black dashed curves).

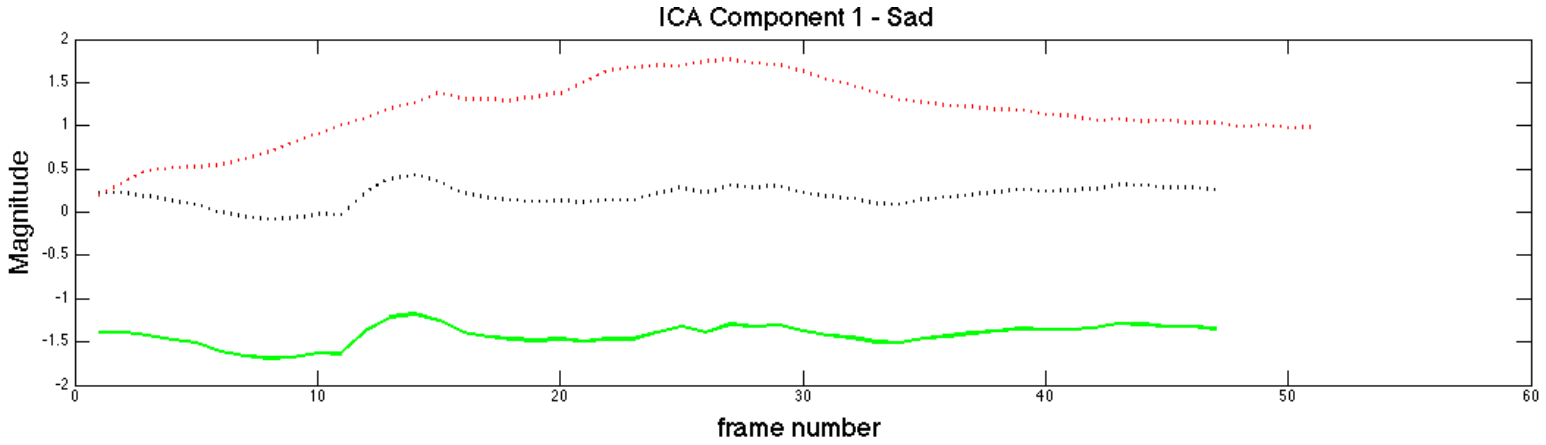


Figure 8.14 Time-varying independent components for a ground-truth neutral sequence (green solid curves), expressive equivalent sequences (red dashed curves), and neutral after expressive modulation (black dashed curves).

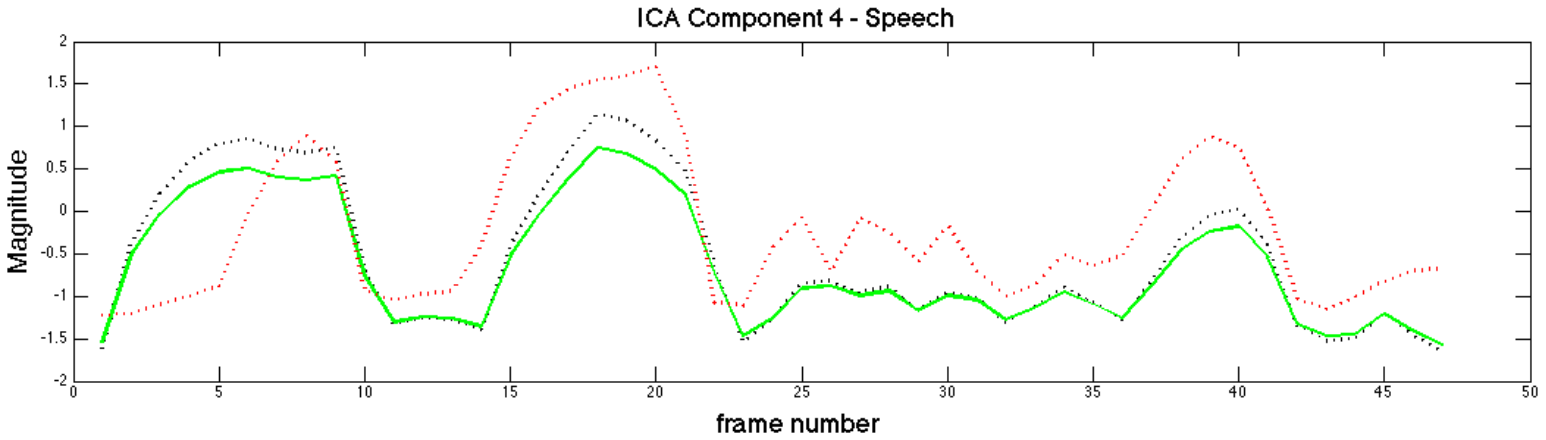


Figure 8.15 Time-varying independent components for a ground-truth neutral sequence (green solid curves), expressive equivalent sequences (red dashed curves), and neutral after expressive modulation (black dashed curves).

Figure 8.16 shows some example frames from rendered output, with expressive ground truth and neutral ground truth on the left and right respectively, with modulated neutral in the middle. Further examples of output created using this technique can be seen in Appendix D

8.4 Evaluation of AAM Output

Three subjective experiments were used to evaluate the approach for transforming neutral to expressive speech. A forced choice test was used to evaluate recognition of expressions, a mean opinion score (MOS) test was used to confirm whether speech articulation was still acceptable after modulation, and a turing test was used to see if participants were able to tell the difference between ground truth and synthesized sequences.

8.4.1 Forced Choice Test

Following the procedure described in Beskow and Nordenberg [2005], a perceptual test was conducted using 7 participants. Participants were each shown 40 synthesized expressive sequences in a random order and asked to classify each as either happy, angry, sad or surprised. The sequences were presented without audio to remove any audible cues that participants might use to identify the emotion. Participants were allowed to see sequences as many times as they needed to. Overall average recognition rate for all expressions was 88.2%. Average recognition for each expression was 100% for happy, 91.5% for surprise, 87.3% for angry and 74.3% for sad. All of these are well above chance (25%). The reason for the lower recognition for sad and angry is that the actor used for data capture looked similar displaying these two expressions. The main differences were in attack and volume of delivery,

Figure 8.16 Each row corresponds to an equivalent video frame for (left) real expressive and (right) real neutral speech. Row 1 shows anger, and row 2 shows surprise and row 3 shows happiness and row 4 shows sadness. The neutral versions transformed to expressive (center) display the *style* of the expressive sequences, but with intact visual speech gestures from the neutral sequences.

attributes which are lost without audio.

8.4.2 Mean Opinion Score Test

A second experiment was conducted to show that the modulation of expression onto neutral speech leaves the speech component intact. Eight participants were each shown 56 sequences in a random order, half of which were ground truth expressive, the other half of which were modulated neutral. Participants were instructed to pay particular attention to the mouth articulation and how well this matched the audio as well as how convincing they thought the facial expression displayed was. The audio tracks were from the left out expressive sequence (as described in Section 8.3), therefore none of the video was from the same sequence as the audio. This was done to remove possible bias which would occur if the ground truth expressive video was played synchronized with the corresponding audio. Participants were told to assign a score to each sequence from 1 to 5, 1 being very poorly synchronized with inappropriate mouth movements or implausible facial expressions, and 5 being perfectly synchronized with excellent facial expressions. Participants were only allowed to see each sequence once. The overall MOS for ground truth expressive was 3.94 and for modulated neutral was 3.45. A Wilcoxon signed-rank test Wilcoxon [1945] was used to calculate significance, and this shows that the difference in the MOS is statistically significant ($p < .05$). It is thought that this difference occurred because although both neutral input data and ground truth expressive data had been warped to the left out expressive audio, it is inevitable that the neutral will have been warped slightly more, since the already expressive ground truth would have needed less processing when warping to a similar soundtrack, whereas the neutral input had more dissimilar dynamics to the target soundtrack.

8.4.3 Turing Test

To evaluate the overall performance of this approach, a Turing Test experiment was conducted. 16 participants were each shown 120 sequences in a random order. Half were ground truth expressive and half were modulated neutral. The participants were told that half the sequences were real and half were synthesized. They were asked to decide for each sequence whether it was real or synthesized. Sequences were all shown alongside the held out expressive audio. Participants were allowed to see each sequence twice. Following the example of Geiger et al. [2003], we calculated the mean contingency table, as shown in Table 8.1. A Chi Square test

		Ground Truth	
		R	F
Response	R	33.6	22.4
	F	28	36.4

Table 8.1 Contingency table

with Yates correction and a Fisher's exact test shows that participants are not able to reliably identify the modulated sequences from the ground truth sequences ($p = 0.0983$).

8.5 Discussion

The expression recognition experiment showed that this approach is successful in modulating the expressions in the training set with neutral visual speech. However recognition scores were slightly lower for angry and sad. Indeed these two expressions were confused for one another on several occasions and account for nearly all the error in this result. Participants reported finding these two expressions difficult

to separate from one another. Inspection of the ground-truth recordings show that they appear very similar. Further testing with a broader range of expressions, and perhaps using different actors would help to demonstrate that this is an issue with the data rather than the approach.

It is less easy to say why the mean opinion scores of the synthesized sequences were lower than for ground-truth sequences. The mean score of 3.45 at least suggests that the speech articulation is left intact, but the fact that it is lower than the ground truth score means that it is not perfect. Although participants were instructed to judge the video sequences on the basis of mouth shape and lip sync, several particpants stated that they had rated movies lower for things like a lack of sharpness in the appearance, or graphical rendering artefacts. Other participants cited greater variance in the expression of some sequences.

It is likely that the output from the Active Appearance Model is at least in part to blame for the lower mean opinion scores and the significant difference observed in the Turing test for two reasons. Firstly, AAM output often suffers from artefacts which are a bi-product of poor fitting during the initial AAM video encoding. Unless lighting is perfect during filming, there is usually a very smooth gradient defining the inner lips. AAMs are poor at fitting to smooth gradients (performing better when there is a sharp definition), meaning that the inner lips are usually not well tracked. This often leads to interference between the shape and appearance portions of the model, resulting in blurring of the inner lips. It is possible that some people may be particularly sensitive to this blurring and therefore perceive the quality of the final output to be degraded. Secondly there is the phenomenon of the uncanny valley Mori et al. [2012]. First proposed in the 1970's, the uncanny valley is the idea that as representations of humans (such as robots or avatars) become closer to being indistinguishable from real humans, at some point there is a sudden drop in viewer's familiarity with the subject material, eliciting a feeling

of "eeriness" or even revulsion. Various explanations have been offered for the effect such as instinctive preference in terms of mate selection, mortality salience and pathogen avoidance. It is possible that the near photorealistic output of the AAM lies within this uncanny valley and therefore adversely effects results.

One obvious way of overcoming this limitation is to retarget the output animation to a graphics model. That is to take the movements from one type of animation model, and have them displayed on another. In this case, taking the animation parameters produced by the procedure described above, and retargeting them to a more pleasing type of graphical model may overcome the issue of the "uncanny valley". Various approaches have been used to achieve retargeting (Sumner and Popovic [2004] Choe et al. [2001a] Theobald et al. [2007a] Kholgade et al. [2011]). The algorithm used in this work is based on Pighin et al. [2006] and is described in Section 6.5.

8.6 Data Transformation

Next the dataset described in Section 4.2 was transformed. Since AAM features are based on Principal Component Analysis, each element of an AAM feature is a global in nature. That is, a change to an element of an AAM feature will have an effect on every part of the face model. Considering the earlier proposition that speech and expression are largely independent, a feature which cannot discriminate between expressive movement (such as furrowing of the brow) and articulatory movement is undesireable. Rather it would be better to have a feature allowing geo-spatially discrete parts of the face to be manipulated independently. For example, consider the expressing of surprise, the key characteristic of which is the raising of the eyebrows. Ideally an animator requires independent control over the eyebrows and the mouth. In an Active Appearance Model, the first orthogonal

mode accounts for the largest variation from the mean. For expressive speech, this is usually some combination of the mouth opening and the eyebrows raising and is therefore not an appropriate method of control to animate speech movements and surprised movements simultaneously. Although Independent Component Analysis does a reasonable job of separating out these different movements (those relating to speech and those relating to expression), it can be seen from Figures 8.1, 8.2, 8.3, 8.4 and 8.5, that no ICA mode is entirely responsible for expression (or speech). However, if the data were to be transformed into a geo-spatially discrete type (where each element of a feature is responsible only for a spatially discrete area of facial control), then it is likely that ICA will perform better in its task of separating speech from expression.

The following section details work to implement such a data transform and to address the problems with using AAMs described in the previous section (artefacts such as blurring or mesh tearing and the uncanny valley). An animation rig was acquired, onto whose controls the movements originally captured by the AAM could be transferred. For this work we chose the Morpheus rig Burton [2010], which is an open source freely available blend shape based rig created in the Autodesk Maya software package. It consists of facial geometry and controllers for deforming the geometry. See Figure 8.17 for some examples of how Morpheus can be manipulated. In overview, the Morpheus geometry was warped to each AAM landmark configuration from the training set. Then the Morpheus rig controls were optimised to best match the rig to the warped geometry therefore transferring the AAM captured video into rig control space, where each frame of animation was simply represented by a vector of activations for the rig controllers and was consequently no longer a global representation.

First the geometry and triangulation data for Morpheus were exported from

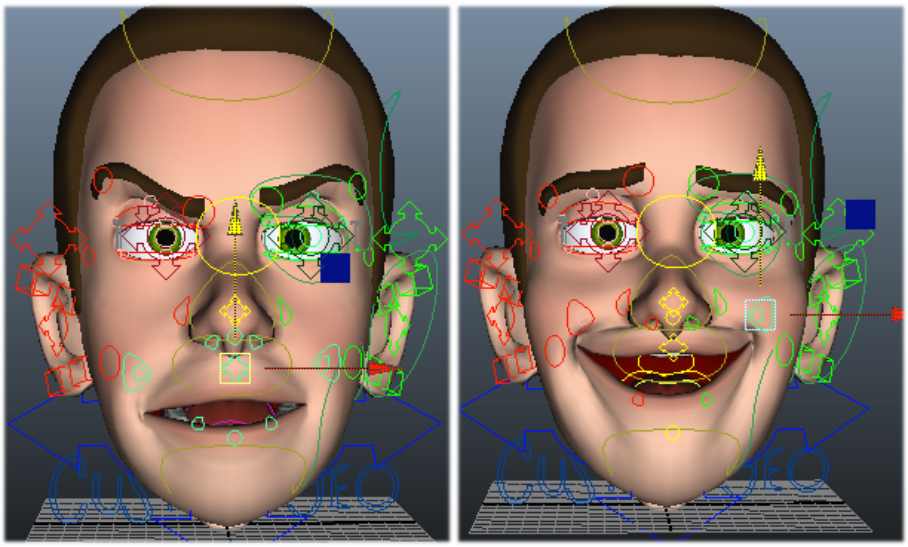


Figure 8.17 Examples of posing the Morpheus facial rig into different poses. The controllers are visible.

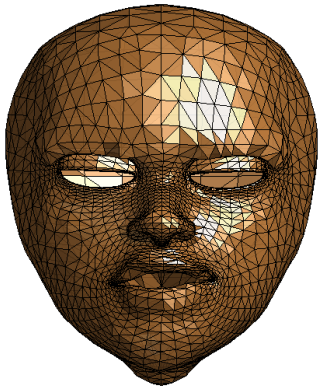


Figure 8.18 A visualisation of the Morpheus geometry having been imported into Matlab, plotted with the "patch" function and illuminated appropriately. Note that the geometry for the ears, hair and eyebrows has not been exported from Maya, leading to a slightly different appearance. This was done to reduce complexity.

Maya into Matlab where they could be manipulated more easily. See Figure 8.18 for an example of the geometry visualised using Matlab's "patch" function. Next,

an appropriate affine transformation was calculated to align the mean landmarks from the AAM model to the Morpheus geometry in the form:

$$\mathbf{p}'_k = \mathbf{p}_k \cdot \mathbf{T} \cdot \mathbf{R} \cdot \mathbf{S} \tag{8.14}$$

where $\mathbf{p}_k$ is the k_{th} AAM landmark, $\mathbf{T}$ is the translation matrix, $\mathbf{R}$ is the rotation matrix, $\mathbf{S}$ is the scaling matrix and $\mathbf{p}'_k$ is the aligned position of point $\mathbf{p}_k$. Since the AAM features originally returned by the tracking algorithm were already normalised for scale, rotation and translation, the affine transformation needed to be calculated once only and could be used to align every frame in the training set. Figure 8.19 shows the mean AAM landmarks superimposed on top of the Morpheus geometry after alignment. At this stage the alignment did not have to be perfect and not all points were used as warping correspondences, therefore the fact that the AAM points marking the eyes were not aligned over the eyes of the Morpheus geometry was of no consequence. The Morpheus geometry was deformed to match the landmark configuration of every frame of AMM captured video data. Corresponding pairs of points on the AAM landmarks and the Morpheus geometry were established. All AAM landmarks for the mouth and eyebrows were paired with points in the Morpheus geometry as well as points at the top and bottom of the head. For each AAM tracked frame, the landmarks were aligned with the Morpheus geometry using the pre-calculated affine transform. Then the positions of the paired points in the AAM landmarks were copied to the corresponding points in the Morpheus geometry. These Morpheus vertices were therefore constrained and contain the *essential characteristics* of the pose since they described the positions of the facial features in AAM space. Scattered Data Interpolation was then applied to calculate new positions for the unconstrained vertices in the Morpheus

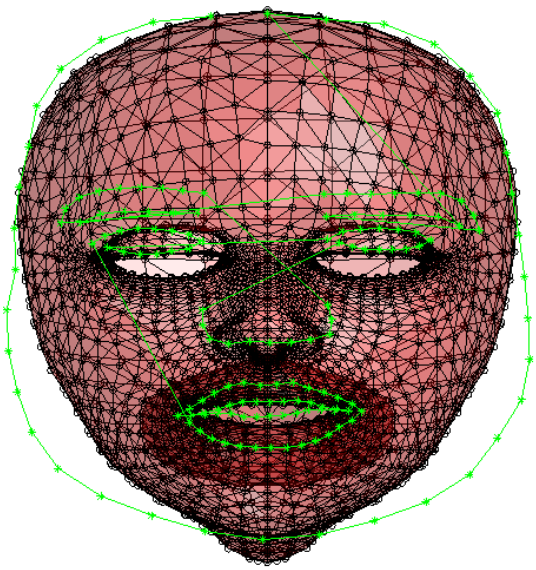


Figure 8.19 The mean AAM landmarks superimposed over the Morpheus geometry after alignment.

geometry thus:

$$\mathbf{u}_j = \mathbf{f}(\mathbf{p_j}) \tag{8.15}$$

where $\mathbf{u}_j$ is the displacement describing unconstrained vertex j's movement, and $\mathbf{f}(\mathbf{p_j})$ is the interpolant, a set of radial basis functions (RBF) of the form:

$$\mathbf{f}(\mathbf{p}) = \sum_i \mathbf{c}_i \phi(||\mathbf{p} - \mathbf{p}_i||), \tag{8.16}$$

where $\mathbf{c}$ is a vector of weights describing each RBF's contribution to the interpolation, $\mathbf{p}$ is the current unconstrained point to be interpolated and $\mathbf{p_i}$ is the current constrained point Pighin et al. [2006]. After experimentation, we define $\phi(\mathbf{r}) = e^{-r/(3/32)}$ as this gives a smooth interpolation. Figure 8.20 shows an example of the process where the Morpheus geometry has been warped to the mean landmark configuration in the AAM. Figure 8.21 shows various frames after warping the Morpheus geometry to different landmark positions captured by the AAM tracker.

The Morpheus mesh was warped to the landmarks from each frame of the original dataset and then saved in a form readable by Autodesk Maya (Wavefront OBJ). The movements from the original video had effectively been transferred onto the geometry of the Morpheus rig. The controllers were solved for, providing a set of activations which minimised the euclidean distance between the rig deformed to the AAM features and the rig deform by the Maya controls. The Nelder-Mead downhill simplex optimisation Nelder and Mead [1965b] was used as it is more efficient than an exhaustive approach, is relatively simple to implement and does not require the derivatives of the function being minimised (which in this case would be the instantaneous change between the controller movement and the movement of vertices in the mesh, information which Maya does not provide). Each $n + 1$

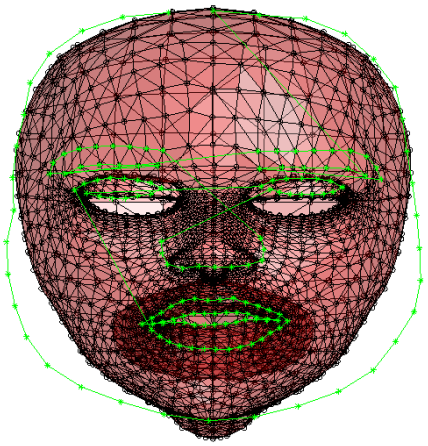


Figure 8.20 The Morpheus geometry warped to the mean AAM landmarks using scattered data interpolation. Note how the geometry (red) now matches the AAM points (green) around the mouth, eyebrows, top of the head and bottom of the chin. The eyes and nose are disregarded as they provide no articulatory or emotional information.

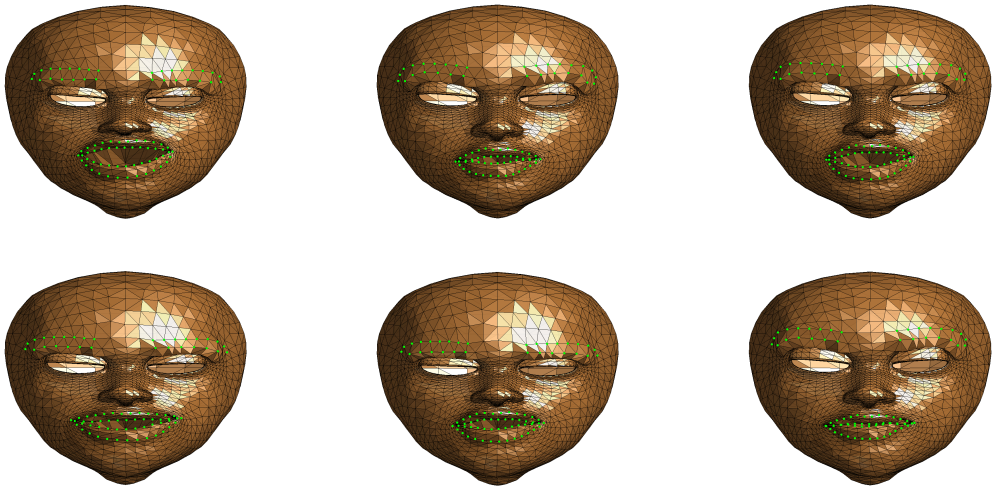


Figure 8.21 The Morpheus geometry warped to various shape configurations as captured by the AAM tracker.

vertex of the Nelder-Mead simplex in n dimensional space represented the inputs to the n dimensional function to be optimised. In this case, the inputs were the continuous activations for the 30 controllers of the Maya rig, and the output was the mean squared error between the rig deformed by the Maya controllers and the rig deformed by AAM features and SDA. The algorithm was implemented as a C++ plugin for Maya. See Figures 8.22, 8.23, 8.24 and 8.25 for examples of the rig after having been fitted to the warped geometry. Further examples can be seen in Appendix E. The original movements from the AAM tracked data were therefore encoded in a geo-spatially discrete feature which was potentially useful in that such features would allow independent control over movements which relate to expression (e.g. movements of the eyebrows and certain types of mouth movement such as curling of the lips), and those which relate to speech i.e. the articulatory movement of the lips.

8.7 Modulation of Neutral Speech in Rig Space

A similar approach to Section 8.2 was used to modulate neutral speech with expressive styles, but instead of AAM features, rig activation features were used. An ICA model was trained on a collection of happy, sad, angry, surprised and Neutral time aligned rig activation features using Equation 8.1 and training data arrange thus:

$$\mathbf{r}_m = \begin{vmatrix} \mathbf{r}_{angry} & \mathbf{r}_{angry/neutral} & \mathbf{r}_{happy} & \mathbf{r}_{happy/neutral} & \cdots \\ \mathbf{r}_{sad} & \mathbf{r}_{sad/neutral} & \mathbf{r}_{surprised} & \mathbf{r}_{surprised/neutral} & \end{vmatrix} \tag{8.17}$$

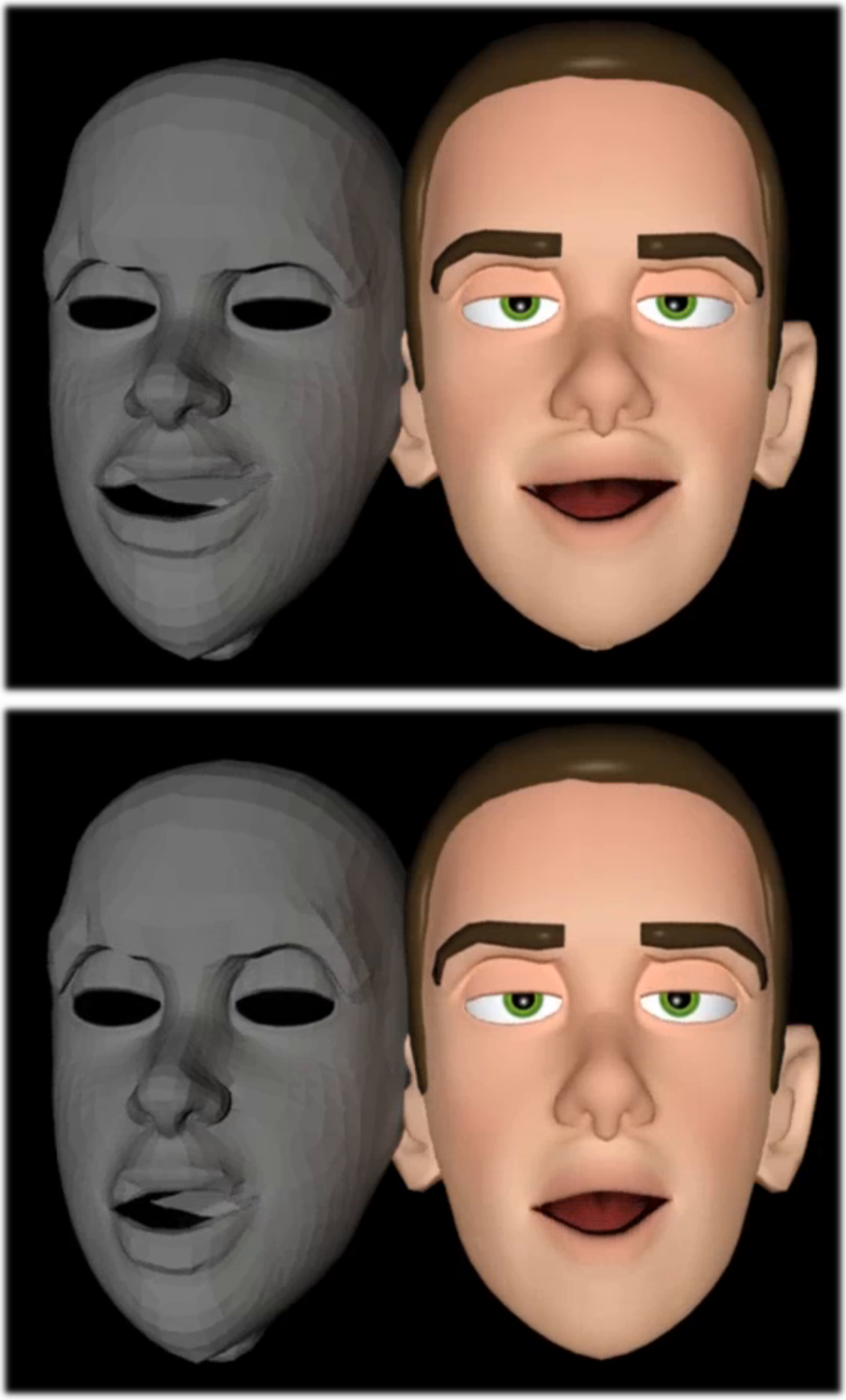


Figure 8.22 Results of fitting the Morpheus rig controls to the AAM fitted geometry. The grey mesh is the AAM fitted geometry having been warped with SDA to AAM captured landmarks, the textured mesh is the Maya rig controlled geometry with controller activations discovered by Nelder-Mead optimisation.

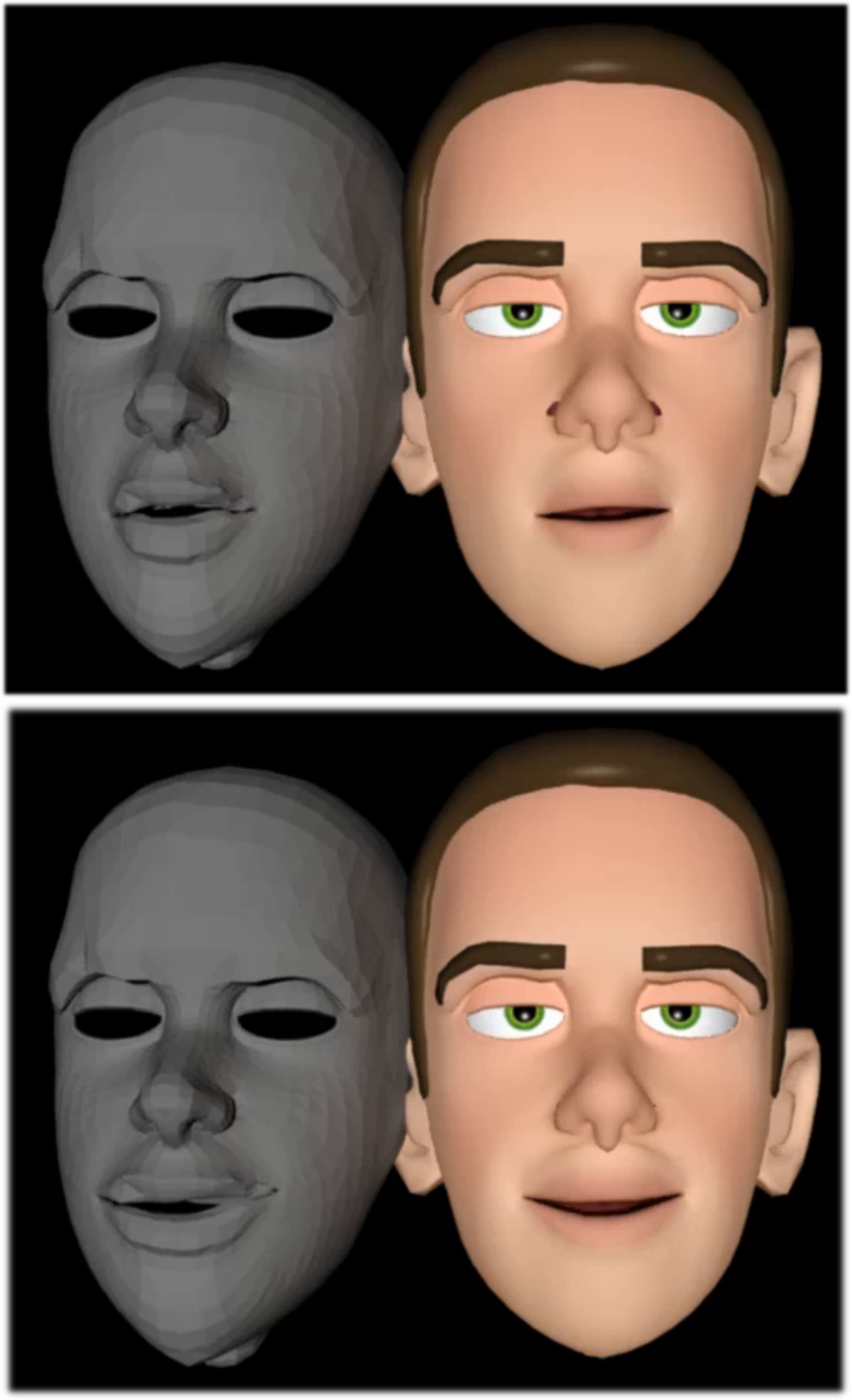


Figure 8.23 Results of fitting the Morpheus rig controls to the AAM fitted geometry. The grey mesh is the AAM fitted geometry having been warped with SDA to AAM captured landmarks, the textured mesh is the Maya rig controlled geometry with controller activations discovered by Nelder-Mead optimisation.

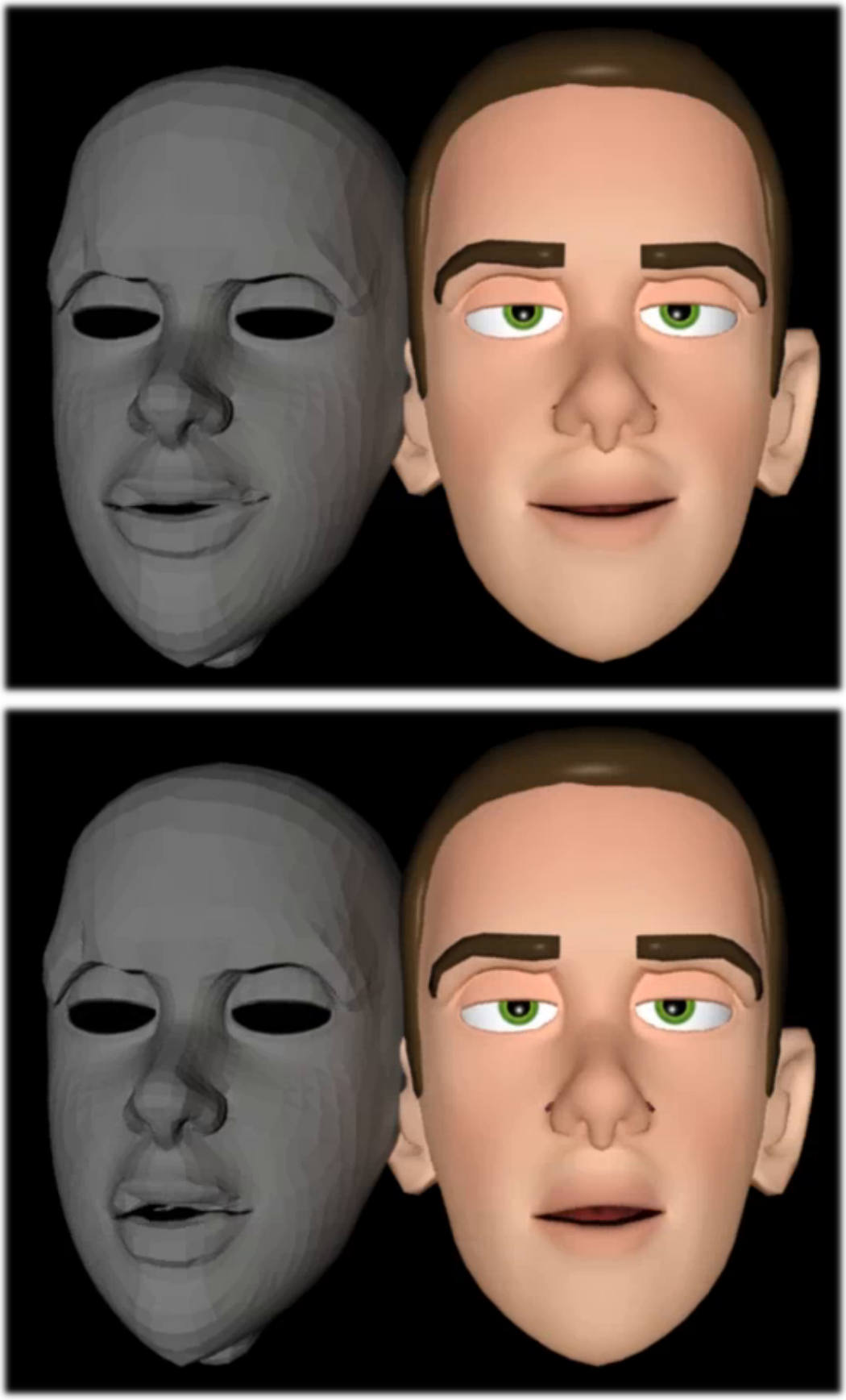


Figure 8.24 Results of fitting the Morpheus rig controls to the AAM fitted geometry. The grey mesh is the AAM fitted geometry having been warped with SDA to AAM captured landmarks, the textured mesh is the Maya rig controlled geometry with controller activations discovered by Nelder-Mead optimisation.

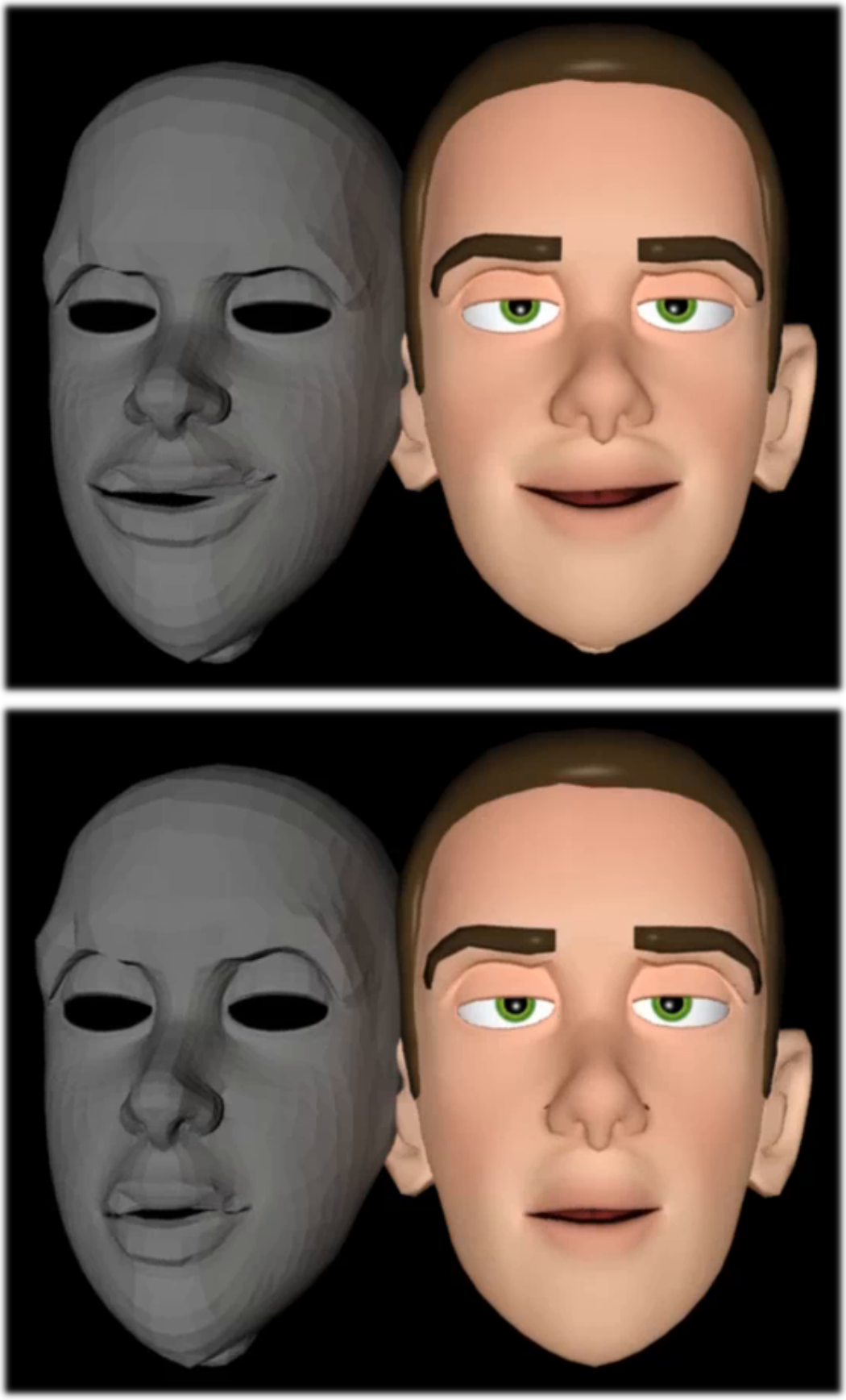


Figure 8.25 Results of fitting the Morpheus rig controls to the AAM fitted geometry. The grey mesh is the AAM fitted geometry having been warped with SDA to AAM captured landmarks, the textured mesh is the Maya rig controlled geometry with controller activations discovered by Nelder-Mead optimisation.

where **r** are the various styles of rig activation features. Appropriate modulation offsets were calculated and applied by projecting time aligned expressive and neutral sequences into ICA space using the model in Equation 8.18

$$r_{ica} = W_{rig} r, \tag{8.18}$$

and calculating the difference between the mean values of corresponding modes in the neutral and expressive styles as in Equations 8.4, 8.5 and 8.6. After these offsets were applied, the modes thought to represent speech were boosted as in Equations 8.10 and 8.11 to account for the increased amplitude associated with expressive speech. The outputs from this process could then be directly applied to the Morpheus model for rendering.

8.8 Results

Since the Nelder-Mead optimisation solved for the positions of 30 different rig controllers, each feature in the dataset representing a single frame was a 30 dimensional vector containing the corresponding activations for each rig controller. In order to avoid any loss of data in the ICA projection (a linear transform), the FastICA algorithm was instructed to return 30 independent components, and square transformation matrices **Q** and **W** of dimension $[30 \times 30]$. Figures 8.26, 8.27, 8.28 and 8.29 show some time trajectories of various ICA modes of decomposed neutral and expressive test sentences. Speech related components and expression related components are shown for each expressive style. Interestingly, mode six appears to be responsible for expression in all expressive styles in the dataset. Each expression also has one or two additional modes which appear to be responsible for the characteristic movements of that style. Expressive modes for different styles are

as follows: for sad expressions, modes 6, 7 and 17; for happy expressions, modes 6 and 11; for angry movements, modes 6 and 18; for surprised movements, modes 6 and 12.

For each frame, the activations for the feature were applied to the corresponding Morpheus rig controllers. This set the rig into the correct pose for each time step. A keyframe was then set so that the controller positions would be saved. The whole animation was then rendered using Maya's builtin rendering module. Figure 8.30 shows the results of the expressive modulations applied to the Morpheus rig. Corresponding frames are shown for Neutral, happy, sad, surprised and angry expressions. Further examples can be seen in Appendix F.

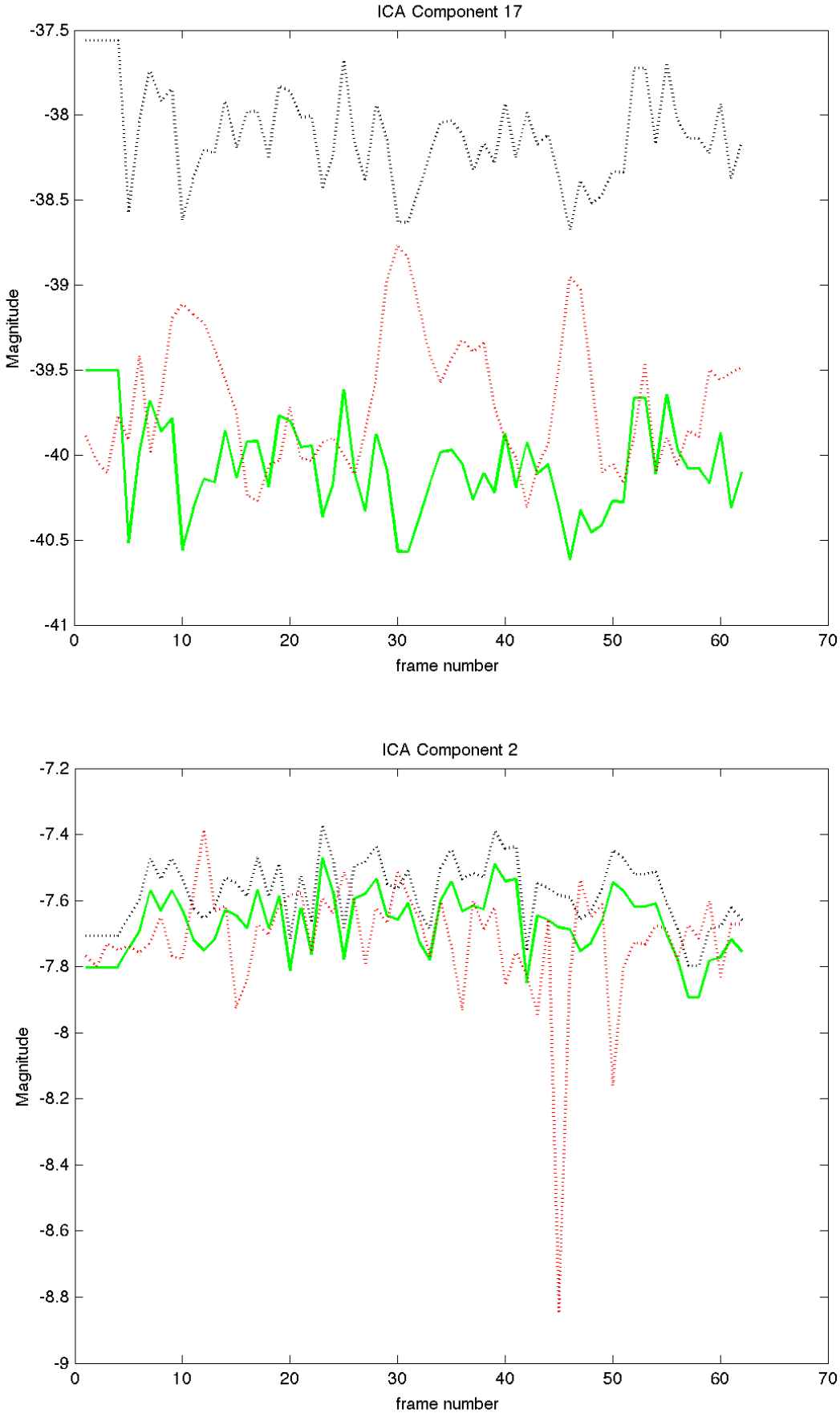


Figure 8.26 Top: Mode 17 capturing sad movements, Bottom: Mode 2 capturing speech movements. Black is neutral ground truth, red is expressive ground truth, green is modulated neutral.

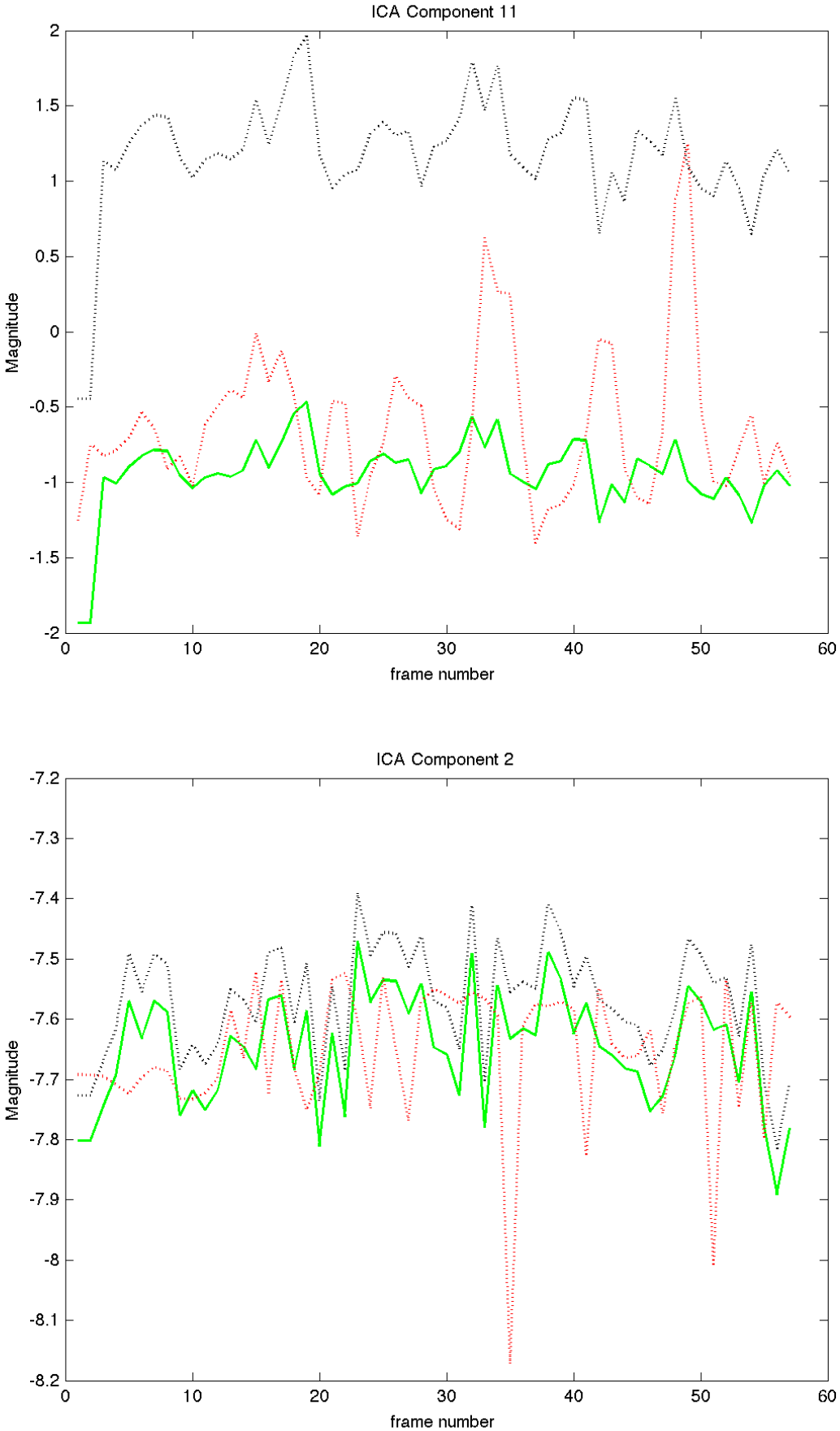


Figure 8.27 Top: mode 11 capturing happy movements, Bottom: Mode 2 capturing speech movements. Black is neutral ground truth, red is expressive ground truth, green is modulated neutral.

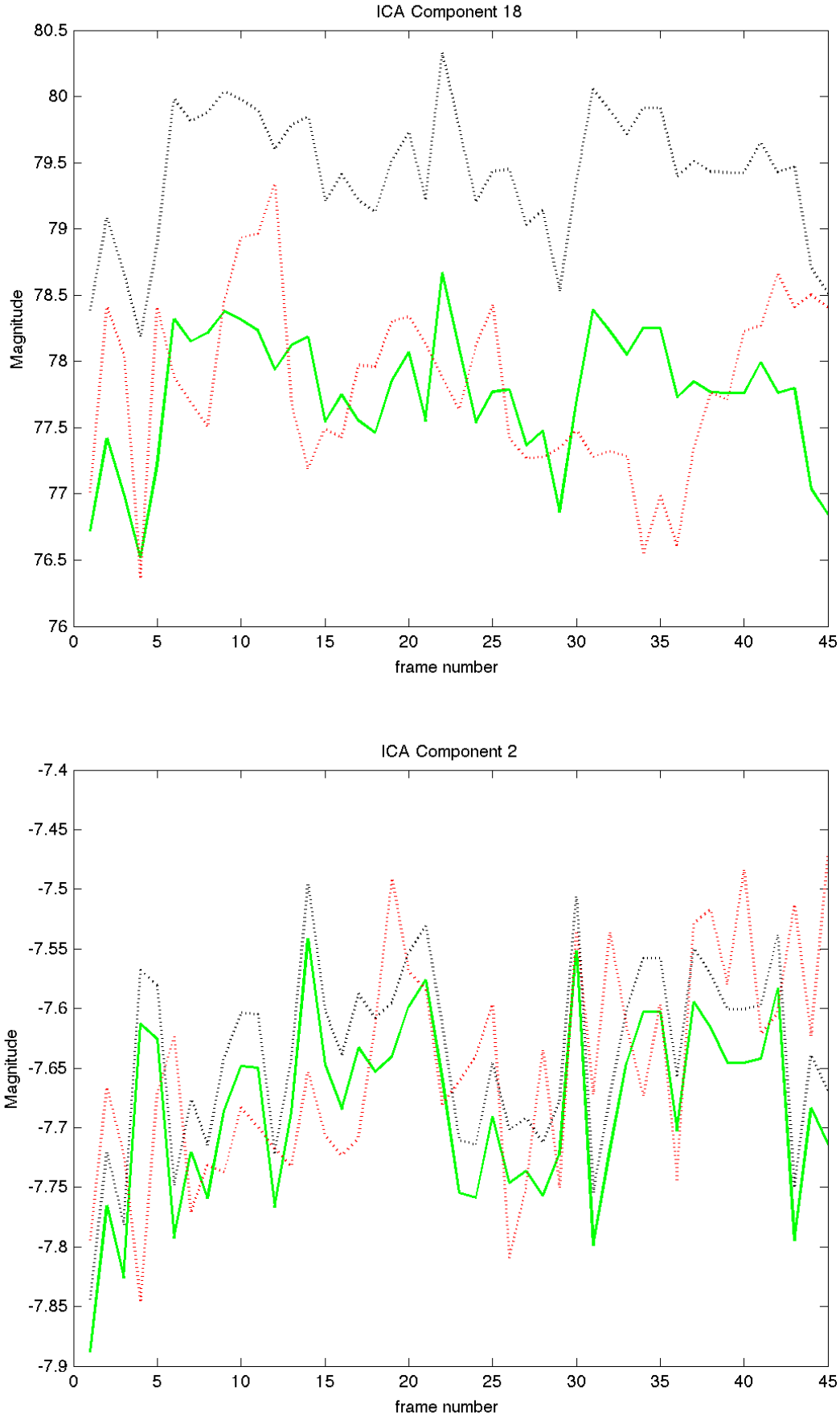


Figure 8.28 Top: Mode 18 capturing angry movements, Bottom: Mode 2 capturing speech movements. Black is neutral ground truth, red is expressive ground truth, green is modulated neutral.

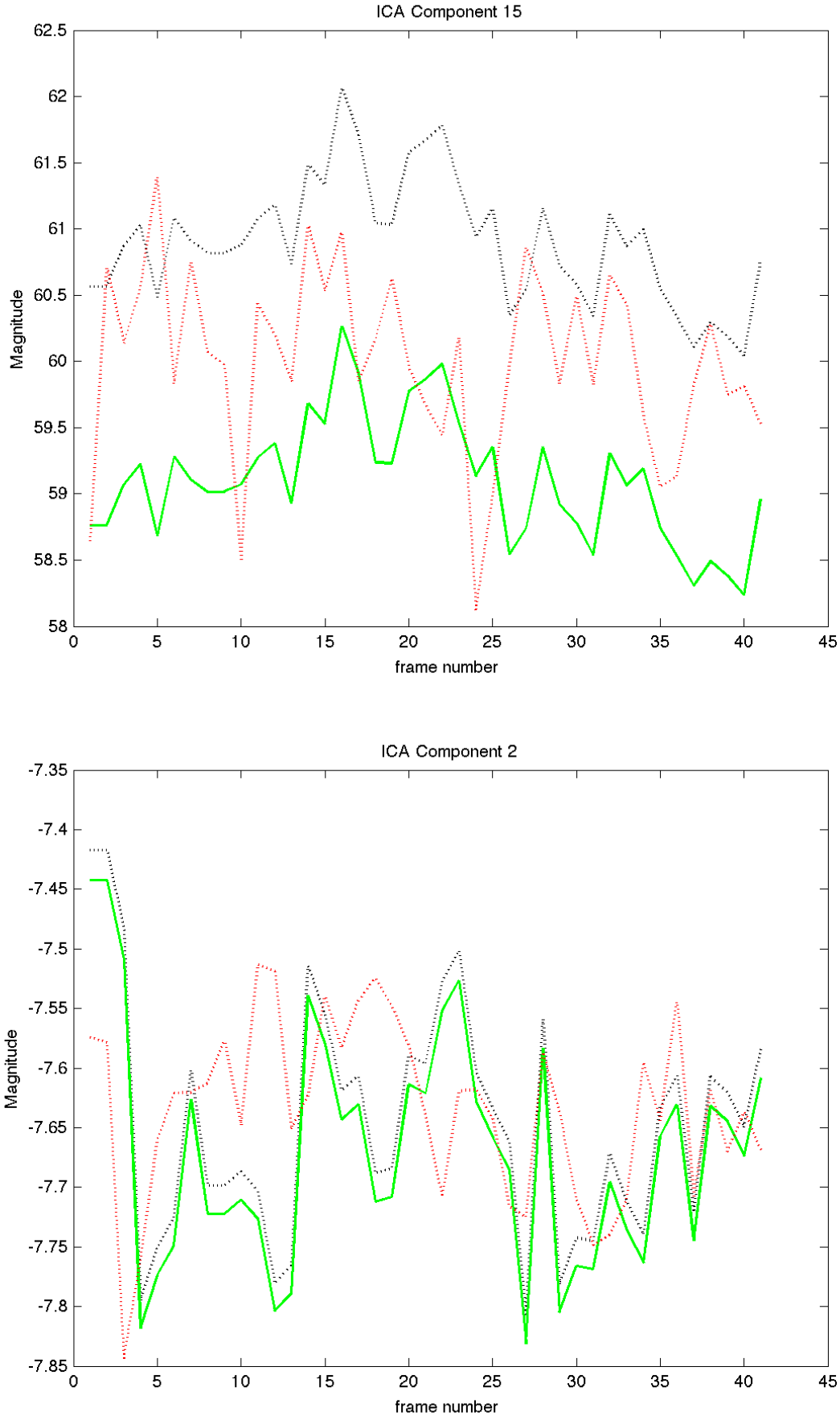


Figure 8.29 Top: Mode 12 capturing surprised movements, Bottom: Mode 2 capturing speech movements. Black is neutral ground truth, red is expressive ground truth, green is modulated neutral.

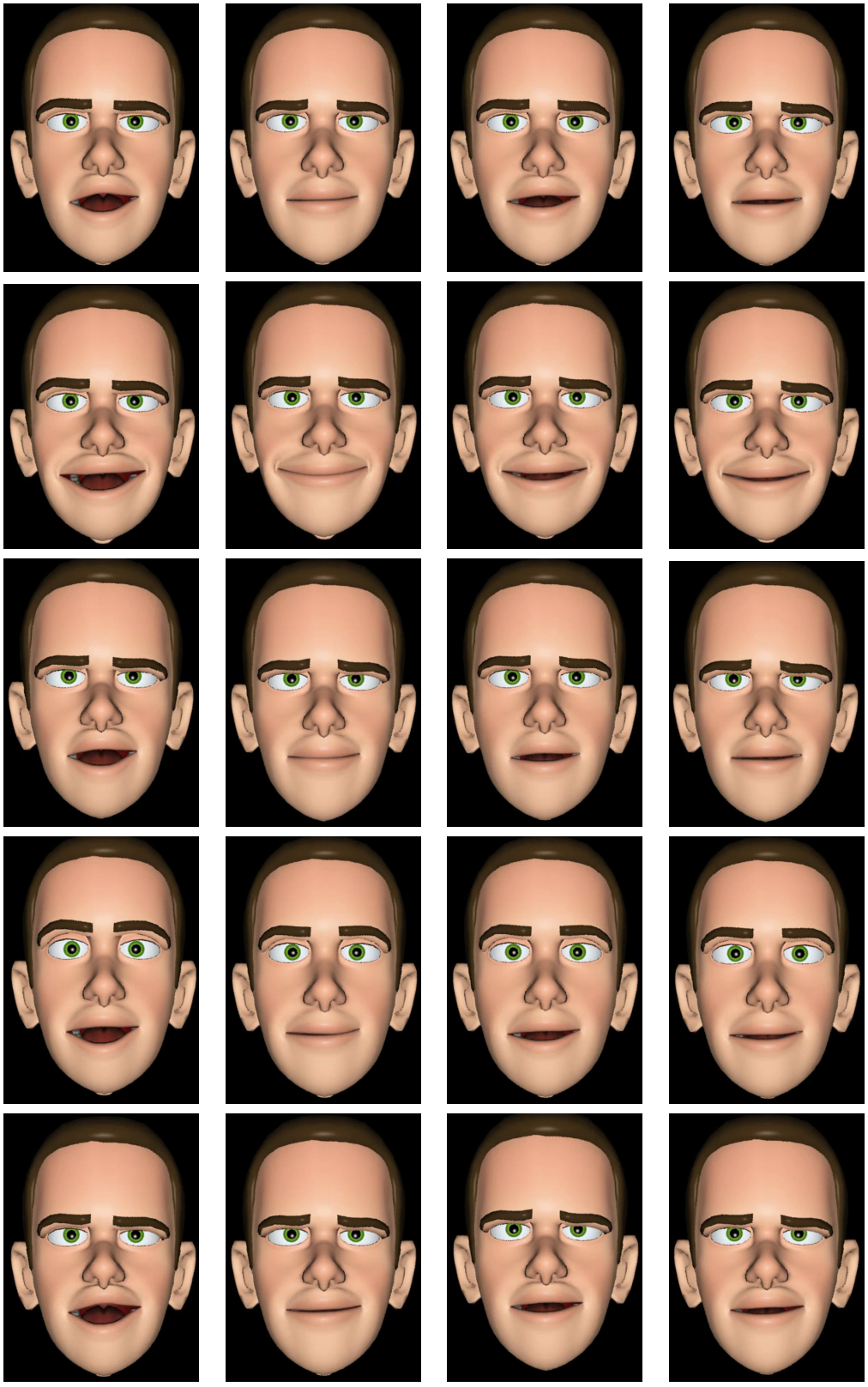


Figure 8.30 First row is neutral ground truth frames showing different mouth shapes, the second row shows the same fames after modulation with happy, third row sadness, forth row surprise and fifth row anger.

8.9 Evaluation of Graphical Output

8.9.1 Turing Test

To evaluate the overall performance of this approach, a Turing Test was conducted. Eight participants were each shown 40 sequences in a random order. Half were ground truth expressive and half were modulated neutral. The participants were told that half the sequences were real and half were synthesised. They were simply asked to decide for each sequence whether it was real or synthesised. Sequences were all shown alongside the held out expressive audio. To determine the statistical significance of this result, we performed Wilcoxon's Signed Rank test. This showed that participants were not reliably able to discriminate between real and synthesised sequences. A p-value of >0.8, means we and cannot reject the null hypothesis that the population mean ranks are the same.

8.9.2 Mean Opinion Score Test

A second experiment was conducted to show that the modulation of expression onto neutral speech leaves the speech component intact. Eight participants were each shown 45 sequences in a random order, a third of which were ground truth expressive, another third of which were ground truth expressive having been time aligned to a different audio track, and the final third of which were modulated neutral. Participants were instructed to pay particular attention to the mouth articulation and how well this matched the audio. The audio tracks for the second and third treatments were from the left out expressive sequences (as described in Section 8.3). This was done to remove possible bias which would occur if the ground truth expressive video was played with its original audio (since this matched perfectly). The ground truth played alongside its original audio was added to the

experiment as a control. Participants were told to assign a score to each sequence from 1 to 5, 1 being very poorly synchronized with inappropriate mouth movements and 5 being perfectly synchronized. The experiment was conducted using a custom test application which randomised the ordering of the movies, and allowed scoring via the use of a slider. The slider ranged from 1 to 5 and recorded the response to 3 decimal points. Participants were only allowed to see each sequence once. The overall mean opinion score for ground truth expressive with time alignment was 3.28 and for the modulated neutral was 3.20. Surprisingly the mean opinion score for for the ground truth with no time alignment and its original audio was 2.75. It is thought that this counter intuitive result is due to the fact that the time alignment process smoothes the animation slightly and that people find this pleasing. This also corresponds to the result reported in section 7.6. An ANOVA (Analysis of Variance) test was used to calculate significance. A p-value of <0.01 means we reject the null hypothesis that the group means are the same. Further, using Tukey's honestly significant difference criterion, we find that there is no significant difference between the synthesised and time aligned treatments, and that there is a significant difference between these two treatments and the ground truth with no alignment. Figure 8.31 shows the mean opinion scores and their 95% confidence intervals.

8.10 Blending of Expression in ICA Space

As was mentioned in Section 8.1, human feeling is much more complicated than the so-called six universal emotions of anger, disgust, fear, happiness, sadness and surprise which Ekman [1992] proposed are common to all human societies, a proposition which is refuted in Jack et al. [2012], where is stated that there is more

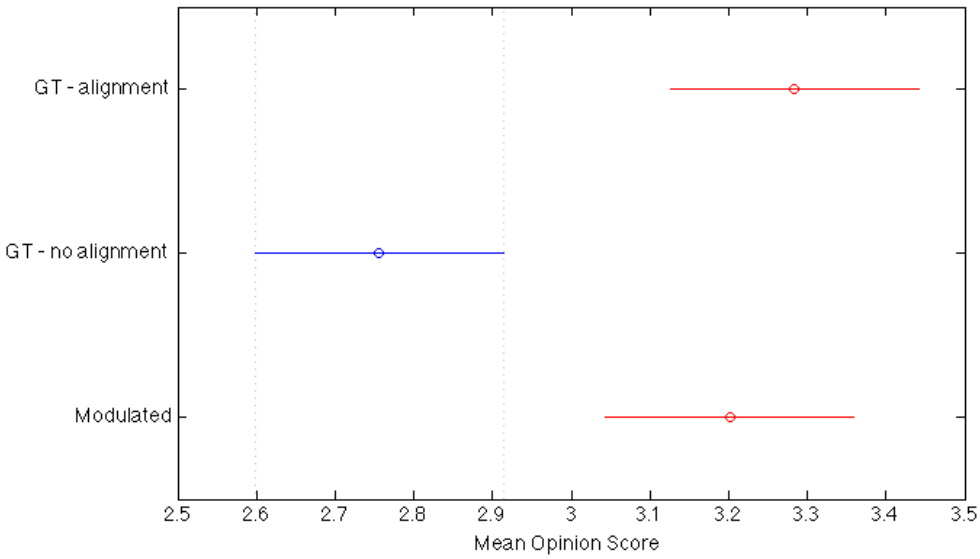


Figure 8.31 Mean Opinion Scores with 95% confidence intervals

variation in the expressions of East Asians than there is in Caucasians. At the very least there are degrees or magnitudes of emotion, from mild contentment to stunning euphoria, passing despondency to heart-breaking despair. But usually, instead of simple categorical emotions, we feel many different things at once. In reality emotion is a subtle mix of many different feelings, thoughts and stimuli. Indeed it is entirely possible to feel two seemingly contradictory emotions simultaneously e.g. happiness at the birth of a child, coupled with sadness that a deceased relative is not present to share in such joy. Consequently, human expression of such mixed emotions is equally intractable. Any technique aiming to synthesise expressive speech animation must attempt to model this gamut of expression. If not, it is condemned to simply create idiosyncratic caricature likely to inspire derision and irritation in its viewers.

This section details an experiment which shows that the ICA techniques described previously are capable of modulating neutral visual speech sequences with

expressions within the gamut of expression on which the ICA model was trained, and not just with the categorical extremes.

This work again used the BiWi audiovisual expressive corpus Fanelli et al. [2010b], chosen for the extremes of its expressive gamut, but also the subtle, emotionally ambiguous sequences which also appear. Firstly a PCA based point distribution model was trained to capture the variance across the entire dataset. Then three different ICA models were produced. One was trained on time aligned matching pairs of neutral and sad sequences another on time pairs of neutral and angry sequences and another on matching pairs of neutral, angry, sad, happy and surprised sequences. Test sentences were selected (held out from the training data) for their unusual or particularly strong expressive content. The dataset contained each sentence in both an expressive and a neutral style, time aligned to account for slight differences in phonetic alignment. For each test sentence, the most expressive frame(s) in the expressive version were considered. The corresponding neutral frames were projected into the ICA space of each of the three models thus:

$$\mathbf{s}_{sad} = \mathbf{W}_{sad} \cdot \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{8.19}$$

$$\mathbf{s}_{angry} = \mathbf{W}_{angry} \cdot \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{8.20}$$

$$\mathbf{s}_{Mixed} = \mathbf{W}_{Mixed} \cdot \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{8.21}$$

where $\mathbf{W}_{sad}$, $\mathbf{W}_{angry}$ and $\mathbf{W}_{Mixed}$ are the estimated un-mixing matrices for the three training sets, $\mathbf{P}$ are the orthogonal basis vectors of the point distribution model describing 99% of the variance from the mean, $\mathbf{x}$ is the current neutral test frame, $\bar{\mathbf{x}}$ is the mean test frame from the training set, and $\mathbf{s}$ are the independent components of $\mathbf{x}$.

The ICA parameters encoding the neutral frame were then used as input to

an implementation of the Nelder-Mead downhill simplex optimisation (see Section 6.6), in which the error between the ground-truth expressive equivalent frame and the neutral modulated frame was optimised for each model thus:

$$\epsilon_{sad} = f(\mathbf{s}_{sad}, \mathbf{W}_{sad}) \tag{8.22}$$

$$\epsilon_{angry} = f(\mathbf{s}_{angry}, \mathbf{W}_{angry}) \tag{8.23}$$

$$\epsilon_{Mixed} = f(\mathbf{s}_{Mixed}, \mathbf{W}_{Mixed}) \tag{8.24}$$

where ϵ_{sad}, ϵ_{angry}, and ϵ_{Mixed} are the minimum RMS errors describing the fidelity with which each of the corresponding ICA models $\mathbf{W}_{sad}$, $\mathbf{W}_{angry}$ and $\mathbf{W}_{Mixed}$, is able to modulate the neutral frame to approximate the ground-truth expressive equivalent frame.

The expectation was that the mixed model would be able to better approximate a wider range of expressive styles and magnitudes of emotion than either of the simpler models since this greater range of expressive variance was present in its training gamut. Producing intermediate poses within this gamut should be possible by regulating the projected ICA components. This is an important feature of the technique as it means that it would be possible to create expressive interpolations which are subtle mixes of the expressions in the training gamut, as opposed to simply creating categorical and unrealistic caricatures of expression.

8.11 Results

The optimised ICA features returned from the Nelder-Mead algorithm can be inverted and projected back into point distribution space thus:

$$\mathbf{x}_{sad} = \bar{\mathbf{x}} + \mathbf{P} \cdot \mathbf{A}_{sad} \cdot \mathbf{s}_{sad} \tag{8.25}$$

$$\mathbf{x}_{angry} = \bar{\mathbf{x}} + \mathbf{P} \cdot \mathbf{A}_{angry} \cdot \mathbf{s}_{angry} \tag{8.26}$$

$$\mathbf{x}_{Mixed} = \bar{\mathbf{x}} + \mathbf{P} \cdot \mathbf{A}_{Mixed} \cdot \mathbf{s}_{Mixed} \tag{8.27}$$

where $\mathbf{x}_{sad}$, $\mathbf{x}_{angry}$ and $\mathbf{x}_{Mixed}$ are the best approximations for the corresponding ICA models in point distribution space, $\bar{\mathbf{x}}$ is the mean configuration in the point distribution model, $\mathbf{P}$ is the set of orthogonal basis vectors, $\mathbf{A}$ is the corresponding ICA mixing matrix for each expressive ICA model, and $\mathbf{s}_{sad}$, $\mathbf{s}_{angry}$ and $\mathbf{s}_{Mixed}$ are the set of ICA features optimised by Nelder-Mead to best approximate the expression in the test expressive equivalent. Figures 8.32 and 8.33 show the renderings of several approximations. Note how although the single expression ICA models have reasonably approximated the ground-truth frame, they still maintain some of the expressive characteristics of the data on which they were trained, whereas the approximation produced by the mixed model is much closer to the ground-truth. We believe that this demonstrates that the mixed ICA model approach is capable of producing the subtle mixes and magnitudes of different expressions required for a realistic representation of human visual speech. See Appendix G for more examples.

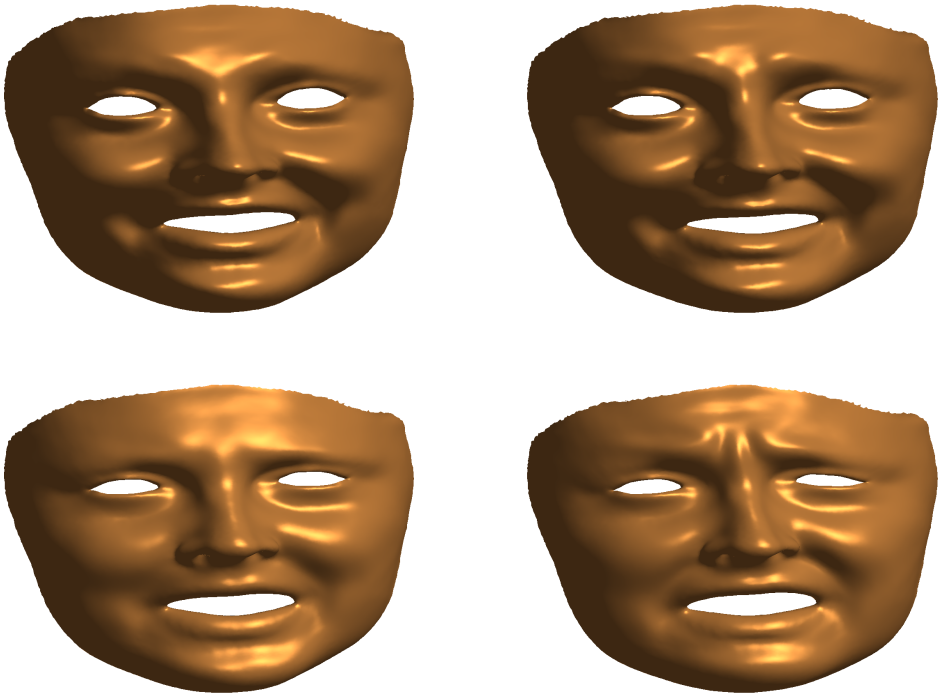


Figure 8.32 Clockwise from top-left: Expressive ground-truth, mixed model approximation, sad model approximation, angry model approximation

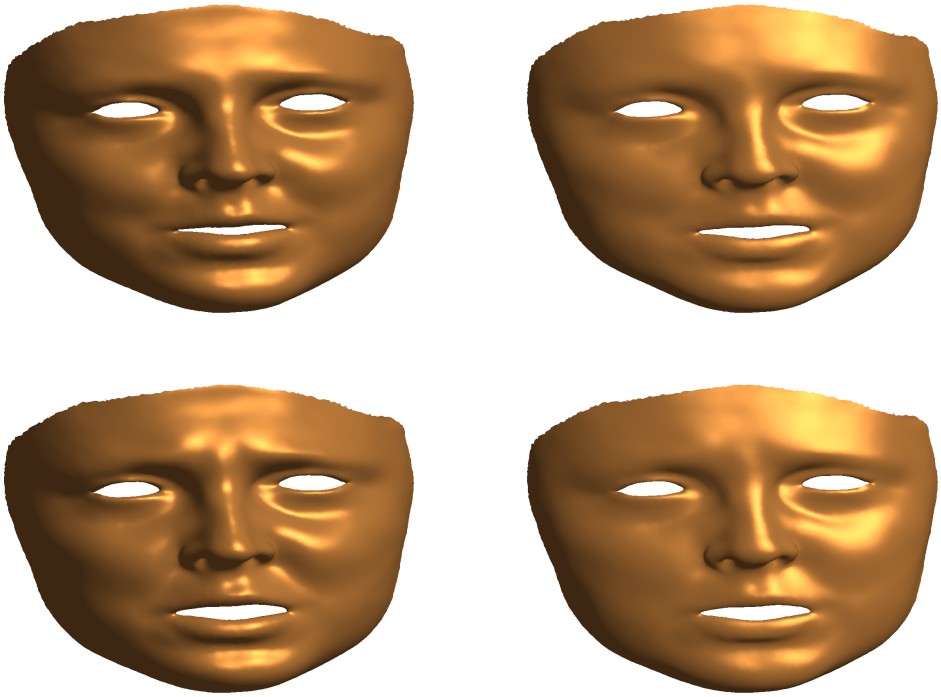


Figure 8.33 Clockwise from top-left: Expressive ground-truth, mixed model approximation, sad model approximation, angry model approximation

8.12 Summary

This chapter has described further extensions to the the work described in Chapter 7, which in turn was based on the work of Cao et al. [2003]. It was shown that when an Independent Component Analysis model is trained on many expressive styles, it discriminates between speech and the component expressions by redistributing the energy within a visual speech sequence onto various modes. The fact that it does this means that speech and the various expressive styles are to a large degree independent from one another. ICA therefore provides controls which allow the independent manipulation of these different characteristics of expressive animation. However, there will inevitably be some co-variation between speech and expression. For example, when a person expresses surprise, the eyebrows often raise at the onset of the sentence which will invariably coincide with some prominent articulatory movement. Therefore, the signals are not entirely independent and ICA is not capable of completely factorising speech from expression. However, it does a good enough job to be useful as shown in the subjective testing reported in Sections 8.9.1 and 8.9.2.

The chapter went on to show that it is possible to project a neutral sequence onto the independent components of an ICA model trained on a variety of expressive styles, and then manipulate the sequence in ICA space in order to modulate the neutral sequence with any of the expressive styles in the original training set. This was demonstrated firstly on video data by using an Active Appearance Model, where AAM features were the input and output to the modulation process. The modulated neutral outputs were subsequently projected back onto the AAM model in order to make video sequences. Subjective tests were carried out which showed that viewers were not able to significantly tell which videos were ground truth and which had been subjected to the modulation process. However, lower than

ground truth mean opinion scores are accounted for by the fact that AAM output is prone to various types of visual distortion such as blurring and mesh tearing. It is possible that it was the AAM output which was responsible for this difference.

To tackle this problem the technique was retargeted to an animated rig. The geometry from a facial rig was warped to the shape of the aligned AAM tracked landmarks for each frame of the entire training set. Then the facial movements of the training set were transferred into the facial rig's controller space using a Nelder-Mead downhill simplex optimisation. This had the interesting effect of transforming the features from a global transform feature, where each element of the feature affects the entire facial pose, to a localised transform feature where each element affects only a localised region of the facial geometry. It is possible that ICA would work more effectively on such a feature set since certain types of expression are only conveyed through localised areas of the face (surprise for example is largely conveyed with the eyebrows) and would therefore only ever manifest through certain rig controllers. After this data transformation was complete and the dataset was in the rig controller activation space, the same procedure was carried out, the ICA modulation taking place on the raw rig controller activations. This time there was no significant difference in the user's preferences during the evaluative testing, meaning that participants were unable to tell the difference between the expressive ground truth and the modulated neutral outputs and displayed no particular preference for ground-truth or modulated output.

To show that the technique is able to produce facial poses representing mixes of different expressions and not just categorical facial expressions (an important feature of any expressive synthesiser), a further experiment was conducted. Three ICA models were trained, one on a neutral/sad training set, one on a neutral/angry training set, and an third on a neutral/happy/sad/surprised/angry training set. A neutral frame from a test sequence was projected onto the independent components

of each ICA model, and the Nelder-Mead downhill simplex optimisation was used to derive the ICA coefficients which minimised the error between the neutral frame and its expressive equivalent. The frames were chosen for either their unusual or their strong expressive content. It was found that the mixed ICA model was much better able to manipulate the neutral frame to approximate the expressive equivalent frame than either of the single expression models. Whilst the single expression models were able to approximate the reference expressive frame with regard to mouth shape, a residual of the single expression on which the ICA model had been trained was still apparent. This demonstrates that the mixed model is able to modulate neutral speech within the gamut of the different expressions on which it was trained at least for static frames. There is however the possibility that temporal effects may come into play if the technique were for video.

Chapter 9

Conclusion

The aim of this project was to investigate methods of producing expressive visual speech animation. Existing methods tended to focus on creating ever more accurate articulatory mouth movements, yet opinion seemed to suggest that viewers still thought animation produced with such methods lacked realism. It was thought adding additional modalities to the animation, such as the synthesis of head movement, eye gaze, gesture and in particular, expression, would improve the perceived realism. Therefore, the automatic production and addition of expression for visual speech became the research focus of this project.

9.1 Aims Addressed

As was stated back in section 5.6, this work sought to address several existing problems with expressive visual speech synthesis. With reference to those sections, the proposed systems solutions are as follows:

1. Coarticulation: Many existing techniques suffered from co-articulation arte-

facts. The articulatory movements we make with our mouths are only loosely correlated to the speech sounds they produce, and a particular sound may be produced by many different lip shapes, depending on phonetic context. The co-articulation problem then is where an inappropriate lip shape is selected to articulate a particular sound. Many existing techniques made the assumption that there is a one-to-one mapping between phonemes and visemes, which has been shown to be invalid Taylor et al. [2012]. There is in fact a many-to-many relationship between phoneme and viseme meaning that techniques which do not model this mapping will often pick an inappropriate viseme to articulate a particular phoneme, leading to a kind of "dubbed" effect. The technique presented in this work circumnavigates the problem of co-articulation. The mixed expressive speech signal is projected onto independent components some of which represent mostly speech and some which represent mostly expressions. By only manipulating those components representing expression, the speech is left intact (as shown by the earlier qualitative evaluation) meaning that no co-articulation artefacts are introduced. Therefore, as long as the input neutral speech signal is free from co-articulation, so will the output expressive speech signal. Therefore in a complete system, the speech component of a synthesis could be provided by an HMM based synthesiser using state to account for sparseness, and providing an input signal to our system which is free of co-articulation artefacts.

2. Training Data: Most techniques for producing expressive visual speech animation require large corpora of training data. For example Anderson et al. [2013] collected 6925 sentences for their PCA based expressive synthesiser. Given a conservative estimate of 5 seconds per sentence, that equals over 9 hours of footage. If the footage was recorded at 25 fps, then there would be over 800,000 frames of input data to label and parameterise. Alternatively,

Cosatto and Graf [2000b] use a much smaller training set of 14 sentences, but the parameterisation process is relatively complex requiring image segmentation and measurement of strategic features. More typically, Tao et al. [2009] describe an HMM based synthesiser which used 900 sentences of training data. Even with this training set, a conservative estimate puts the number of frames needing to be labelled and tracked at well over 100,000. Accurate labelling of training data along with automated parameterisation is a time consuming and error prone process, requiring considerable amounts of tedious labour.

Our technique by contrast requires only very small amounts of data. The experiments presented in this work used at most around 16,000 frames of training data or 14 sentences in each of the expressive styles (based on time aligned neutral, sad, angry, happy and surprised, recorded at 30 fps with an average sentence length of 5 seconds). Indeed we have observed that the system works with considerably less training data, with as few as three or four pairs of expressive and neutral sentences producing realistic expressive output.

3. Expressive Controls: As has been repeatedly shown throughout this work, ICA provides the means to separate speech and expression into different independent components. These components can then be manipulated independently of speech. Furthermore, in an ICA model trained on a mix of different expressive visual speech styles, different modes can be manipulated to provide a continuous mix of the expressions on which the model was trained. This therefore provides the animator with a set of controls which can be used to change the expressive style of unseen neutral footage.

4. Expression Blending: An important aspect of expressive visual speech syn-

thesis is the ability to model the subtle mixes of human emotions which are visible through our facial expressions. Humans rarely feel categorically happy or sad or angry about something. More typically there is a mix of feelings.

By capturing the "extremes" of a categorical expressive gamut, we have shown that it is possible to interpolate between these extremes to not only create less intense versions of the categorical expressions, but to blend expressions within the gamut in order to output new expressions not originally in the training set, which better mimic the way humans outwardly express their complex internal set of mixed feelings and emotions.

5. Time Complexity: Another issue with examples of previous work in this field is time complexity. For example, the system proposed by Sifakis et al. [2005] solved for the position of 32 transversely isotropic pseudo muscles in a mesh of 370,000 tetrahedrons using gauss-newton gradient descent. In 2005 on a single Xeon 3.06 Ghz CPU the technique took 8 minutes per frame to solve the muscle activations. In Terzopoulos and Waters [1990] the authors report that a Silicon Graphics Iris 4D-240GTX workstation was able to render their system of tri-layered tetrahedral spring loaded meshes at 8 Hz which is nothing like real time. Although computer technology has moved on significantly, the techniques need to be judged in terms of the hardware that was available at the time, since as the power of hardware increases, so does the complexity of animation models and the demands of the viewing public.

   By contrast, our technique is computationally straightforward. Estimating the ICA mixing and unmixing matrices is the most computationally intensive part of the process. The FastICA algorithm implemented in Matlab completes this task on a training set of several thousand frames in a matter of seconds on a Macbook pro with a 2.4Ghz Intel Core i5, with 8GB of RAM.

Once the ICA model has been estimated, modulating neutral speech with expression simply involves two linear projections, operations for which Matlab is highly optimised. For example, running on the hardware described above, Matlab is able to multiply 2 randomly generated matrices of size (100 x 100), twice in 0.001291 seconds. Doing this 25 times takes 0.025824 seconds (for real-time applications, the process needs to be able to run at least 25 times every second). Therefore statistical techniques such as these allow for applications to be run on standard consumer level equipment (i.e. laptops and tablet computers) and for a broader range of applications such as retargeted conversational agents in a therapeutic setting such as described in Leff et al. [2013] and Huckvale et al. [2013].

6. Flexibility: Many of the existing techniques for producing expressive visual speech animation are tightly coupled to their training data. For example, a concatenative synthesiser selects either whole images or parts of images to be stitched from a corpus of existing images. Therefore the output is tightly coupled to the identity of the actor in the training data. If another actor is required, a whole new training set of images is required or further processing is needed for retargeting such as that described in Vlasic et al. [2006]; Curio et al. [2006]. Similarly, for physically based geometry models, outputs will always be in terms of the geometry. Although it is possible to warp such geometry to the identities of different actors such as in Noh and Neumann [2006]; Sumner and Popovic [2004]; Der et al. [2006], the process can be complex, manual, computationally intensive and is not guaranteed to produce valid output.

   Since the method presented in this work is purely data driven, it is largely agnostic to things like actor identity, expressive content type and data type. As has been shown in chapters 7 and 8, the technique works on PCA reduced

point cloud data, Active Appearance Model features, and continuous pseudo muscle activation features. This therefore means that the technique can be applied to a range of data types encoding actor identity, expression and speech which may be encountered in an animation pipeline.

7. Simplicity: Overly complicated methods which are difficult to understand are never popular in science. Particularly if they don't demonstrably work. The most popular ideas tend to be those which are simple, easy to understand, modularly fit into existing problems and provide a more predictable solution than existing methods. The work presented in this thesis is arguably one such method. It is certainly easy to understand, simple to implement and above all, it works as we have demonstrated with three different types of data (AAM features, point cloud data and muscle activation features). It reliably provides a method of converting neutral speech into expressive speech and therefore can be used as a modular component in an animation pipeline.

8. Subjective Testing: Testing and evaluation is notably absent from much of the literature in visual speech synthesis (both expressive and non-expressive). Most commonly, authors simply provide still images as evidence that their techniques work. Although these usually look at least expressive, often the exact expression is ambiguous. It is possible that they would be less ambiguous if full video was provided but this is usually not available. In any case, this can hardly be considered rigorous evaluation. Other papers offer objective evaluations for their techniques, usually comparing feature trajectories or RMS error to ground truth. However as was stated in Section 5.5, Theobald and Matthews [2012] show that there is often little correlation between objective measures of output quality and the subjective opinions of viewers. They found that the objective measure which correlated most highly

with viewer opinion was Dynamic Time Warping (DTW) distance between synthesised and ground truth data. The DTW metric is not used in any of the expressive visual speech synthesis literature.

Visual speech animation and expressive visual speech animation (much like other forms of synthesis such as music synthesis and computer graphics) balance a tightrope between science and the arts. Scientific and technical aspects combine to produce a final product which is consumed by humans, often for pleasure. The most important test for such products is therefore whether human beings find them pleasing or at least plausible. Since there is no known correlate between an objective test and that which humans find pleasing, the only way to properly evaluate expressive visual speech is to produce a quantitive scientific test which is based upon subjective human opinion.

Even in this regard, the literature generally takes a sub-optimal approach to subjective evaluation, usually opting for mean opinion score measures and expression identification. Mean opinion scores are somewhat unreliable. As we observed carrying out our research, opinion for a single sequence can vary wildly as different people tend to concentrate on different aspects of the sequence. For example, although our viewers were asked to judge the sequences based on lip synch and expressive realism, some people reported having marked sequences poorly for things like lack of image sharpness or lip artefacts which are a consequence of AAM output. We therefore put more stock in our Turing test result where viewers were shown a sample of real and modulated sequences and simply asked, for each sequence, which they were watching. This is a simple question which covers all of the possible features of the animation which the participants might be focussing on. Essentially the question elicit's the viewer's *gut* feeling about the sequence, and eliminates

issues to do with rendering artefacts since both ground truth and modulated sequences were projected onto the same output model. We regard this simple test as rigorous since it directly compares ground truth data to synthesised data, and it should be noted that it is a test which very few others are willing to perform.

9.2 Future Work

The work described in this thesis should be considered a proof of concept. We have shown that it is possible to decompose an expressive visual speech signal into components which mostly represent expression, and components which mostly represent speech. This decomposition provides two things. Firstly it allows the expressive components of the signal to be manipulated independently of the speech components. This means that as long as the speech components are largely free from co-articulation artefacts, any output from the system will also be free from such artefacts. Secondly, it provides an intuitive set of controls (where the user can quickly understand what a control will do), which can be adjusted which control only expression. By varying the contribution of these controls, the expressive gamut to which the model was trained may be traversed.

9.2.1 Improved Modulation

So far the work described has used simple techniques for adjusting the modes responsible for expressive content (scaling and translating). These work well and produce convincing expressive output, however this could probably be improved by better learning the relationship between speech and expression in the training data. For instance, a recurrent neural network could be trained to map between

speech components and expressive components in ICA space, thus allowing for temporal expressive contexts to be modelled. Some preliminary work towards this was undertaken but the results obtained were "jerky" and not particularly convincing and have therefore been left out of this report.

9.2.2 ICA Analysis

There are several avenues for further study. With ICA decomposition, it was observed that the expressive modes would tend to represent a spectrum of a certain type of expression. For example, altering a particular mode $\pm 3\sigma$ from its mean would yield expressions which looked either very happy or very distressed. This makes intuitive sense since smiling and raised eyebrows are the characteristic expression of happiness, whereas curling the lips down slightly and frowning are the characteristic expression of distress. So it seems that the ICA identified this type of movement as independent from speech and projected it into a separate mode. More study into exactly what the ICA algorithm is identifying as an independent signal would be interesting and could lead to an enhanced control set for the expressive components of an expressive visual speech synthesiser.

9.2.3 I-vector Analysis

In a similar vane, alternative factorisation techniques could be investigated. DeMarco and Cox [2013] reported some success in discriminating between different regional accents for native English speakers from the British Isles. The work (conducted in the audio domain) used i-Vectors and Linear Discriminant Analysis Fisher [1936] to factorise speech and accent information from a universal background model and then performed Gaussian Mixture Model based classification on the accent information. Since this technique was successful in separating accent

from speech, it is likely that it would also separate expression from speech in the visual domain and could potentially be used instead of ICA to create a model in which to analyse speech and expression independently.

9.2.4 Psychological Experiments

Techniques similar to this have been used in a Psychological context. For example, Theobald et al. [2007b] used similar techniques to transfer facial movements from an actor to a different face model. A participant sitting in another room then had what they thought were several conversations with several people via a video screen. They were in fact having several conversations with the same person, but that person's facial movements were retargeted to several different face models (i.e. of varying gender and ethnicity). Therefore the psychologist was able to observe how the participant behaved differently to what they believed were different types of people, but in fact were all the control. This technique could be used in similar psychological type experiments, for example to observe how people react to men or women displaying different types of emotion.

9.2.5 Training

Another possible use and cause of further work could be training. For example, intelligence agencies and police forces often have the difficult task of detecting deception in suspects during interrogation. It would be possible to build a similar model to ours which is trained on video of people talking truthfully and trying to deceive. Therefore sequences of synthesised truth and deception could be produced in order to train agents to better detect when a suspect is lying.

9.2.6 Animation Pipeline

As has been alluded to throughout this thesis, the aim for this work was to streamline, automate and improve techniques used in the animation and computer games industries to produce realistic talking heads with a particular focus on facial expression. Since the method takes neutral visual speech as input and gives expressive visual speech as output it can be considered a modular tool and we envisage it being used in the later stages of an animation pipeline. The pipeline would work as follows. An actor would be recorded saying a pre-designed corpus of training data. The corpus would have the same utterances spoken in extreme categorical styles. From these training data, an ICA model could be trained. This would then provide the controls needed to interpolate across the actor's recorded expressive gamut. In order to realistically capture a person's expressive styles, real conversational data could be captured. This would require the actor to be filmed having real conversations with people. The conversational material would be predesigned to elicit certain types of natural expression from the actor. Of course only certain types of expression could be captured in this way, for instance it would be unethical to genuinely elicit fear in the actor. Once this natural expressive information has been collected, it can be projected onto the previously trained ICA model. As was shown in Section 8.10. By segmenting the expressive modes which "fall out" of the model by phonemic boundaries defined by the accompanying audio, it should be possible to obtain phoneme level expressive features. These could then be clustered into expressive groups and used to train a Gaussian Mixture Model (GMM), or a Hidden Markov Model (HMM), modelling the mean and variance of each expressive cluster. At the same time, the audio could be analysed in order to establish a correlation between audio features and expressive ICA features. This could be done either by analysing things such as fundamental frequency, attack, decay, overall speed etc. or by attempting to perform some kind of factorisation

in the audio domain. In either case, the end result would be a mapping from audio space to the expressive clusters representing the natural expressive speech. An audio-to-visual speech synthesiser Theobald and Wilkinson [2008], or text-to-visual speech synthesiser Taylor et al. [2012] could be used as the first stage in the pipeline. An actor would talk expressively to the audio to visual speech synthesiser, which would produce neutral visual features based on dynamic visemes. At the same time another module would analyse the speech for expressive content and map the audio to one or more of the expressive clusters. The neutral speech features and the information on expressive style would then be passed to the expression modulation module, which would project the neutral speech into ICA space, and produce appropriate expressive features to be modulated with the neutral speech. An overview of a potential expressive speech synthesiser is shown in Figure 9.1. Such a system would provide a full automated pipeline to produce expressive facial animation with realistic lip synchronisation from an expressive audio input system. Each section of the pipeline could easily be implemented for example as a Maya plugin and simply added to an existing project, trained with the already collected data, and be ready to go in a production environment. We think that with further development, this work could be of much use to the animation industry in helping to produce high quality automated expressive animation with complex lip movements, saving time, money and more importantly freeing animators from the drudgery of current techniques, allowing them the freedom to make their artistic visions become reality.

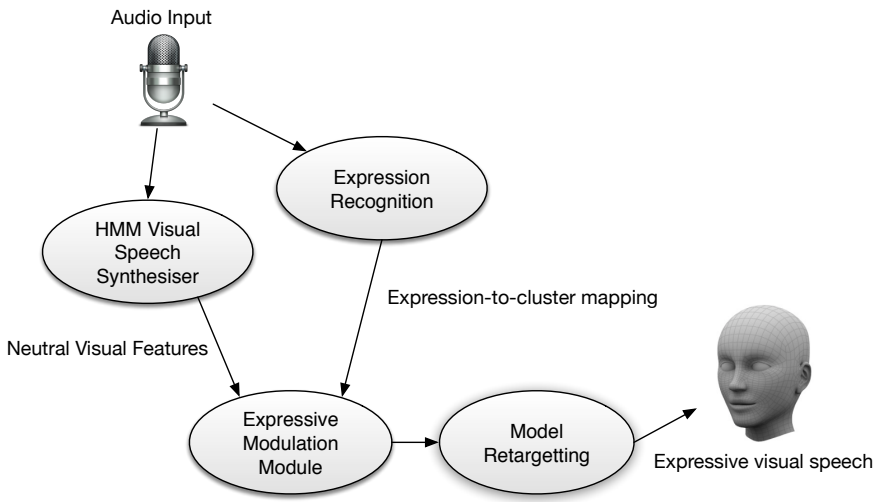


Figure 9.1 A proposed workflow for producing expressive visual speech from audio, using an HMM based speech synthesiser and ICA expressive modulation.

Appendices

Appendix A

Experimental Technique

These are the steps in a condensed form needed to repeat either of the modulation techniques described in Chapters 7 and 8.

1. Collect and Prepare Expressive Data.

   The data can be of different types and in this work, was in the form of 1080p video, 3D point cloud data and rig controller activation vectors. The data should display at least the extremes of expression to be modulated and neutral with preferably an expressive / neutral pairing of each sentence. The point cloud data and the video data was compressed using PCA models. All these models used the general form of:

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}), \tag{A.1}$$

   for projection into PCA space, and:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{A.2}$$

for projection out of PCA space.

(a) Obtain Features.

A parameterised representation of the features must be created.

i. Video.

The video in this work was tracked with an Active Appearance model which had been trained on 22 frames from the captured video representing the extremes of facial expression and neutral. Each frame was landmarked with 156 points in 2D space. The AAM was built leaving 95% of the variance in the shape and appearance. This yielded eigenmatrices which were of size [312x12] and [22834x6] respectively. Therefore each shape feature was of size [12x1] and each appearance feature was of size [6x1], giving a joint AAM feature vector size of [24x1]. A combined model of this joint feature was then made. Firstly the shape features were normalised by multiplying each by a scaling factor thus:

$$\mathbf{w} = \sqrt{\frac{tr_a}{tr_s}}, \tag{A.3}$$

where tr_a is the trace of the covariance matrix of the appearance features and tr_s is the trace of the covariance matrix of the shape features. Using this scaling factor a combined PCA model of shape and appearance was trained. Leaving 95% of the variance, the PCA process provided an eigenmatrix of size [24x5]. Therefore each combined feature vector was of size [5x1].

ii. Point Cloud.

The 3D point cloud data was simply projected onto a PCA model. The raw data was composed of 23370 vertices in three dimensions.

These were stacked giving a vector of size [1x70110] per frame. The model was trained on 39 frames covering the full expressive gamut as well as neutral and retained 99% of the variance in the training set. The returned eigenmatrix was of size [70110x18], therefore providing a projected PCA feature of size [18x1].

iii. Controller Activations.

The third data type used in this work were vectors of rig controller activations. These are simply the normalised (to the interval [0 1]) positions of each of the 30 controllers for each frame of training data. Therefore each element of the [1x30] feature represents the position of the same controller.

2. Dynamic Time Warping.

If using paired neutral and expressive sentences, then the expressive versions should all be warped to the neutral using Dynamic Time Warping (DTW). It is preferable to use the audio from the sequences if available, to create the warping path since this is at a higher resolution than the video (e.g. 44,100 Hz for audio as opposed to 25-60 Hz for video).

3. Independent Component Analysis.

Independent Component Analysis (ICA) is critical to the workflow allowing speech and expression to be factorised and independently manipulated. All cases in this work use the general form of:

$$\mathbf{s} = \mathbf{W}\mathbf{b}, \tag{A.4}$$

for projection into ICA space, and:

$$\mathbf{b} = \mathbf{A}\mathbf{s}, \tag{A.5}$$

for projection out of ICA space, where **s** are the recovered independent components, **b** are the mixes of speech and expression contained in the PCA or controller activation features, **W** is the orthogonal unmixing matrix and **A** is the mixing matrix which is the pseudo-inverse of **W**. To train the model, neutral and expressive time aligned PCA feature vectors were stacked and passed into the FastICA Matlab function. The settings for FastICA were as follows:

- Number of Independent Components (NumOfIC) - Equal to the number of PCA components.
- Non-linearity function used (g) - Gaussian
- Approach (whether to find the independent components in parallel or not) - deflate
- Maximum iterations before stop - 2000
- Epsilon (stopping criterion) - 0.001
- Stabilization (jolt out of local minima) - on

A random starting guess was used each time the models were estimated. This step provided the mixing and unmixing matrices **A** and **W**.

4. Modulation

   In the modulation step, a neutral / expressive pair was projected onto the same ICA model. The resulting ICA components were used as input either into the scaling technique described in Section 7.4, or the translation technique described in Section 8.2

5. Rendering

   After the ICA components have been altered with either of the techniques above, they must be projected out of ICA space using Equation A.5. If a PCA

model has also been used, these features must additionally be projected out of PCA space and into their original graphical representation i.e. landmarks and pixel intensities for video, or 3D co-ordinates for point cloud data, using Equation A.2. This provides the original neutral visual signal, modulated with expression.

Appendix B

Custom Dataset Sentences

The list of sentences used in the custom dataset. Each sentence was designed to make as much sense in Happy, Sad, Angry and Surprised expressive styles. This was done to try an minimise language bias i.e. we wanted to the user to be able to understand the expression portrayed only by the facial characteristics, and not through the linguistic meaning of the sentence. Each sentence and style was shown to the performer with a short back story which provides context for why each particular expression should be displayed.

1. "Why did you do that, what were you thinking?"

   (a) Happy

   Terry's friend placed a bet for him but accidentally placed the money on the wrong horse. Fortunately the horse came in anyway, and Terry won 1000.

   (b) Angry

   Terry's 14 yr old son has just been expelled from his second school for fighting. You call him to you and demand to know what he was doing.

(c) Sad

Terry's father has cancer. He didn't tell you before because he didn't want to worry the family and has been bottling up his anxiety inside.

(d) Surprised

After Terry have spent an hour looking for his wife's car keys, they discover that she threw them in the rubbish bin

2. "If I had a penny for every time I've heard that I'd be rich by now."

(a) Happy

Hannah's daughter has asked again for some money to buy clothes. She promises to pay the money back as soon as she gets a job. Hannah has a good job and has just received a bonus so doesn't mind.

(b) Angry

Hannah's son has yet again kicked in his bedroom door in a rage. A few hours later he has calmed down and comes downstairs to sheepishly apologise and promise never to do it again. Hannah however is not in a forgiving mood.

(c) Sad

Hannah's abusive husband has come (sober) after a two day drinking binge. He promises never to do it again. Hannah suddenly realises that she doesn't love him anymore and their marriage is over.

(d) Surprised

Hannah works as a team leader in a creative PR company. Surprisingly, given their usual mediocrity, one of her team has come up with a truly original idea. She hopes they?ll get the joke when as she says it.

3. "You know there's almost no difference between you and me."

(a) Happy

Janice notices that her 6 year old daughter loves plaiting her hair. Janice has always loved playing with her own hair.

(b) Angry

Filled with self-loathing, Janice thinks her anti-psychotic drugs aren't working. She looks into the mirror and yells.

(c) Sad

Janice's father dies five years ago from drink induced liver cancer. On the anniversary of his passing, Janice stands and pours some whiskey on his grave, and takes a good swig for herself.

(d) Surprised

Janice's father is dying of a drink related liver condition. After years of telling he father that he is a waste of space, Janice finally realises that she has been a hypocrite for many years.

4. "I need to go to Australia."

(a) Happy

Harry has had an academic paper accepted in a prestigious journal. He has been invited to a conference in Australia to present the work.

(b) Angry

Harry's feckless son has gone travelling and has been arrested and jailed for possession of drugs. Yet again, Harry will have to travel halfway round the world to bail out his son, and deal with legal proceedings.

(c) Sad

Harry's father (who emigrated to Australia after he retired) has just passed away. He must go to claim the body and deal with his father's estate.

(d) Surprised

Harry's business meeting which was supposed to be in Bulgaria, has been rearranged to be in Australia

5. "The carpet was covered in paint."

(a) Happy

The insurance office which last week refused to pay out for Paul's recent claim has been flood damaged overnight. On looking in through the window on his way to work, he saw the damage and recalls it to a friend over coffee.

(b) Angry

Paul came home to find that his teenage sons have been spray painting a model aeroplane. Even though they put newspaper down, the paint has seeped through onto the carpet.

(c) Sad

Paul has no money to buy his children Christmas presents. After an accident involving a can of plaint, he pleads to the insurance company who don't want to pay out.

(d) Surprised

Paul's wife tells him his brother (normally a tidy freak and a DIY expert) has been careless enough to paint the ceiling without putting a dust sheet down. Paul's reaction is one of surprise.

6. Hasn't this just been the most memorable day we've had in years?

(a) Happy

Pauline and her husband have little money. Their children have clubbed together to send them to Paris for their golden wedding. On the first

day, after a visit to the Louvre and a boat trip along with Seine followed by a candlelit meal at a good restaurant, Pauline gives her husband a kiss and comments on the day.

(b) Angry

Robert and his wife have booked a hotel in Rome for a short holiday. On the way to Rome, he mislays his boarding pass, loses the details of their hotel and drops a heavy suitcase on his wife's foot. After they have unpacked, they go out for meal which turns out to be very expensive and almost inedible. They return to the hotel to find that Robert has lost the pass key. His wife had had enough and sarcastically lets off steam.

(c) Sad

Melanie and her husband have been out dinner together for the first time in years. They have a severely disabled child who takes all their time. Although they love their child, they are keenly aware of the effect on their relationship of constant care. On the way home, Melanie comments to her husband.

(d) Surprised

Pat has worked at a boring job for three years. At the end of a supremely boring training day, the managers announce that staff are to be given a big bonus. Pat turns to her colleague.

7. She turned and looked straight at me; I couldn't believe it.

(a) Happy

Shy young Henry who has been fancying a pretty colleague but believed she had never noticed him, is delighted when she shows an interest.

(b) Angry

Amy sees her ex-boyfriend out with the girl for whom he left her. The new girlfriend gives Amy a triumphant stare. Amy later tells a friend about the encounter.

(c) Sad

Jake is sure that someone is spreading an unpleasant rumour about him. He believes it to be a colleague he has always tried to help and cannot understand why she should be trying to cause him trouble. They are in the canteen at lunchtime and he notices her, expecting her to glance away embarrassed. He later returns to his desk and tells another colleague about his suspicions.

(d) Surprised

Marion, now in middle age, has not seen her step-daughter in years. On a day out Christmas shopping in London, Marion comes across her working on the shop floor of Liberty's and later tells her sister.

8. The doctor told me he can't find anything wrong.

(a) Happy

Ed has been suffering from stomach pains, and has worried he has cancer. He goes to the surgery to get the results of the tests he has had and is reassured that none of them has revealed anything wrong. He hurries home to share the news with his wife.

(b) Angry

Gillian, who constantly complains about being ill, has gone to her GP yet again. She is told that there is nothing wrong and is made to feel she is being a time waster. She goes home feeling aggrieved and when her husband asked her what happened at the surgery she reports the gist of the conversation with considerable annoyance.

(c) Sad

Hannah is worried that her one year old child may be autistic. Her friends have suggested that perhaps he might have a hearing problem. She has taken him to have his hearing tested but the paediatrician can find no obvious physical cause for the child's inability to relate to people. One of the friends drops round to find out what the GP said.

(d) Surprised

Lily is about to move into sheltered accommodation. She has recently been having chest pains and is convinced she has a heart condition. She goes to her GP who listens to her heart and reassures her that it is fine, and that the chest pains are most likely to be due to anxiety about the coming change. She tells her daughter there is nothing to worry about.

9. Did you hear that the police have shot the dog?

(a) Happy

People have been confined to their homes for the past six hours while police are hunting for a rabid dog who has already bitten a child. The local news channel flashes a report that the police have been successful and it is now safe to go out. Beth phones her husband at work to tell him.

(b) Angry

The Lees are travellers. There are several dogs among the community that look as though they might belong to proscribed breeds. They are reported, and the police arrive. The Lees' dog gets loose and runs towards the police. It is shot. Patrick Lee says the dog was just being friendly and was harmless. An hour later his son gets back from his job at the nearby fruit farm and Patrick greets him with the news.

(c) Sad

Fred has a fatal heart attack as he is out walking his dog. The dog will let no-one near the body and bites a paramedic who gets too close. Eventually the police call in a marksman who shoots the dog. Later the man's daughter is informed of her father's death. She phones her brother and as well as telling him about their father's death also tells him about the dog.

(d) Surprised

As above. A local vet speaks the next day to his colleague. He begins with the given statement, and suggests that the dog could and should have been sedated.

10. It was at that moment I realised what must have happened.

(a) Happy

Tanya loses her engagement ring. She goes over in her mind what she had been doing that day and retraces her steps but she cannot find it. The next day a flash of sunlight glints on something down the side of her car seat. She slips her hand in and pulls out her ring. Then she remembers pulling off her gloves in a hurry to put her credit card into the toll road pay machine.

(b) Angry

Russell is contacted by the council about a loft extension he is supposed to have had done. Although he has done a little work in his loft, it does not amount to anything needing planning permission. He reads the letter from the council again, and notices a phrase that only his estranged brother uses. At this point he realises that his brother has once again tried to cause trouble.

(c) Sad

There is a family get-together and Laura and Diana are enjoying a drink while they look at their parents' wedding album. They notice a shadowy figure on a photo they have never seen before and ask their parents who it is. Their mother suddenly leaves the room. Their father looks at the photo and breathes a man's name before following their mother. One of the sisters remembers a whispered story about their mother having once been engaged to a man who turned out to be married. Together they speculate that this man had learned of their parents' wedding and had secretly gone to the church. Later the sisters discuss the occurrence with their brother and Diana explains how she worked out who the figure in the photo must have been.

(d) Surprised

A teacher asks Jamie why he has not done his homework. He says he handed it in the day before. She goes through the pile of essays for the class, but his essay is nowhere to be found. She is about to demand that he do the essay for the next day, when she remembers a fire drill had taken place in the previous lesson. She realises that the essay was in her hand when she left the classroom and that it must be inside the register. At lunchtime she laughingly relates the incident to her colleagues.

11. Is this really you in the photograph?

(a) Happy

Joan shows her carer a photograph of herself when she was twenty years old. She is very pretty, and the carer is delighted to be able to talk about happier times.

(b) Angry

Kevin is sent a grainy but compromising photo purportedly of his wife. He immediately confronts her with it, and while she examines it, storms out of the house without waiting for an explanation or denial.

(c) Sad

Ellis has spent half an hour arguing and telling his parents what he thinks of them and their parenting skills. He leaves for school and as his mother closes the front door behind him, her gaze fixes on the framed photograph of a sweet smiling little boy. She wonders what ever happened to her beautiful son.

(d) Surprised

A family looks at old photos. Lilian, a rather strict grandmother, sees a photo of herself when she was young; she is wearing a very short skirt and low cut top and enormous hoop earrings. She tries to hide it but her grandson swoops on it ...

12. The board turned down their offer.

(a) Happy

A takeover bid that would have led to redundancies has been rejected. Sam, an employee to whom news of the takeover had been leaked, informs his colleagues that their jobs are safe.

(b) Angry

Nick, one of the owners of a manufacturing firm, is hoping to sell it off and retire but is outvoted in a rancorous meeting by the other members of the board of directors. He drives home after the meeting and his wife appears as he opens the front door.

(c) Sad

As for 'Angry'. Nick's wife, Esther, tells her friend that her husband

won't be retiring yet after all and that dreams of holidays in hot places have been put on hold.

(d) Surprised

Wilsons, an old fashioned DIY firm, is approached by a larger firm for a takeover deal. Justin, the regional manager for the larger firm, has automatically assumed that their offer will be accepted. However, he has not reckoned with old fashioned values of loyalty to employees, and at the end of the meeting phones his boss with the news.

13. You never told me you worked here.

(a) Happy

Phil starts a new job and bumps into the husband of one of his wife's friends.

(b) Angry

Becky, a young teacher, starts a new post at a comprehensive school and sees an ex-boyfriend working there. She would not have applied for the post had she known he had also taken a job at this school because the relationship had ended very badly. She bumps into him in the staffroom on her first day.

(c) Sad

Trevor, a Health and Safety inspector has been called in to inspect a restaurant which seems to be the focus of an outbreak of food poisoning. He finds that the manager is the son of his golfing partner.

(d) Surprised

Roy has told his wife he is working late, but goes to a strip club with some colleagues. A scantily clad waitress goes to their table to take a drinks order and he sees it is the daughter of his next door neighbour

who is supposed to be working evening shifts as a care assistant in an old people's home.

14. She's pregnant again.

    (a) Happy

    Chris tells his mother that his dog is expecting a second litter of pups.

    (b) Angry

    Debs has recently learned that her IVF treatment has failed for the third time. She bumps into an old schoolfriend who moans that she is expecting her third child in four years. Debs feels that life is dreadfully unfair and rushes round goes to see her mother and tell her of the encounter.

    (c) Sad

    Verity has just lost her husband in a car crash and discovers she is six weeks pregnant. She phones her mother who relays the news to her husband when he returns from work.

    (d) Surprised

    Jess and Nathan have had two children and had not planned to extend their family further. Jess is taking a contraceptive pill, but she had a stomach bug and she learns that she is pregnant again. Nathan calls round to see his parents to give them the news.

15. Of all people, you are the last one I'd expected to see here today.

    (a) Happy

    At a school reunion forty years on, Alec re-encounters Hilary, his first girl friend. He had joined the army and she had gone to university and

then to live abroad. He has been widowed for three years. Alec has immediately recognised Hilary and goes over to speak to her.

(b) Angry

Susie experienced an acrimonious divorce when her husband told her he wanted to marry another woman. She suffers from depression for several months but eventually plucks up the courage to go to a singles encounter group session and there she meets her ex-husband.

(c) Sad

After years of non-contact, Carl, the birth grandfather of an adopted child, notices a young boy skateboarding who looks so like his son looked when he was young, that he approaches the boy and asks whether his name is Danny. The boy confirms his name and asks whether he was his grandfather and Carl continues the conversation.

(d) Surprised

Sally friends meets an old school friend at a political rally. Although she had always held radical beliefs, her friend had shown no interest in social protest, and had always poured scorn on the usefulness of such action. Sally greets her.

Appendix C

Simple Modulation

Here are some example sequences using the technique described in Section 7.

C.1 Neutral

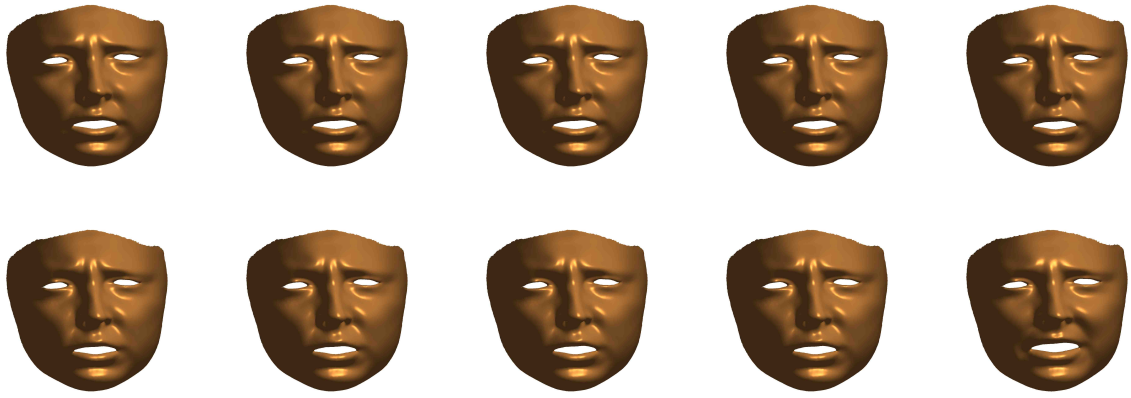
This is a sequence of neutral ground truth data.

C.2 Modulation with sad

This is the sequence from Section C.1 after having been modulated with sadness.

C.3 Sad Ground Truth

Here is the expressive ground truth for the original sad sentence which corresponds to the sequence shown in Section C.1.

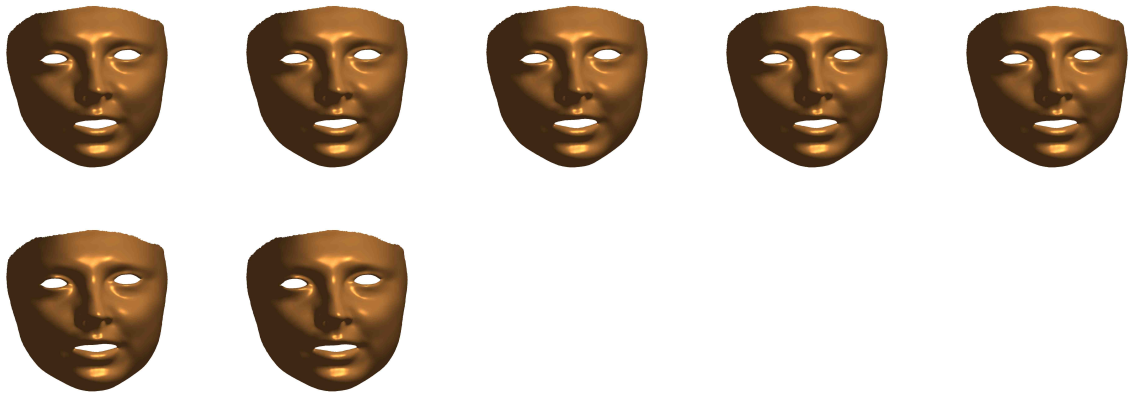


C.4 Neutral

Here is another neutral sequence of ground truth data.

C.5 Modulation with happy

Here is the sequence shown in Section C.4 after modulation with happiness.



C.6 Happy Ground Truth

Here is the original ground truth happy sequence which corresponds to the neutral sequence shown in Section C.4.

Appendix D

Mixed Modulation with AAMs

D.1 AAM based Neutral

This is a sequence of ground truth neutral images.

D.2 AAM based - modulation with happy

Here is the sequence shown in Section D.1 after DTW and modulated with happiness.

D.3 AAM based - modulation with anger

Here is the sequence shown in Section D.1 after DTW and modulated with anger.

D.4 AAM based - modulation with surprise

Here is the sequence shown in Section D.1 after DTW and modulated with surprise.

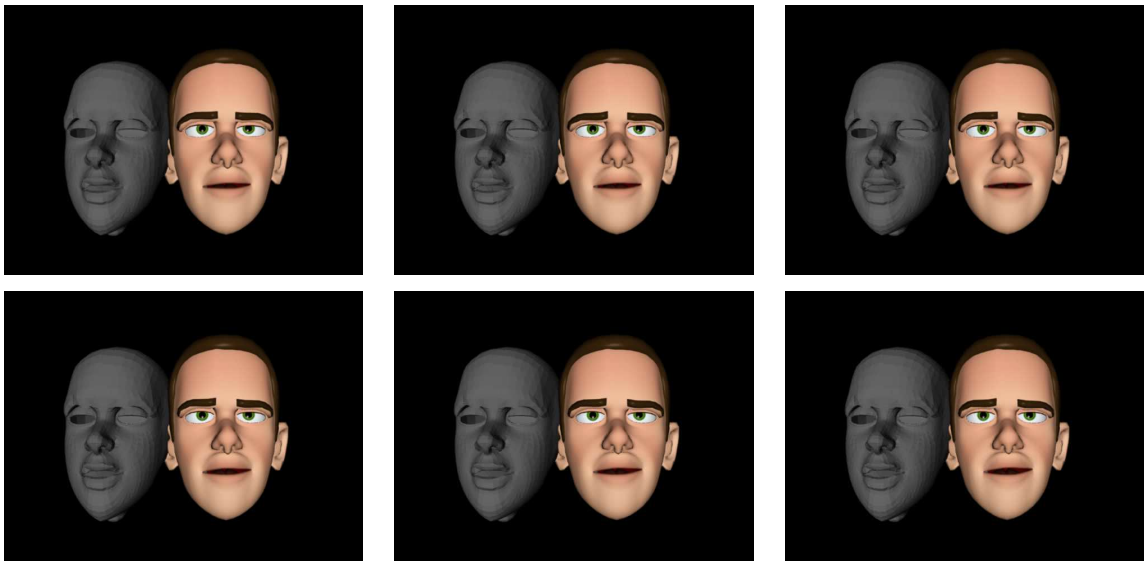
D.5 AAM based - modulation with sadness

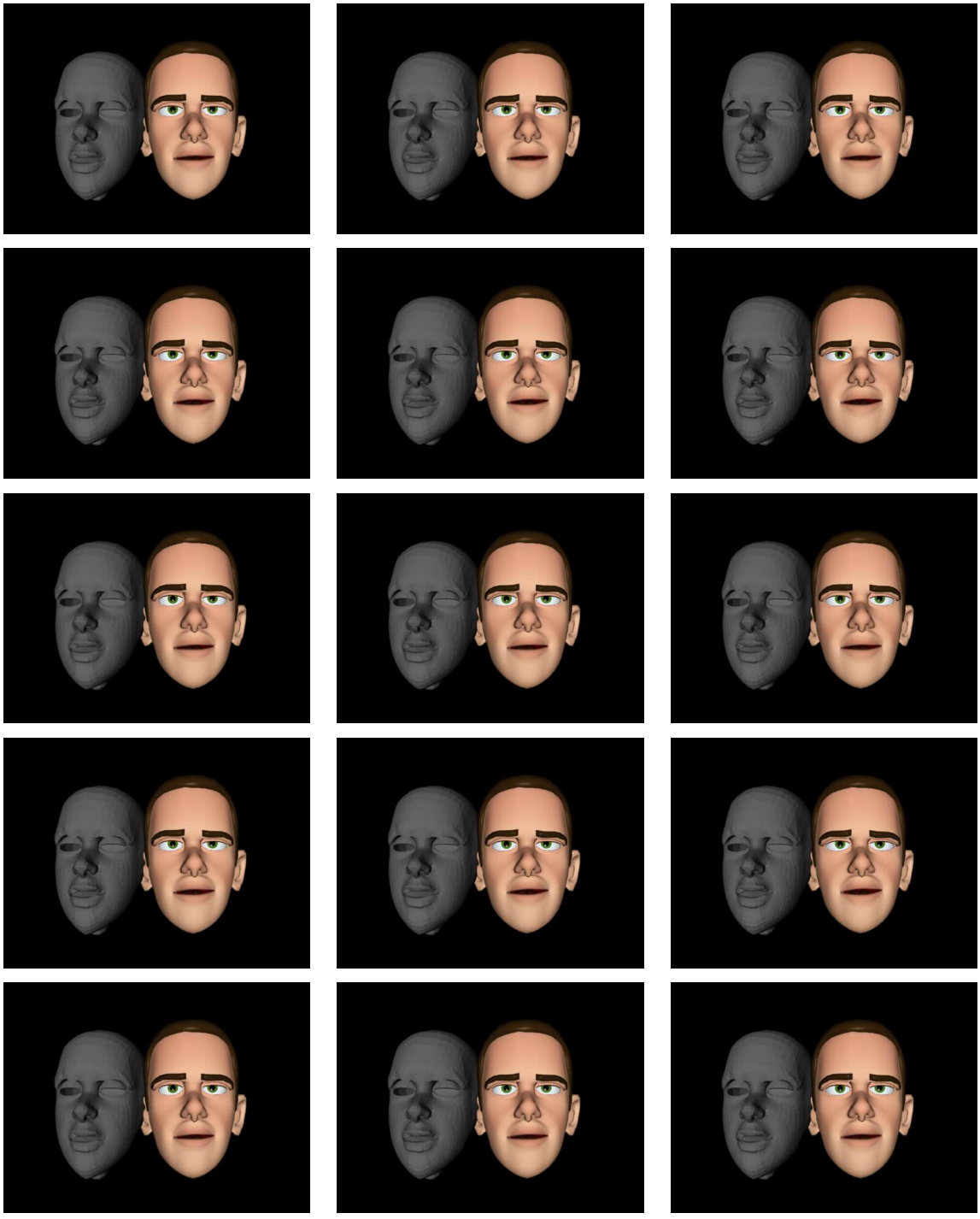
Here is the sequence shown in Section D.1 after DTW and modulated with sadness.

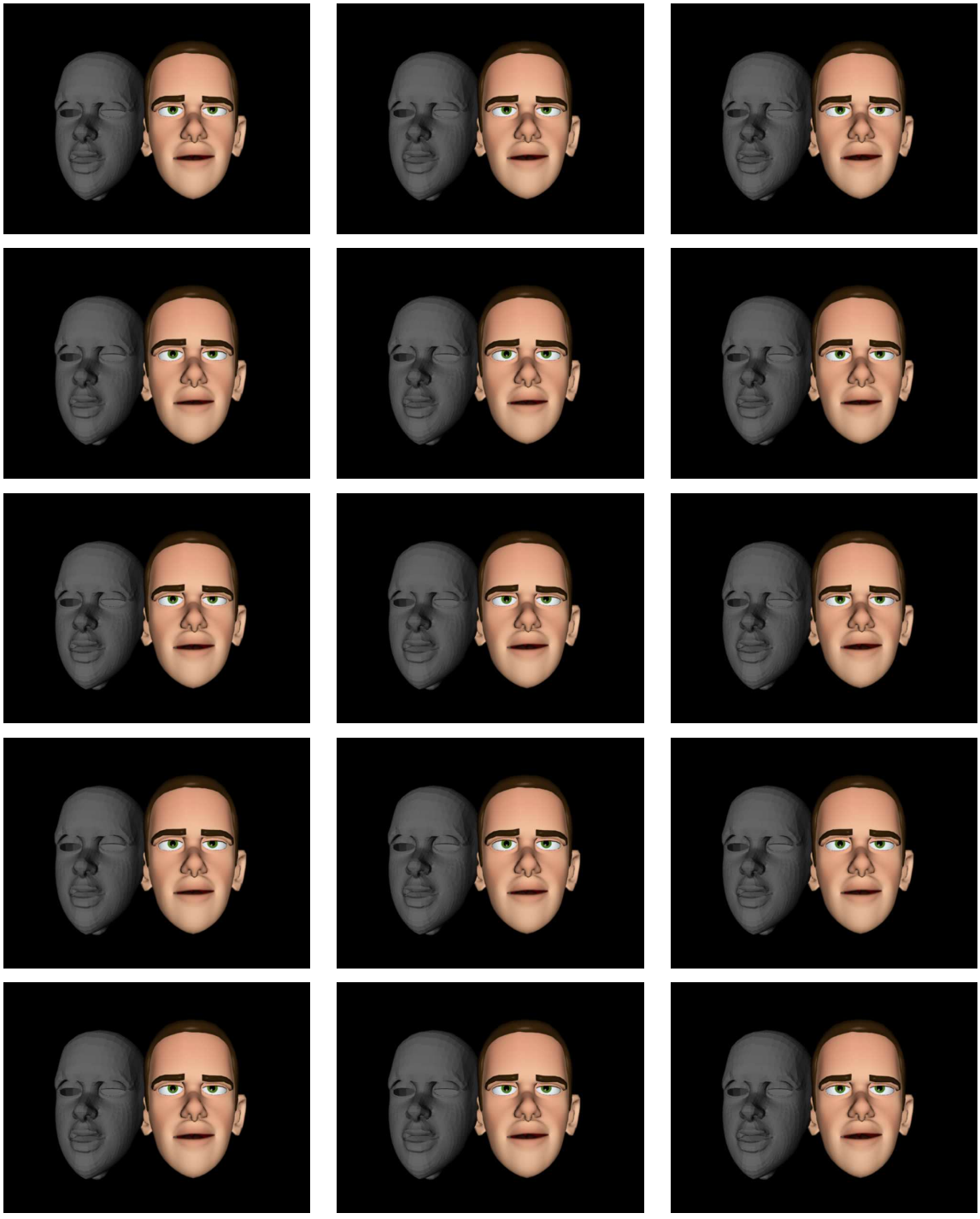
Appendix E

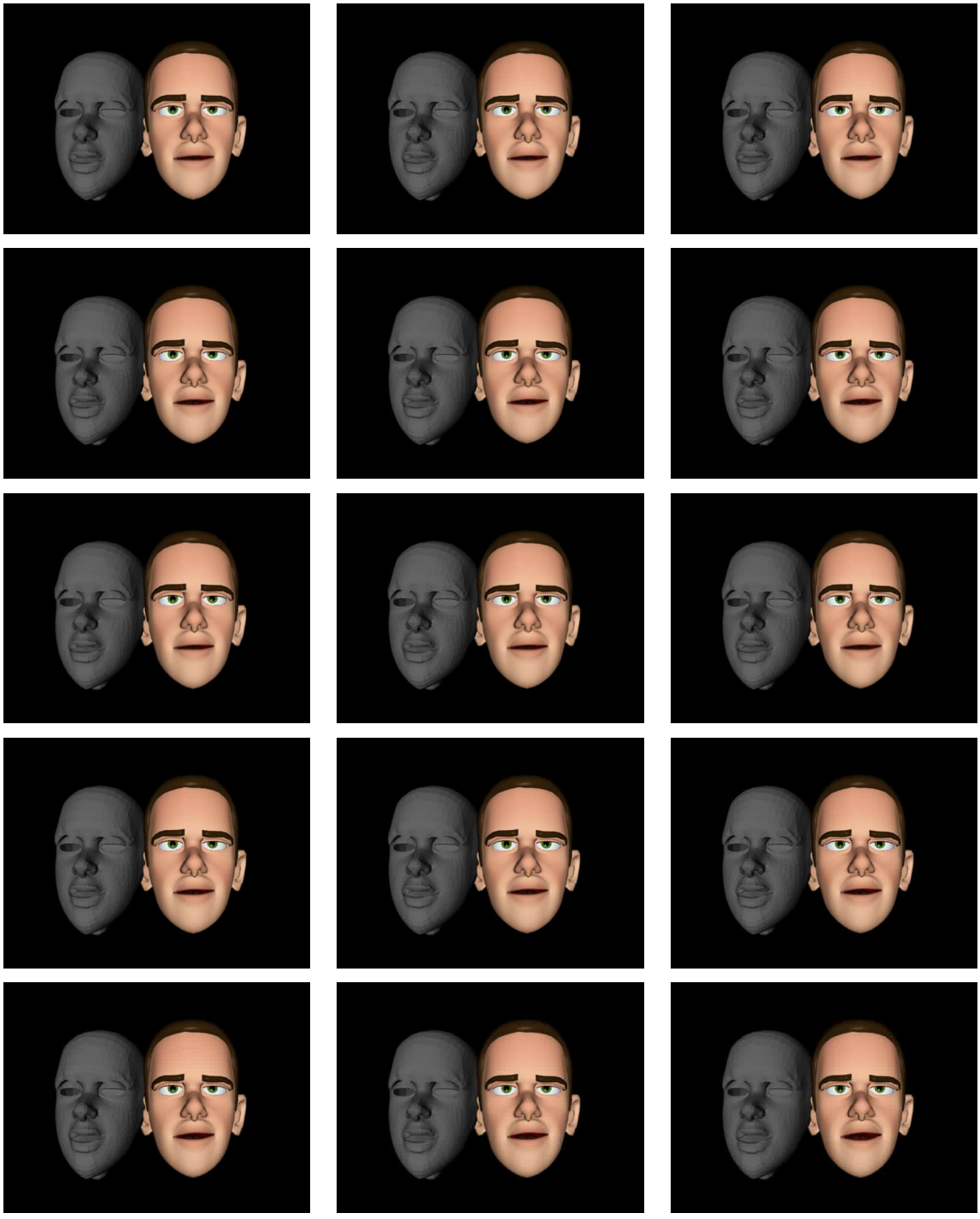
Nelder-Mead Fitting

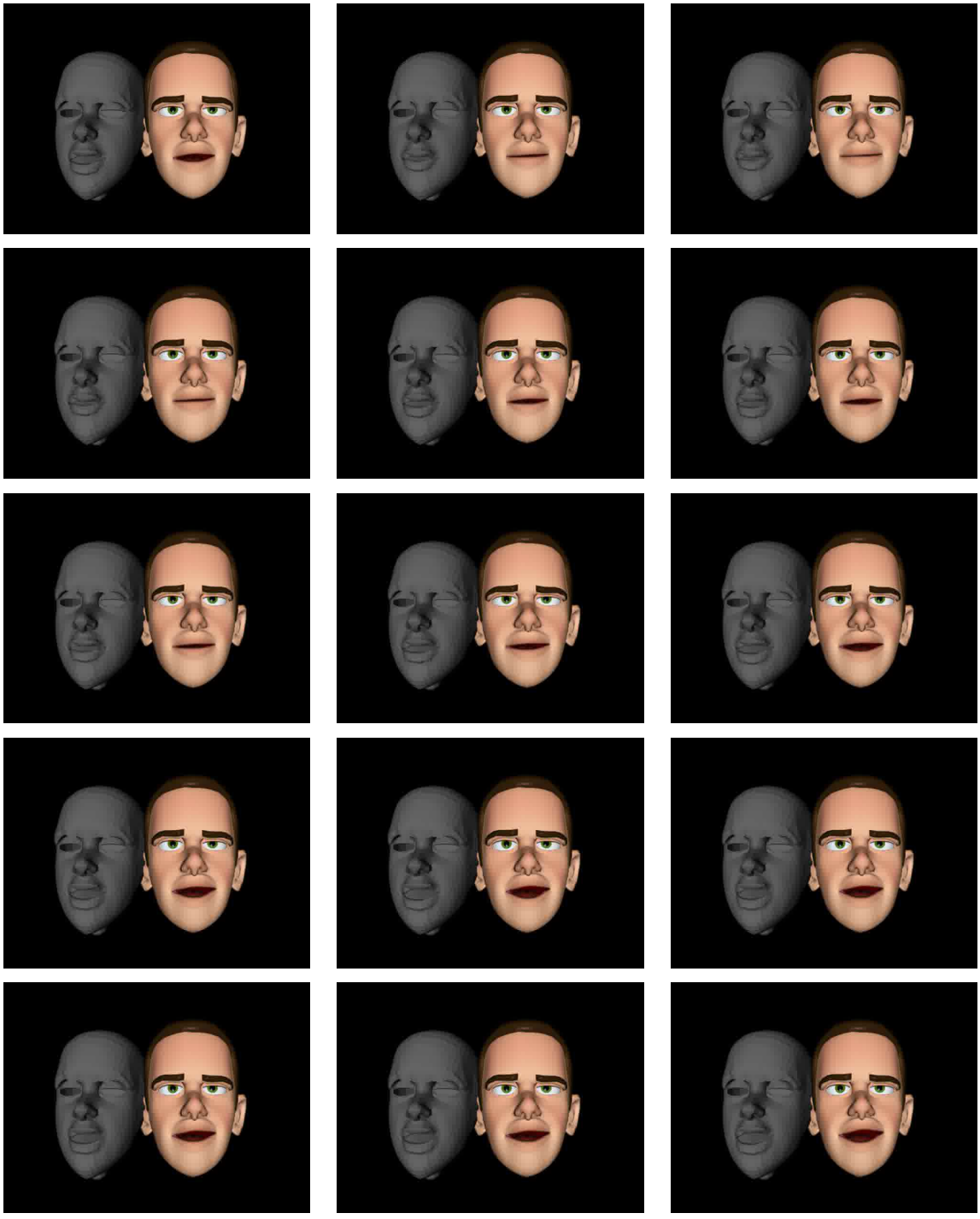
These frames show the results of fitting the rig controlled mesh to the AAM warped mesh for an entire sequence.

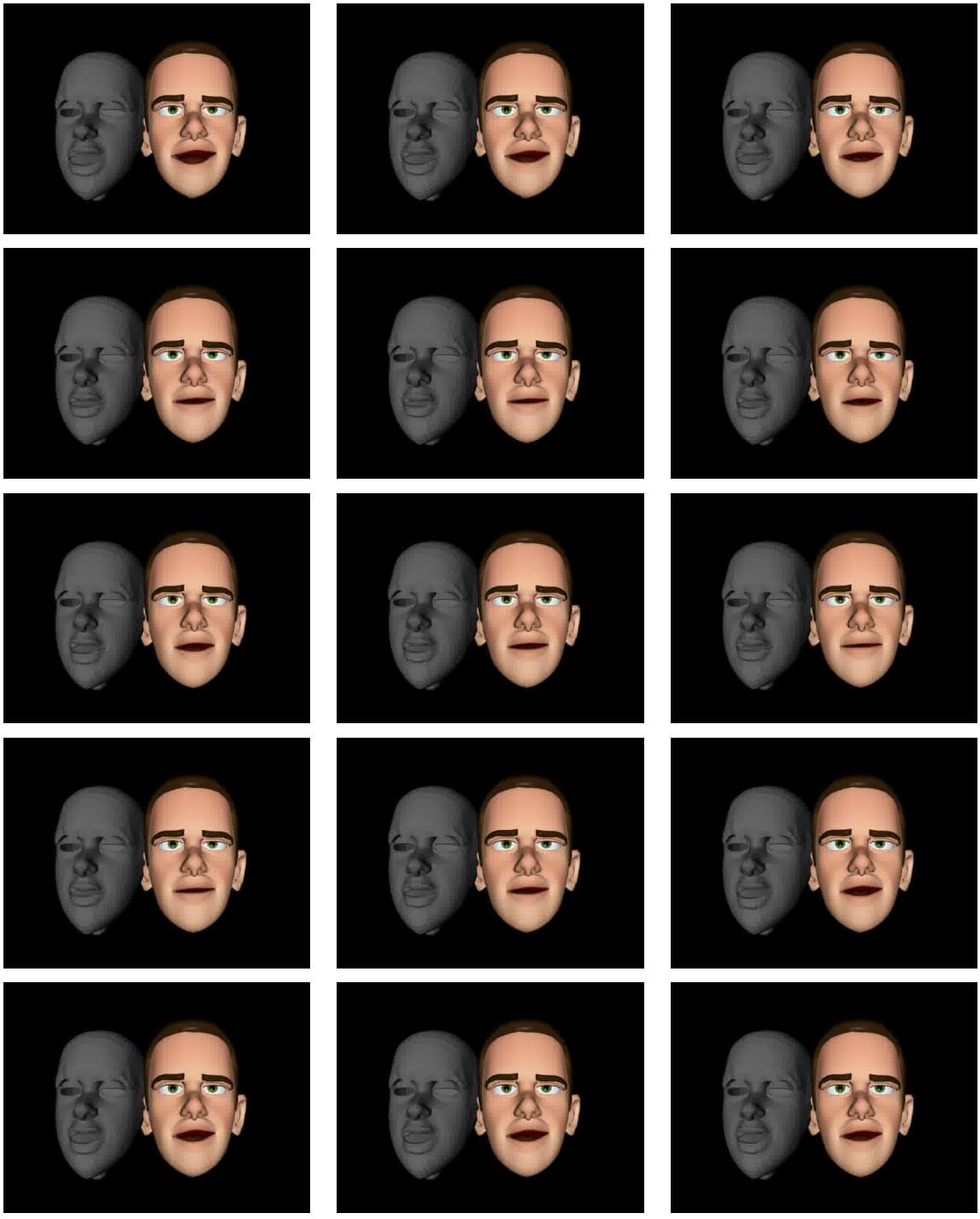


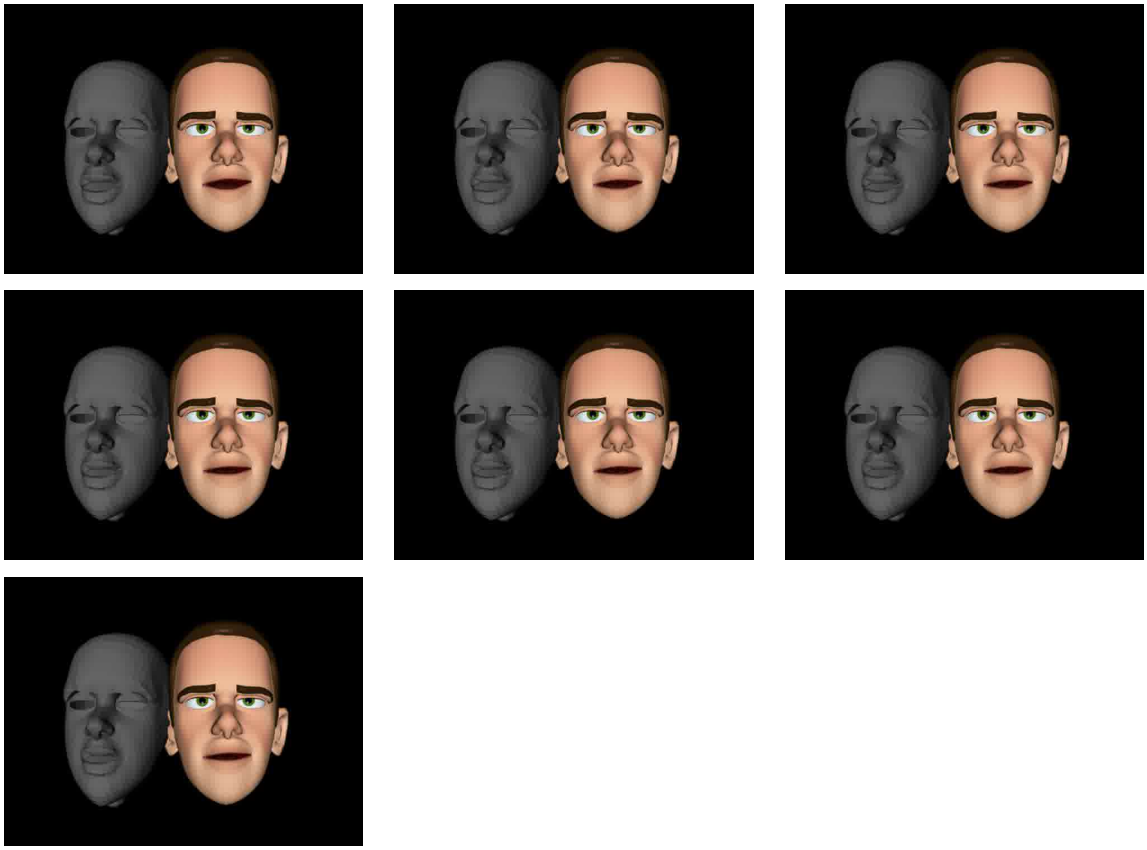










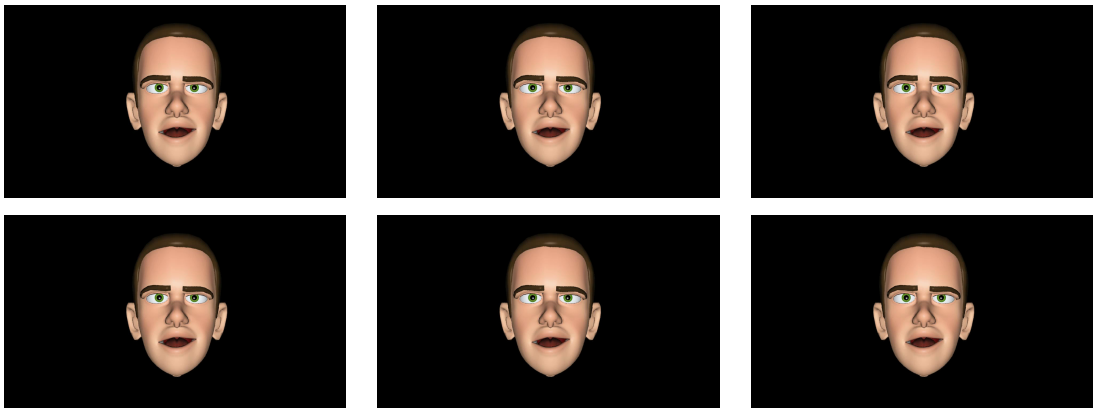


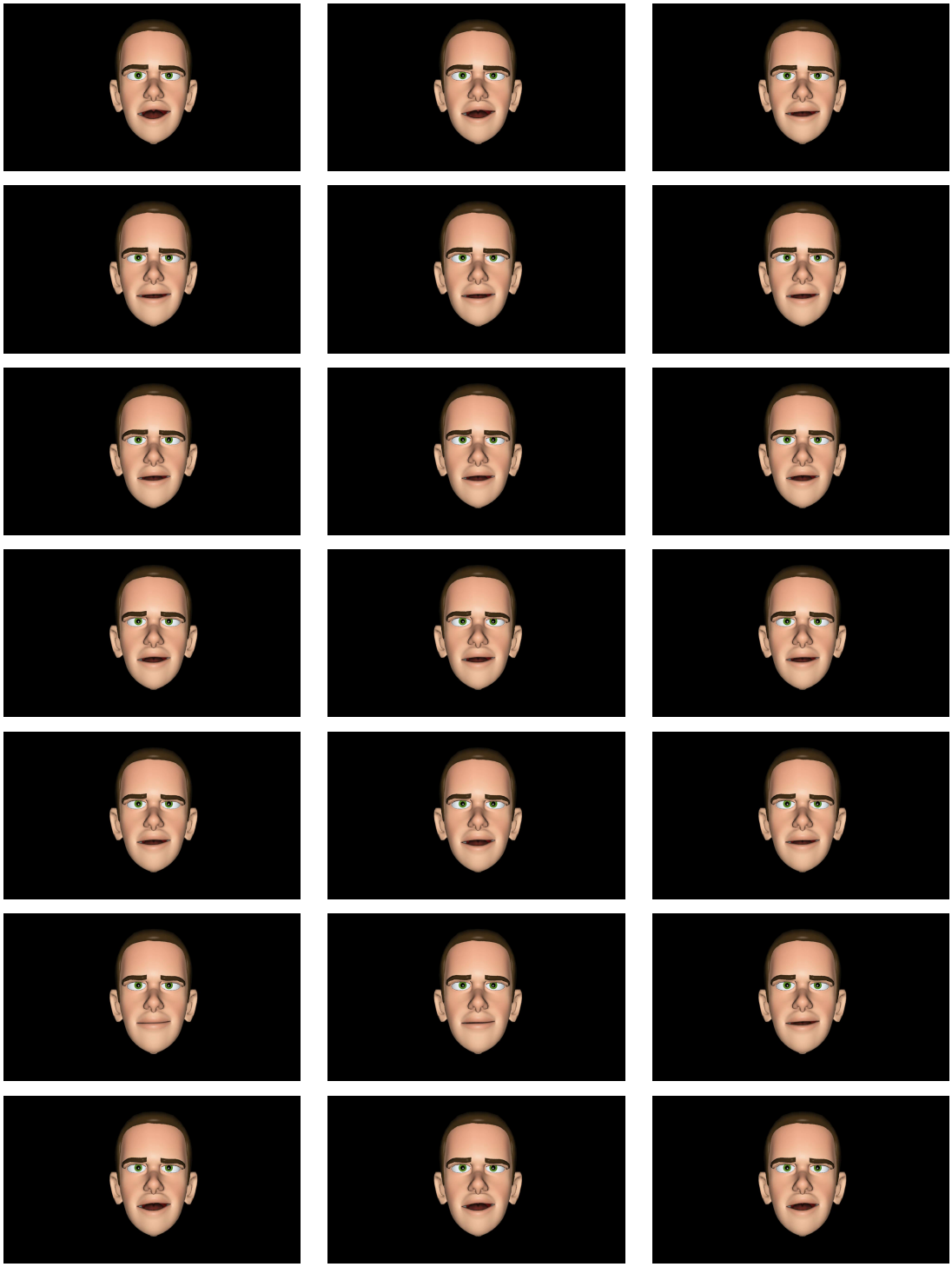
Appendix F

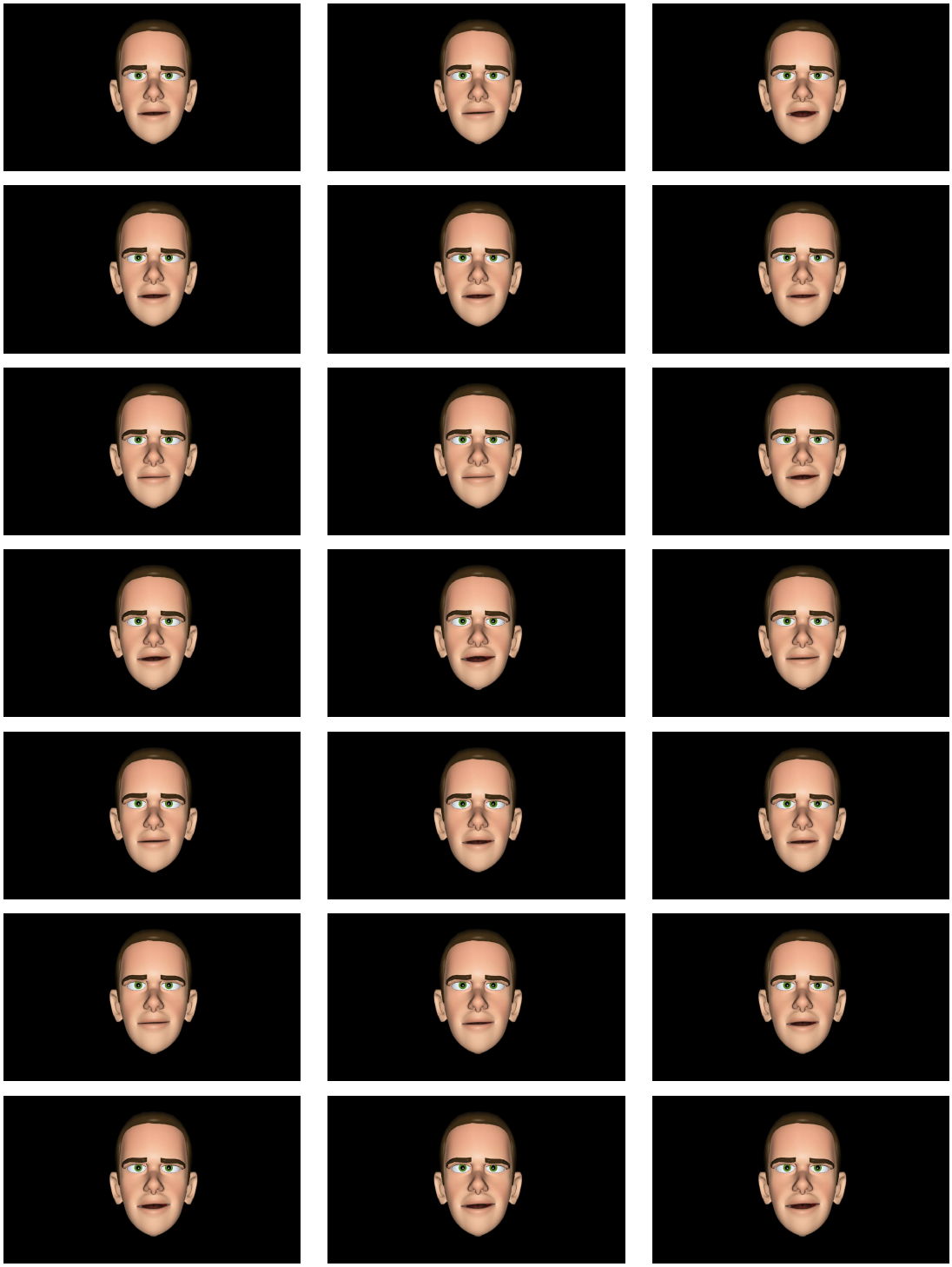
Mixed Model Modulation - Rig Based

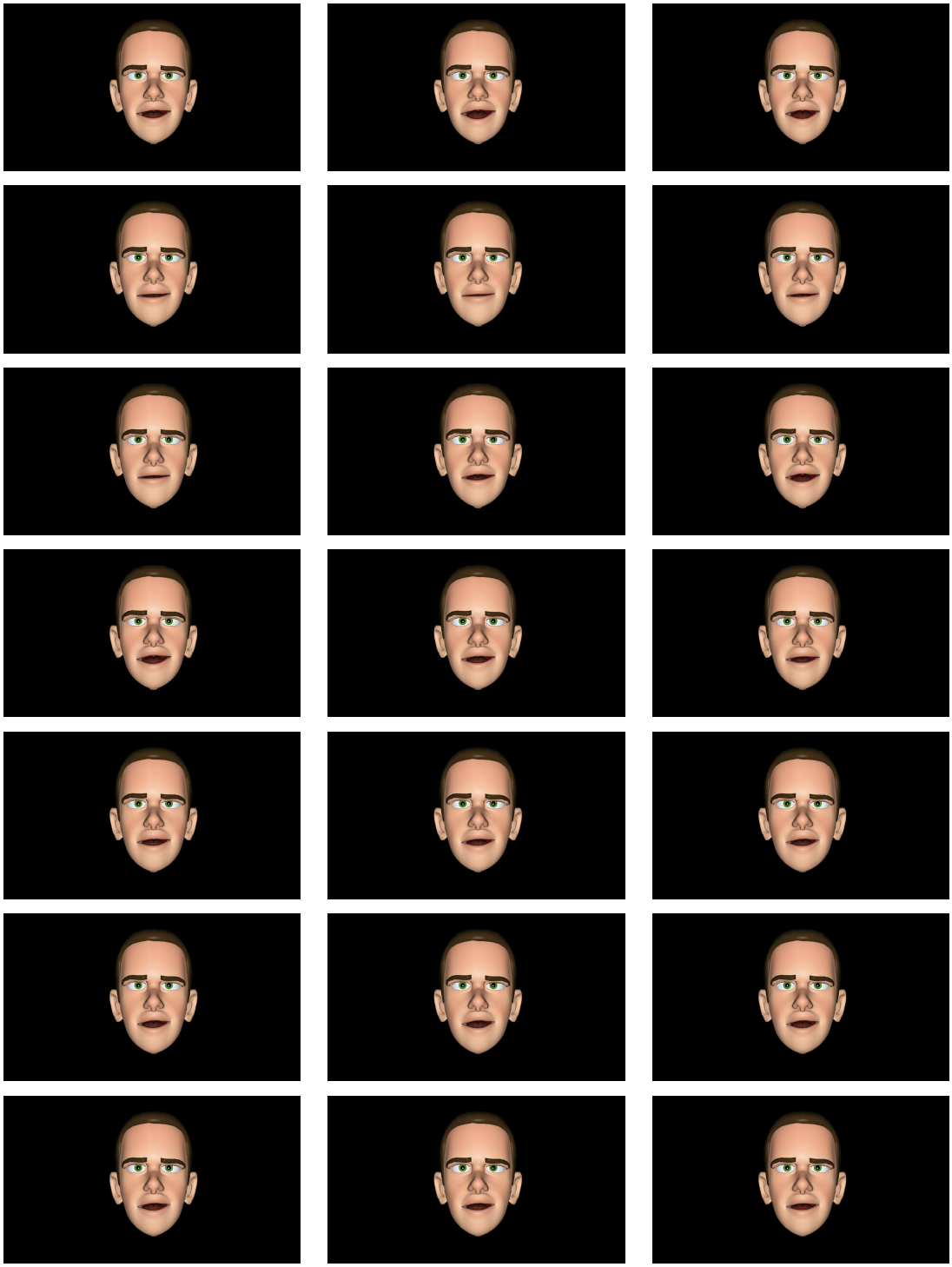
F.1 Rig based - Ground Truth Neutral

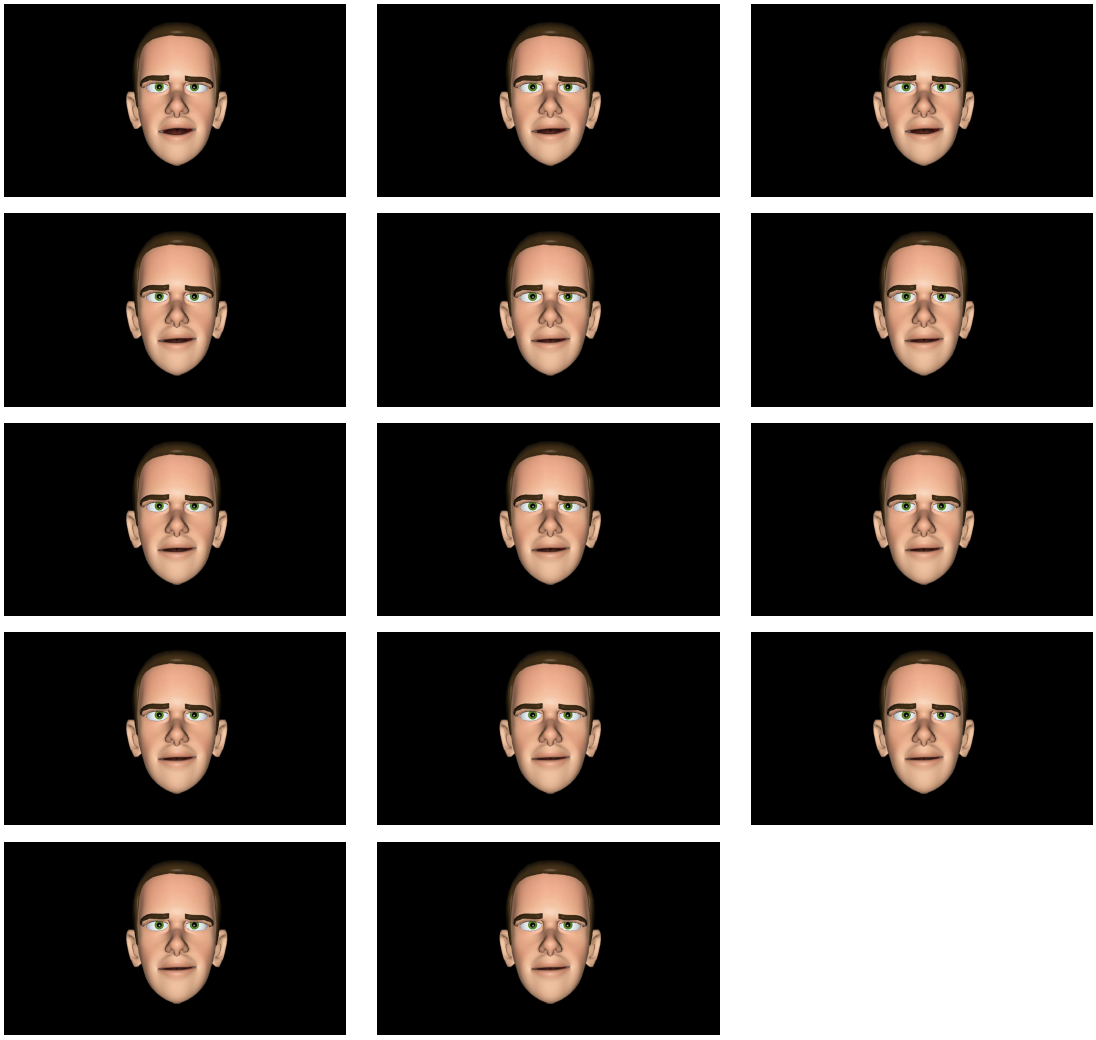
Here is a sequence of ground truth data transformed from AAM space and projected onto the Morpheus rig.





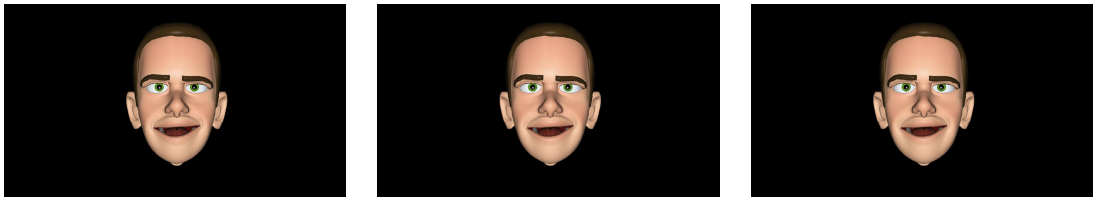


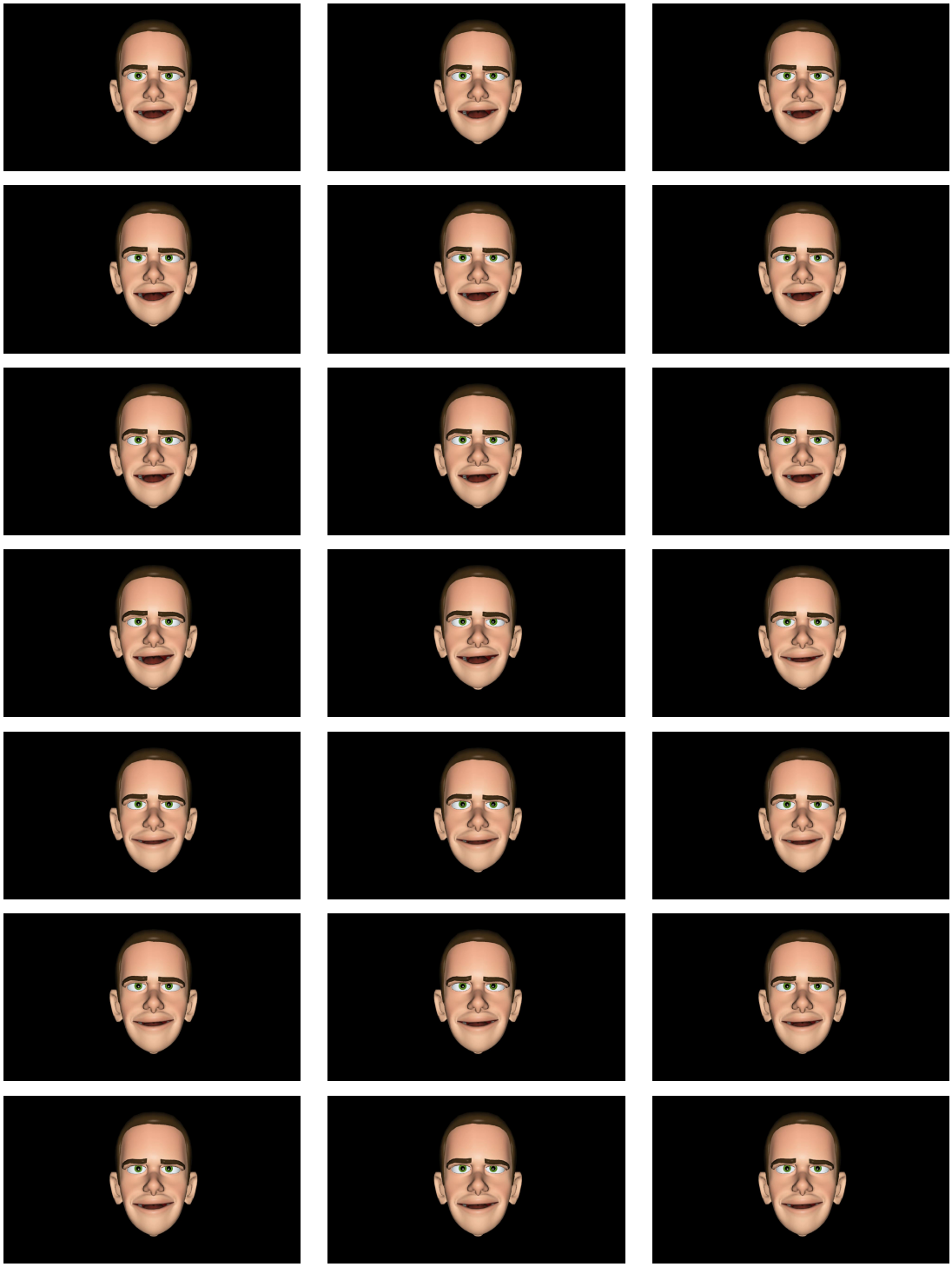


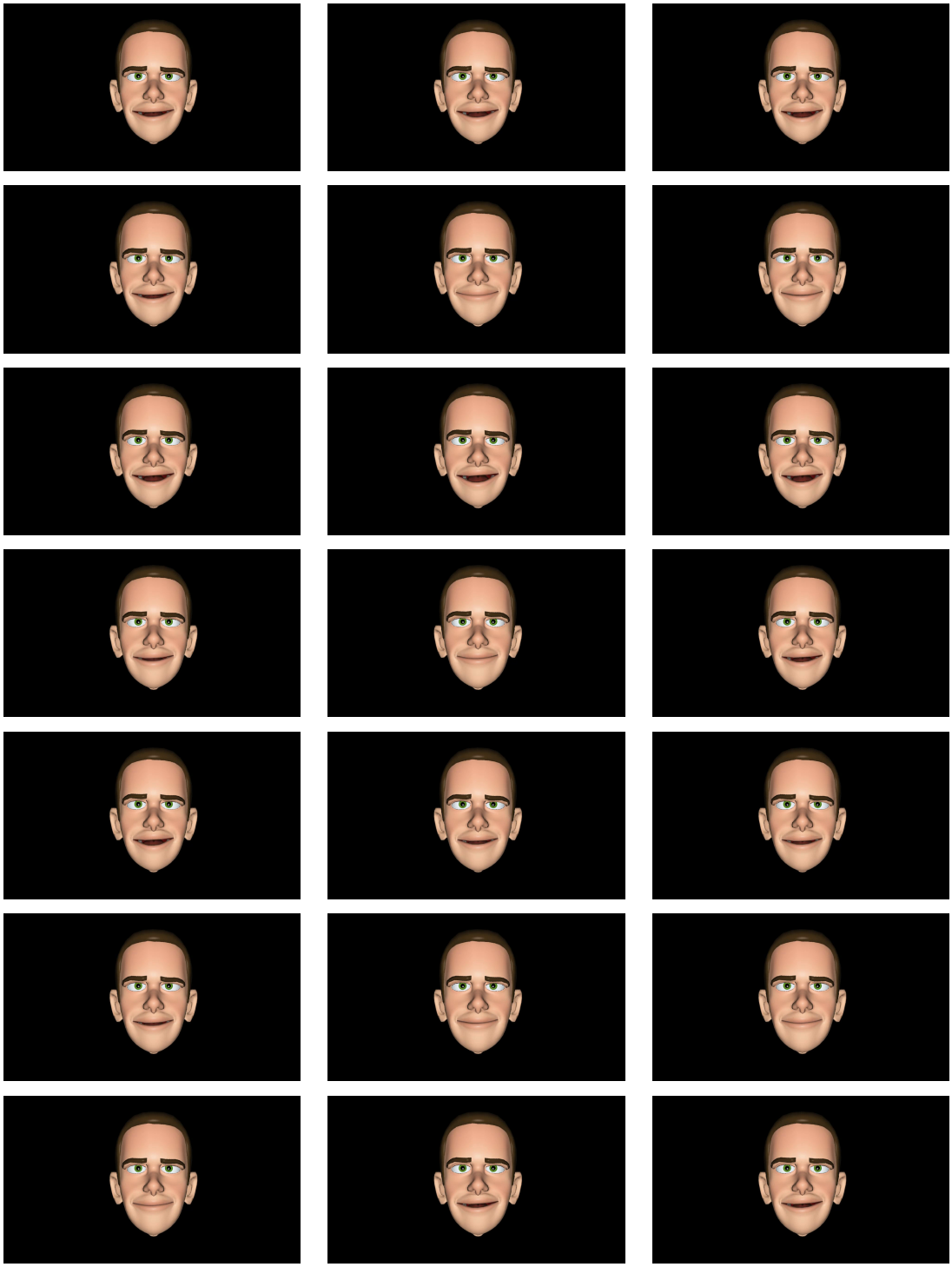


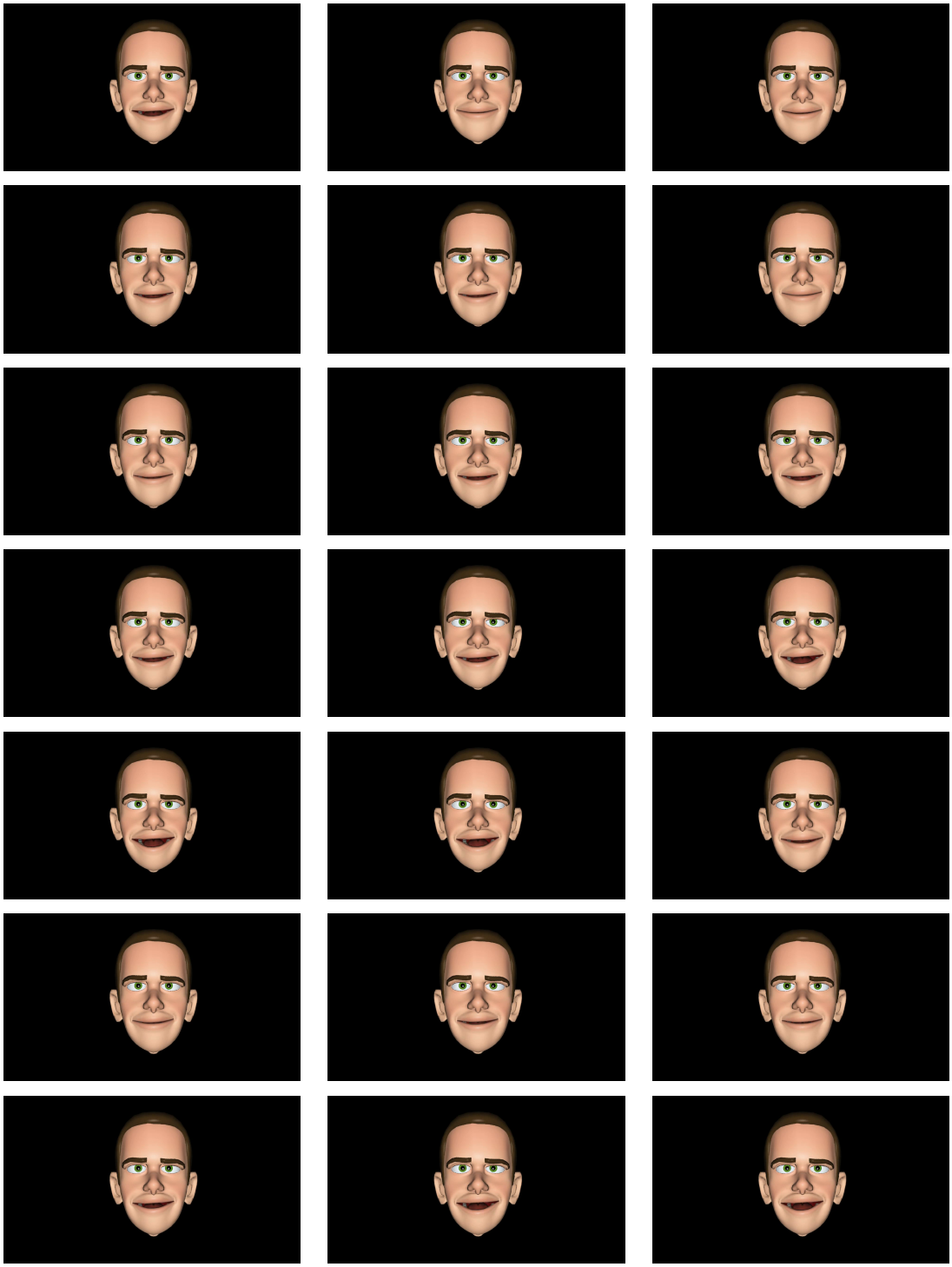
F.2 Rig Based - Modulation with Happy

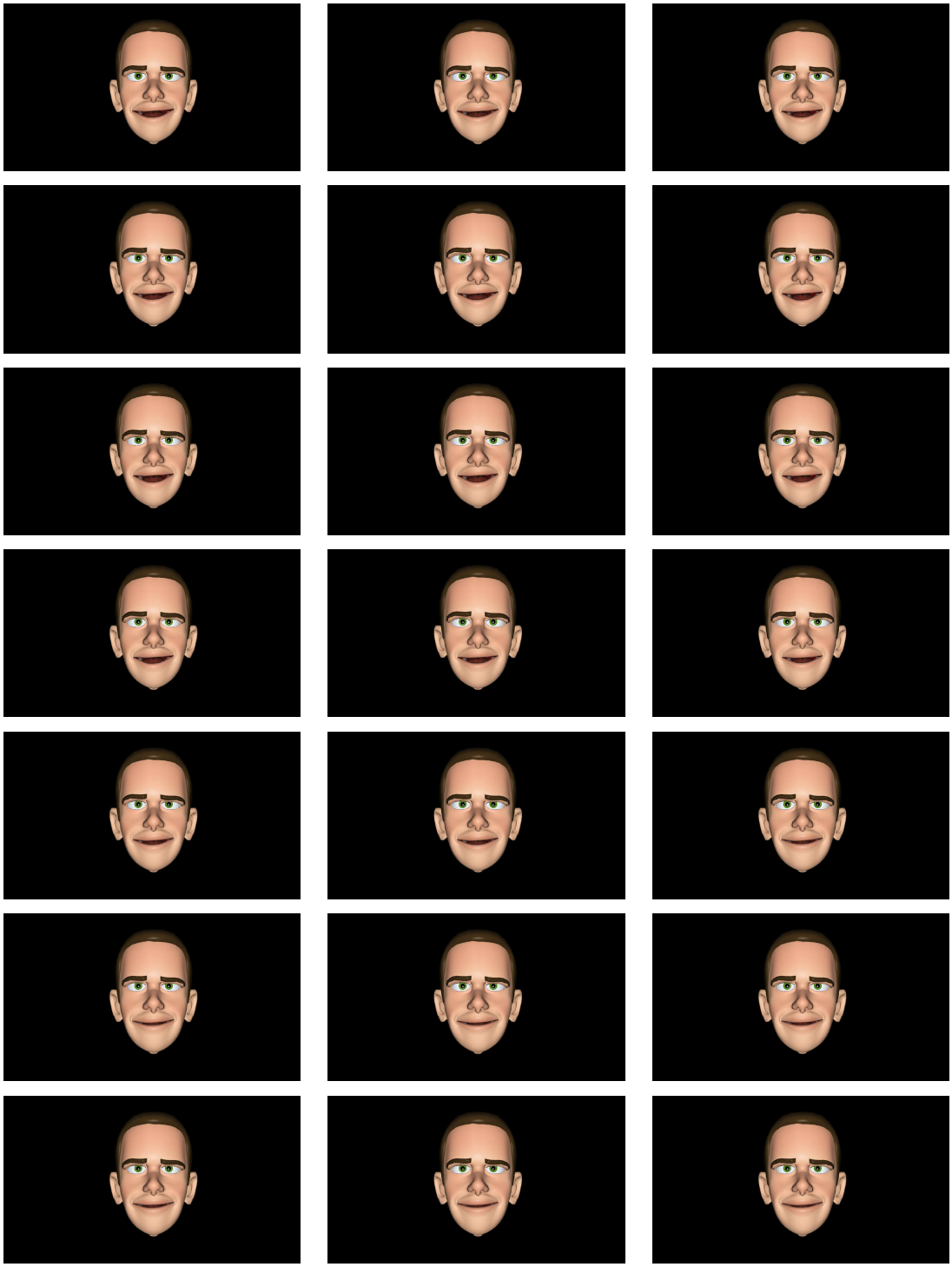
Here is the sequence shown in Section F.1, after DTW and modulation with happiness.

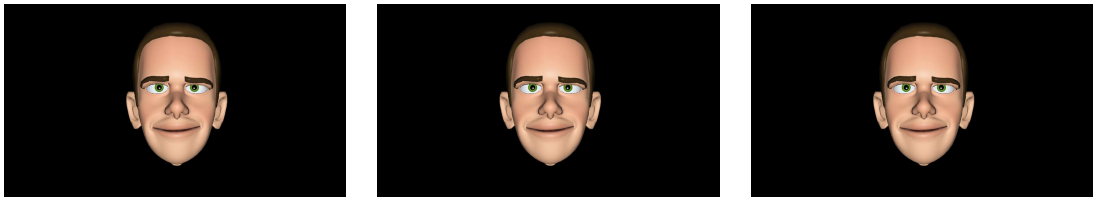






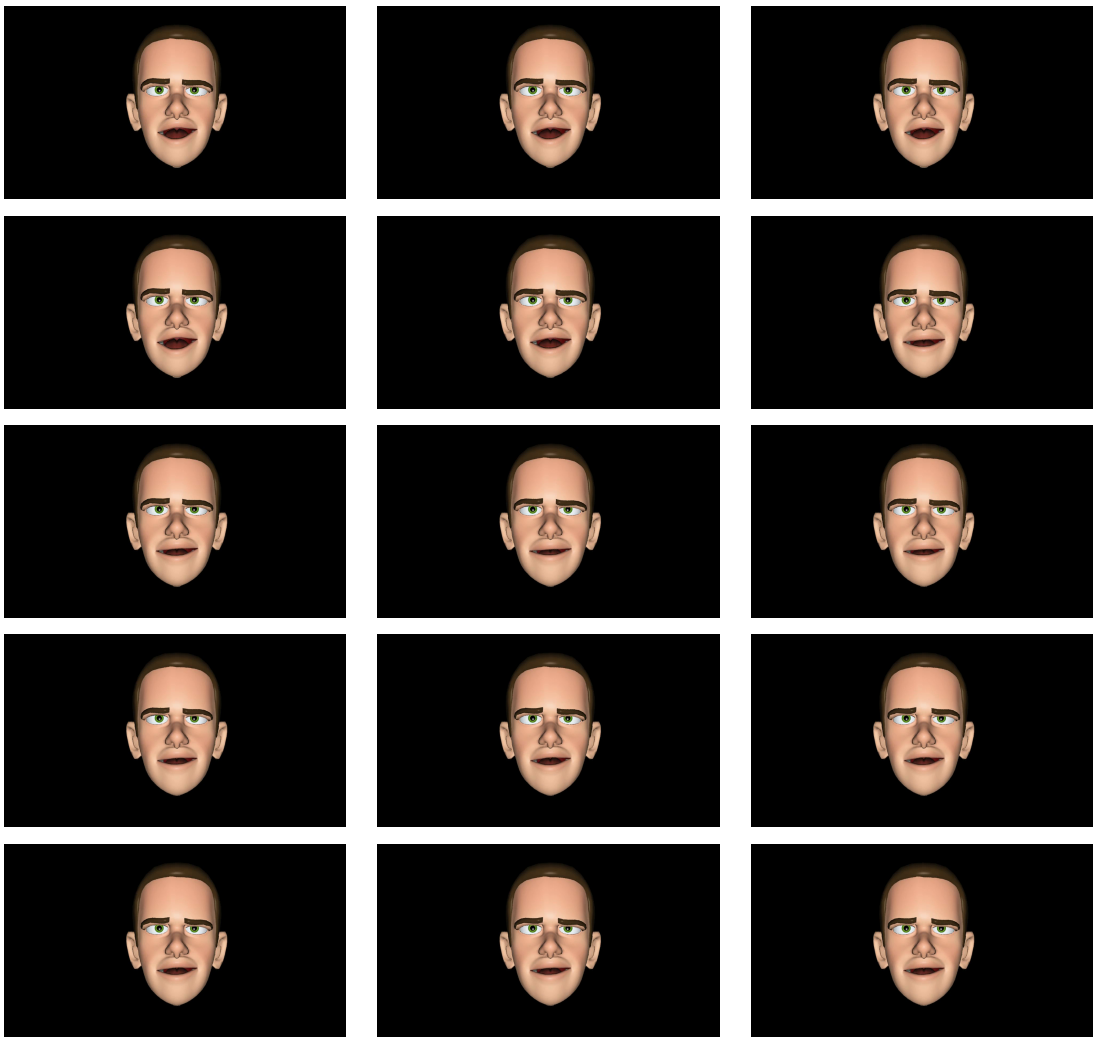


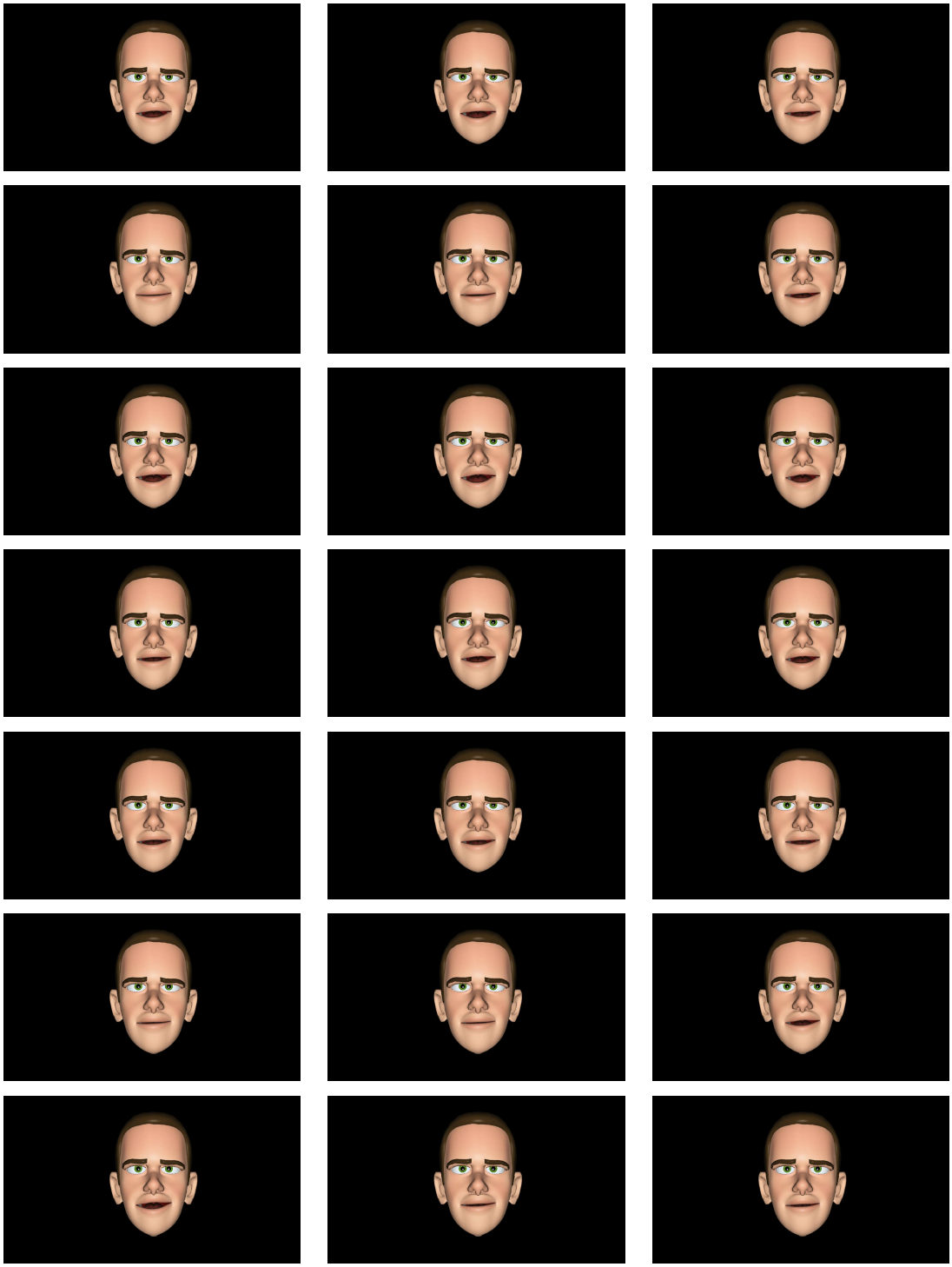


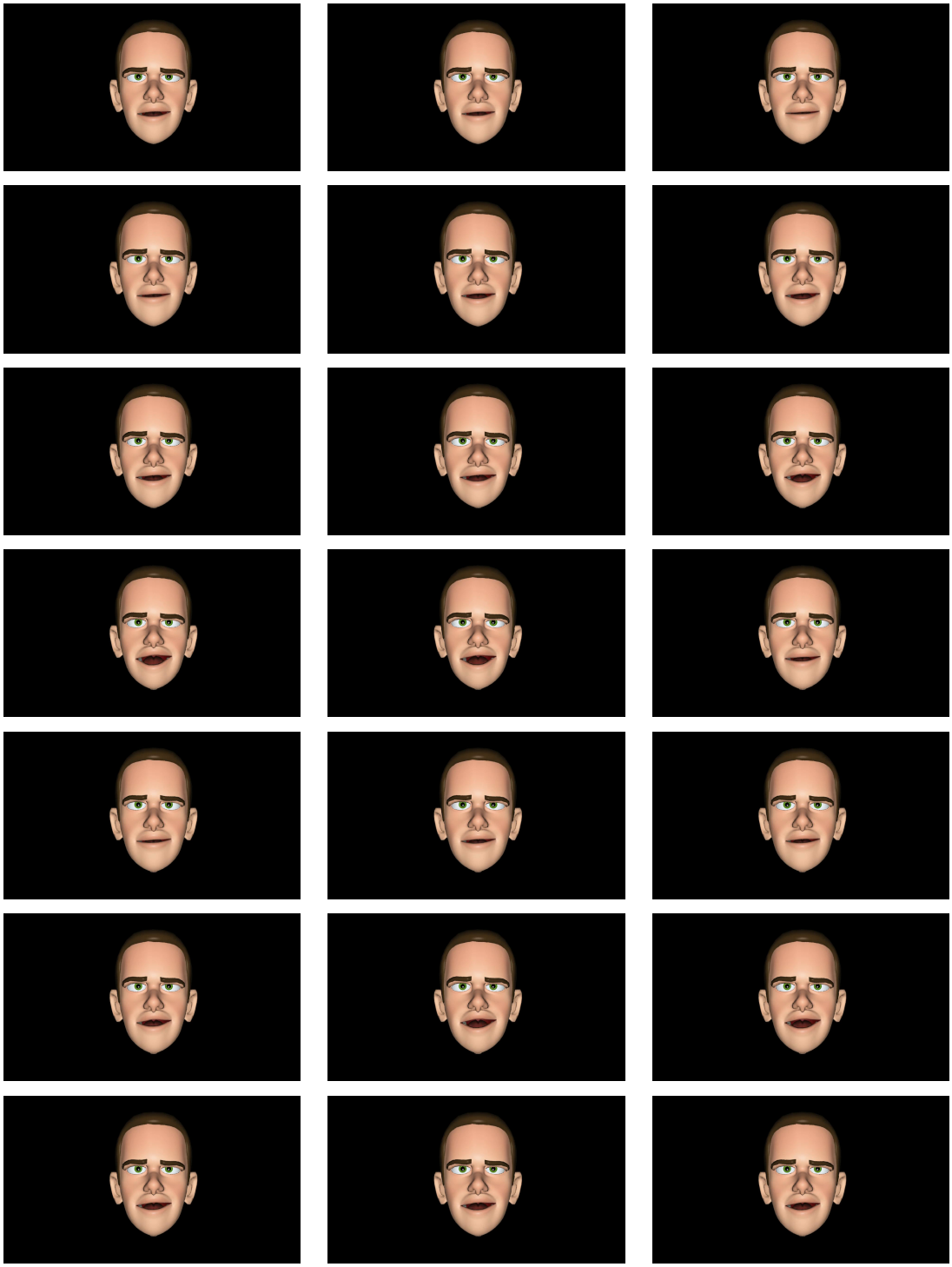


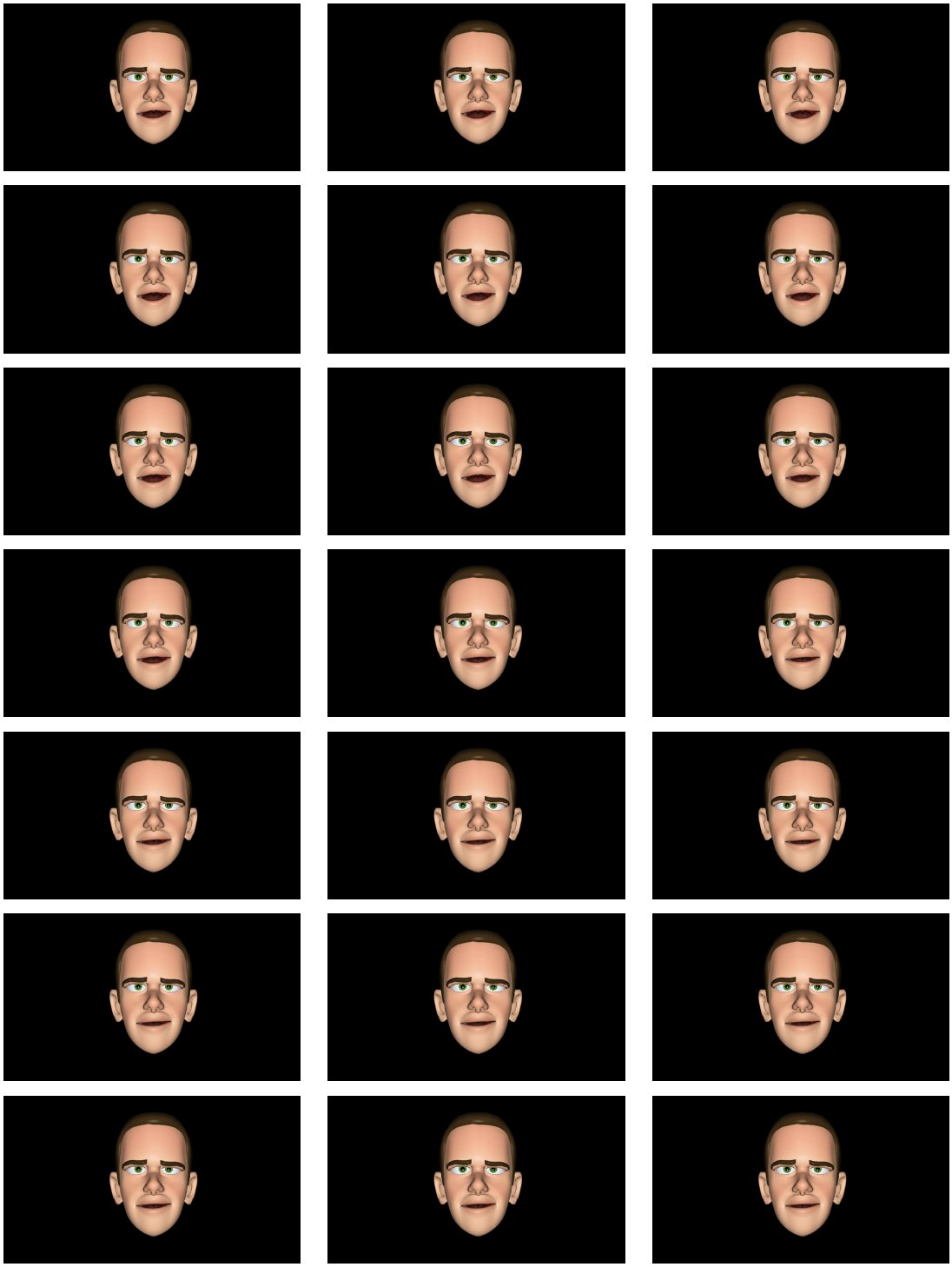
F.3 Rig Based - Modulation with Angry

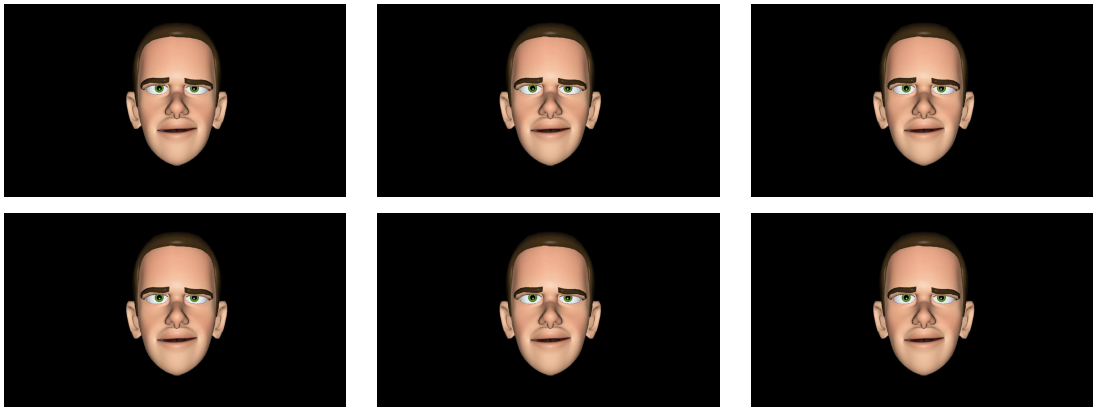
Here is the sequence shown in Section F.1, after DTW and modulation with anger.





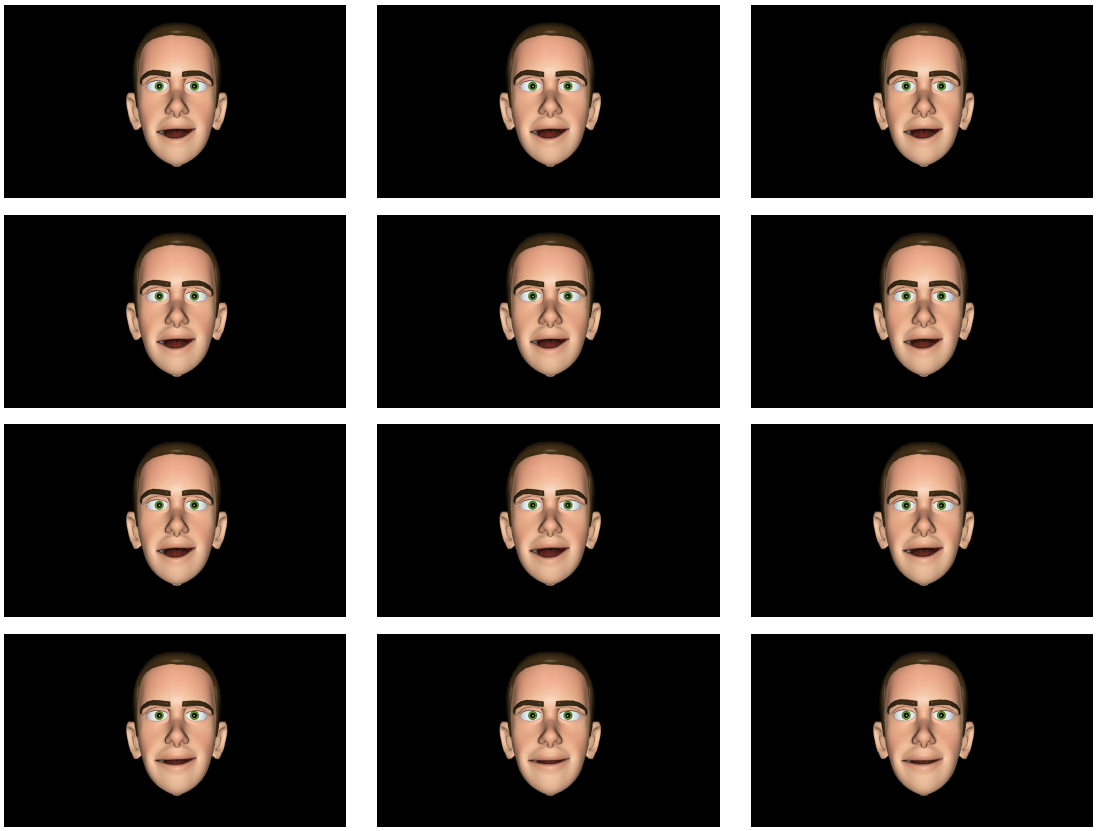


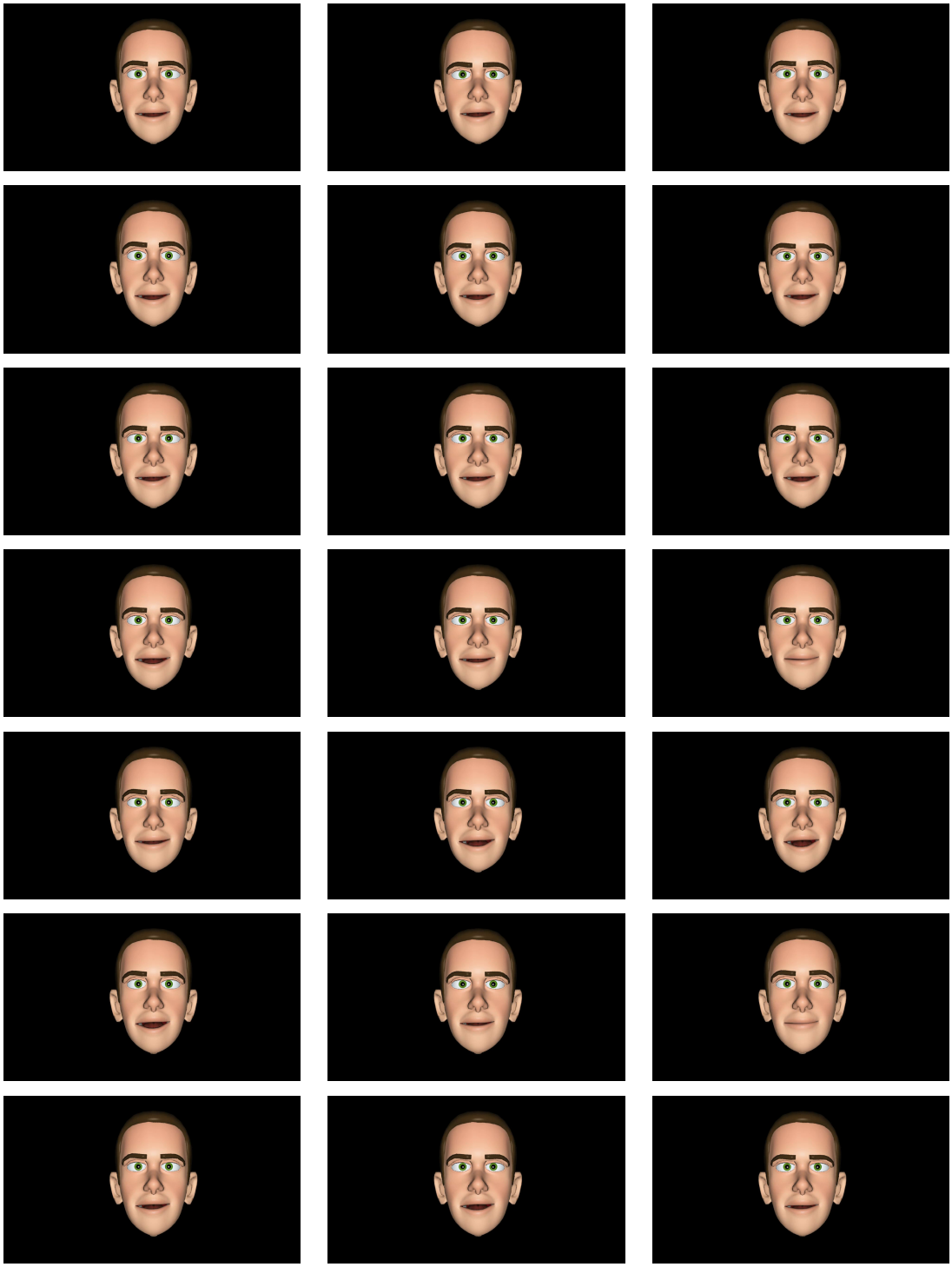


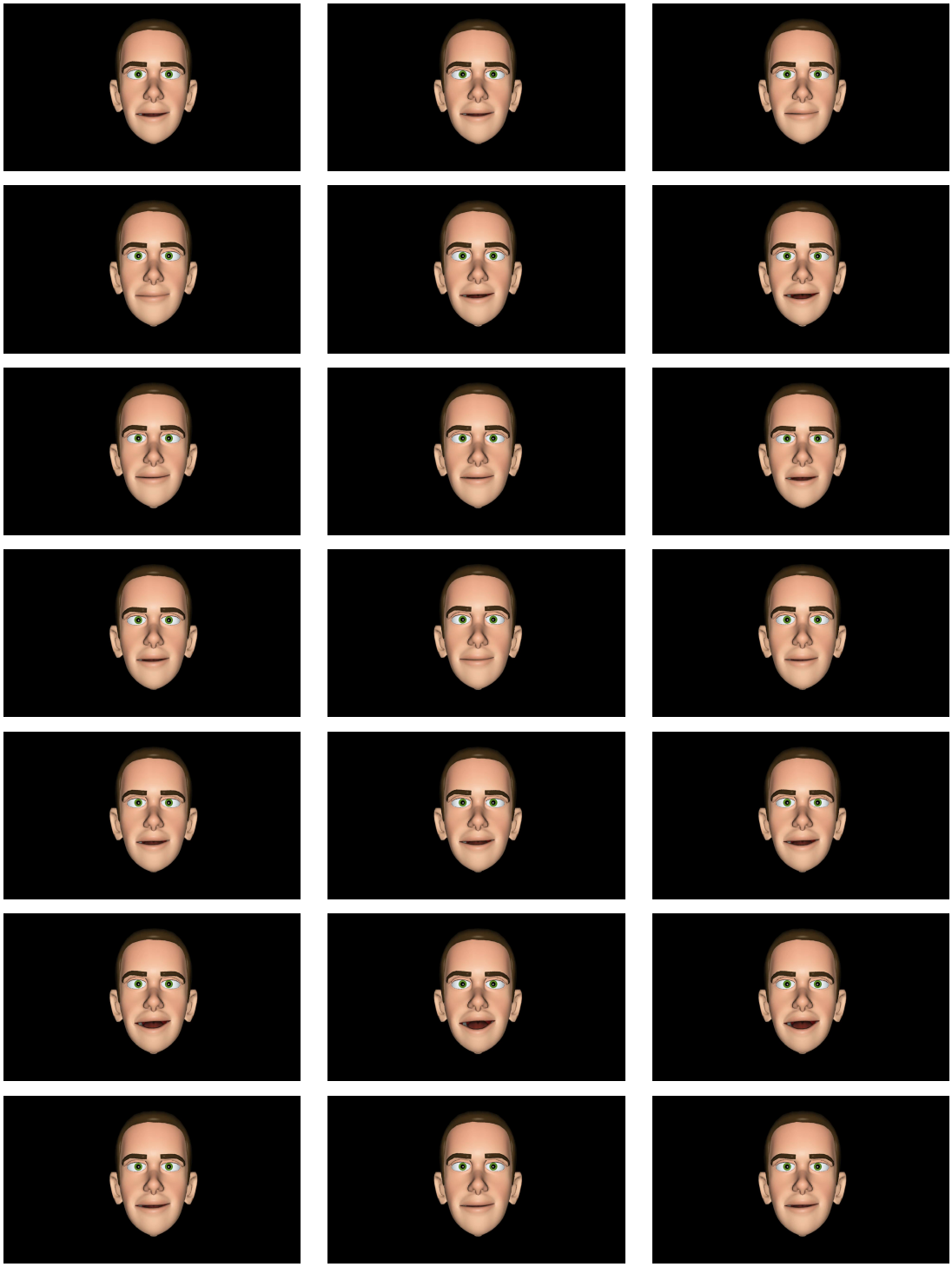


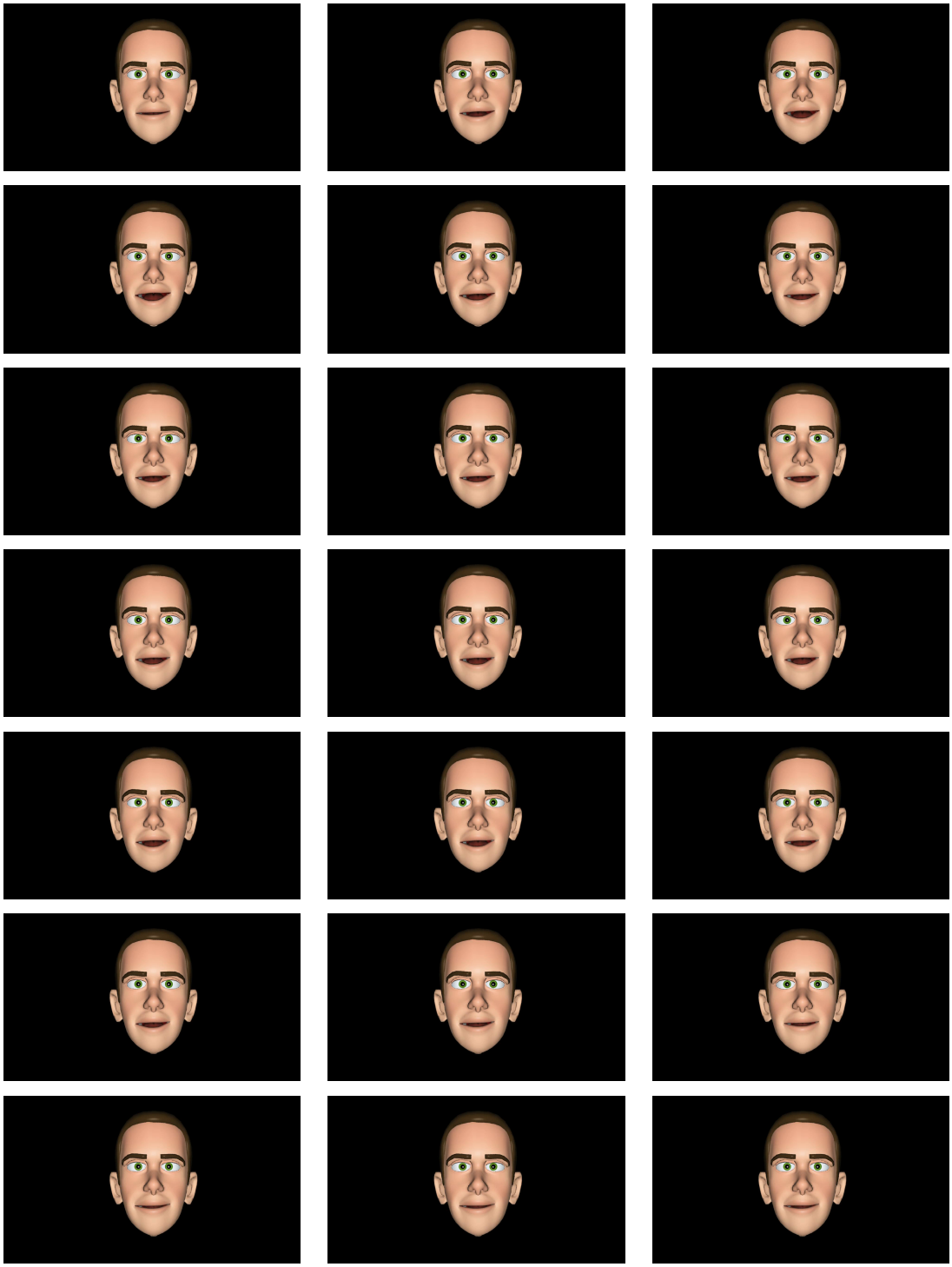
F.4 Rig Based - Modulation with Surprise

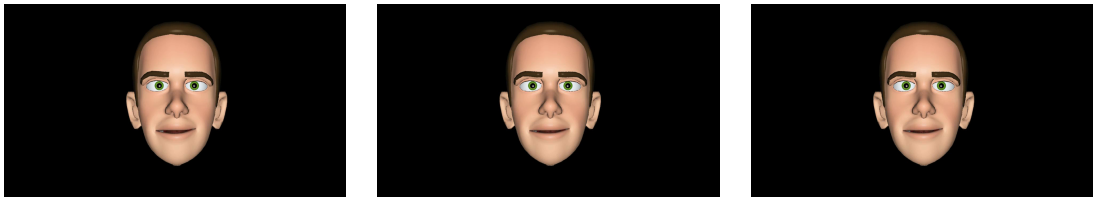
Here is the sequence shown in Section F.1, after DTW and modulation with surprise.





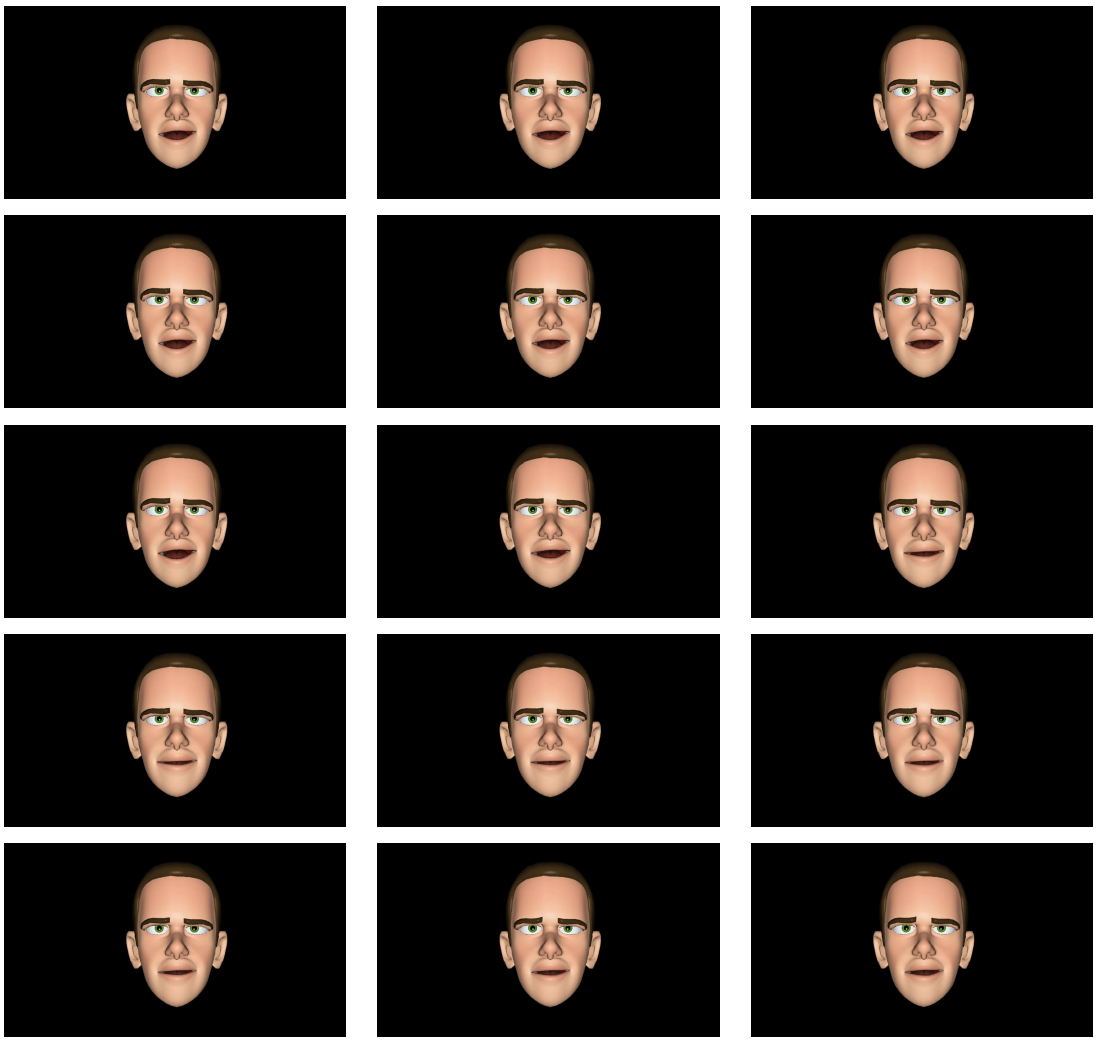


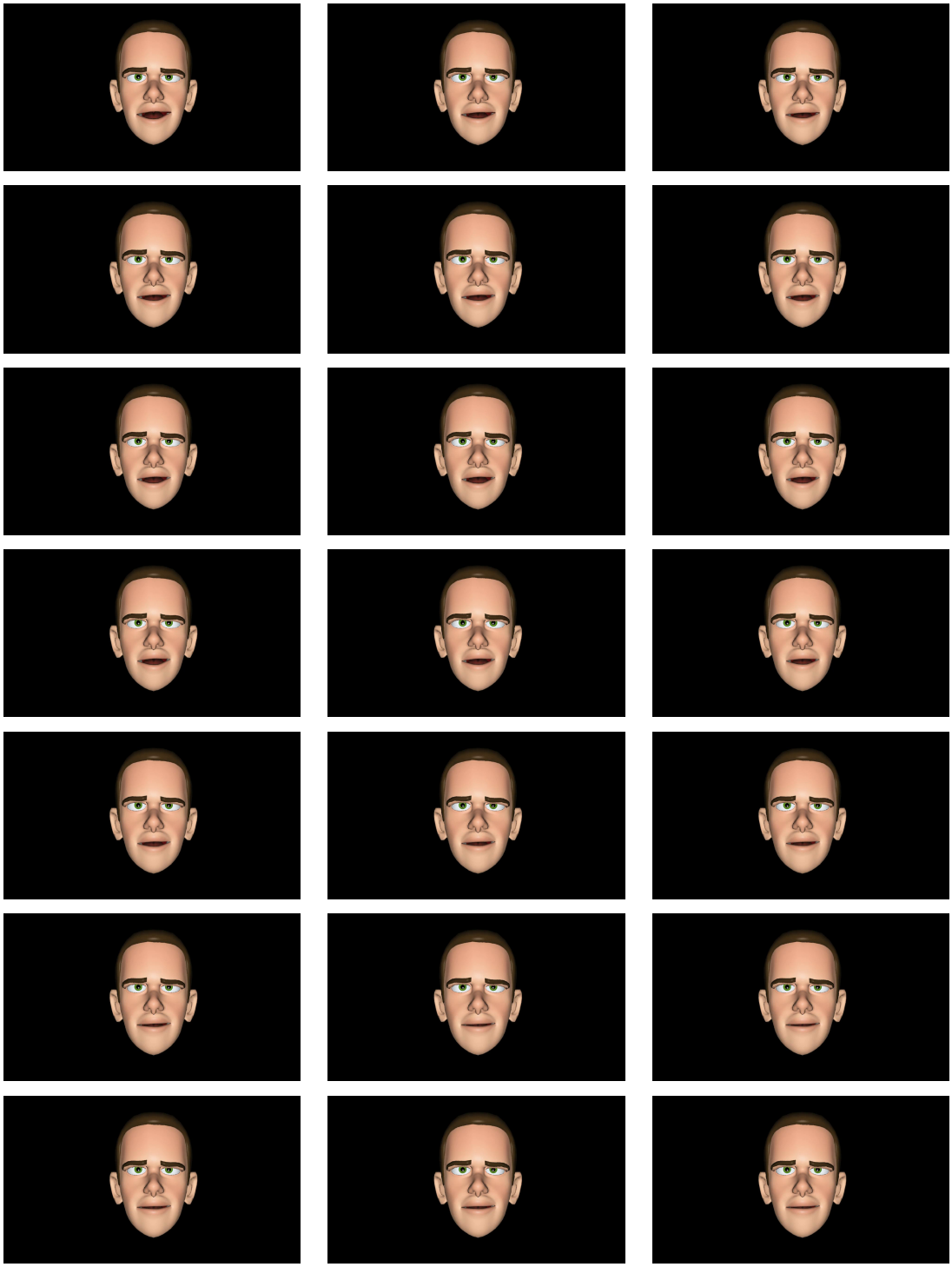


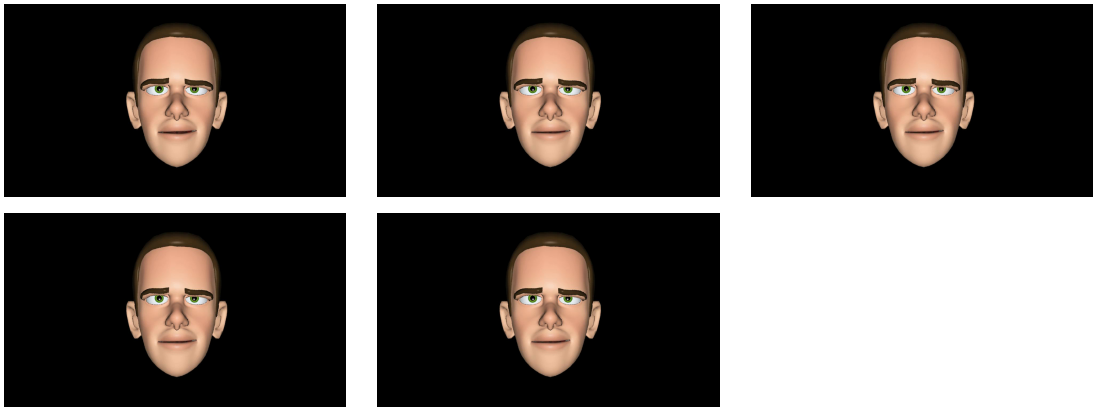


F.5 Rig Based - Modulation with Sadness

Here is the sequence shown in Section F.1, after DTW and modulation with sadness.







Appendix G

Expression Blending

Here are some examples of using a mixed model to interpolate of the gamut of training expressions to create new expressions.

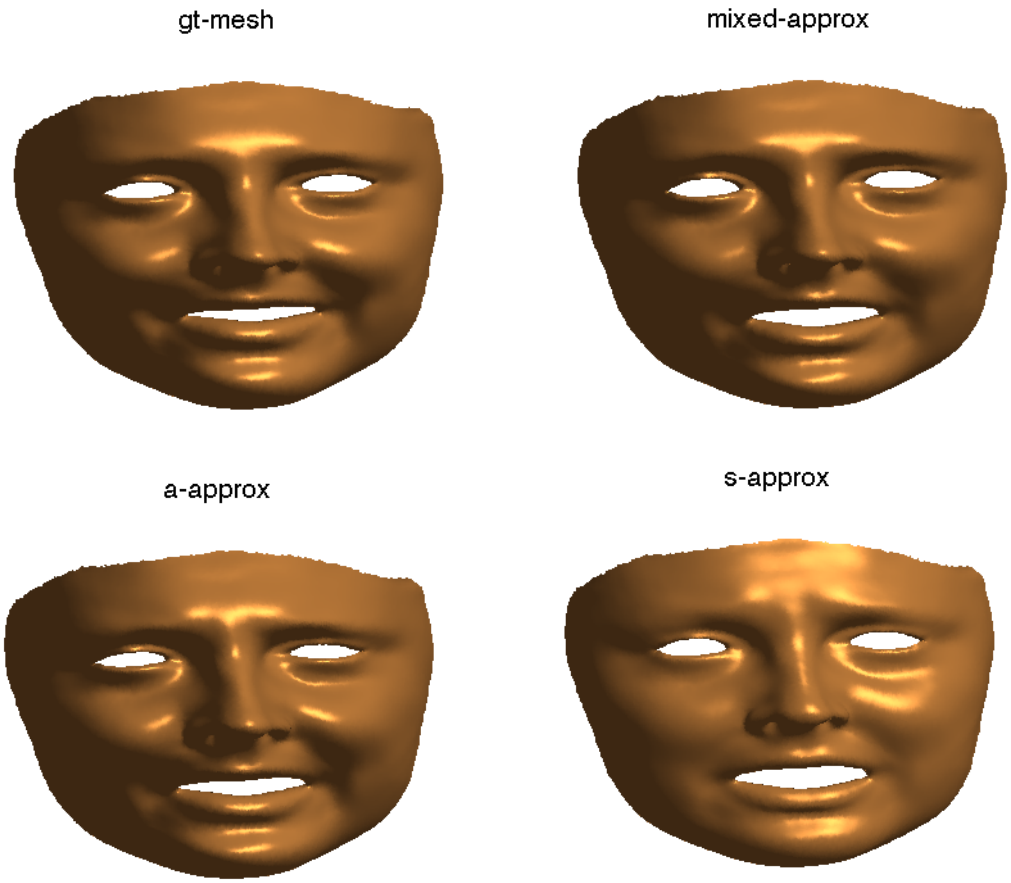


Figure G.1 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, sad model approximation

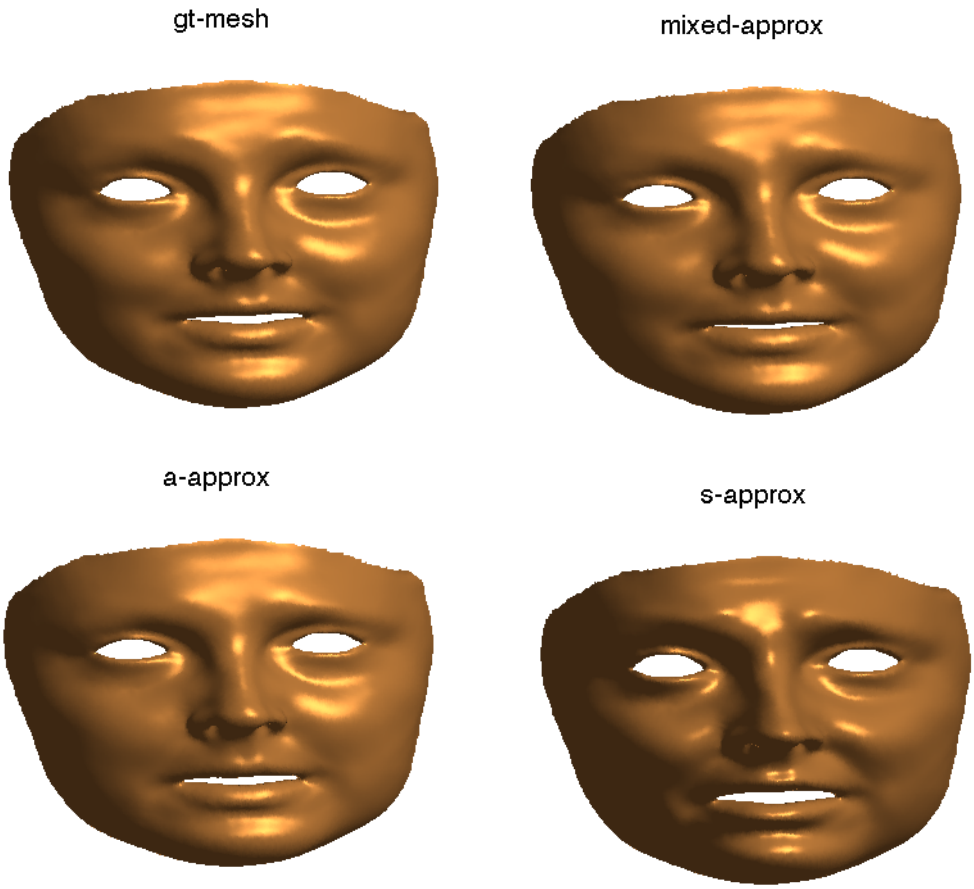


Figure G.2 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, sad model approximation

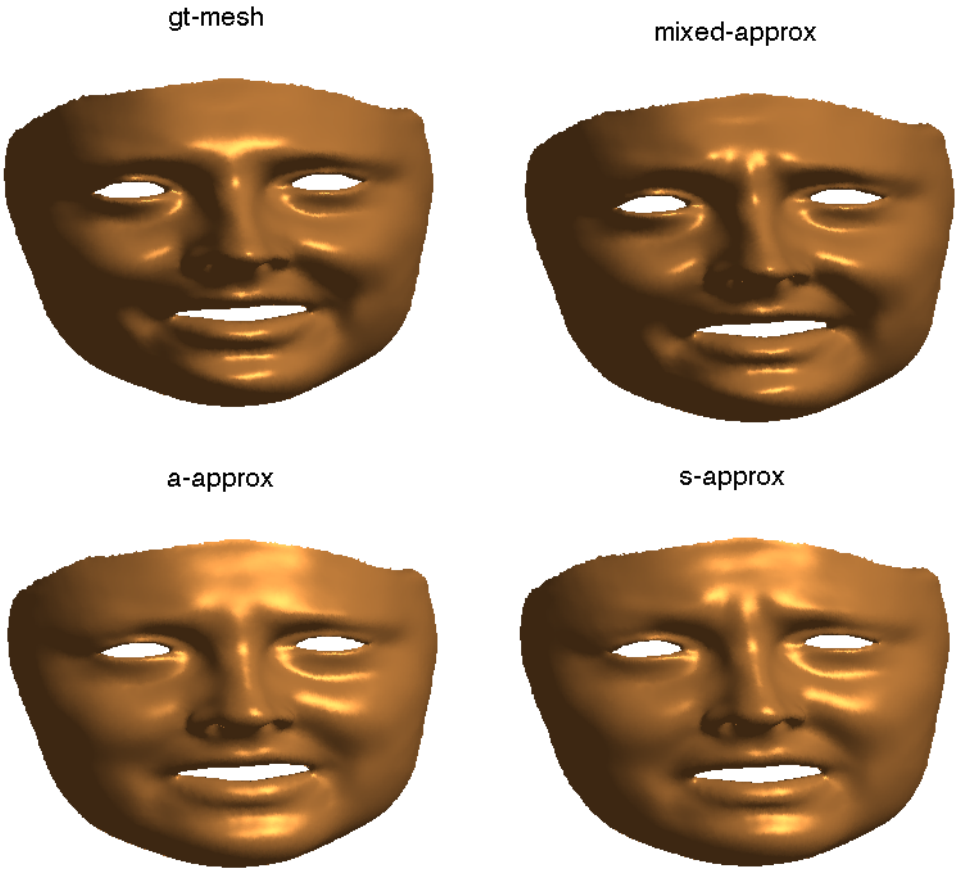


Figure G.3 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, sad model approximation

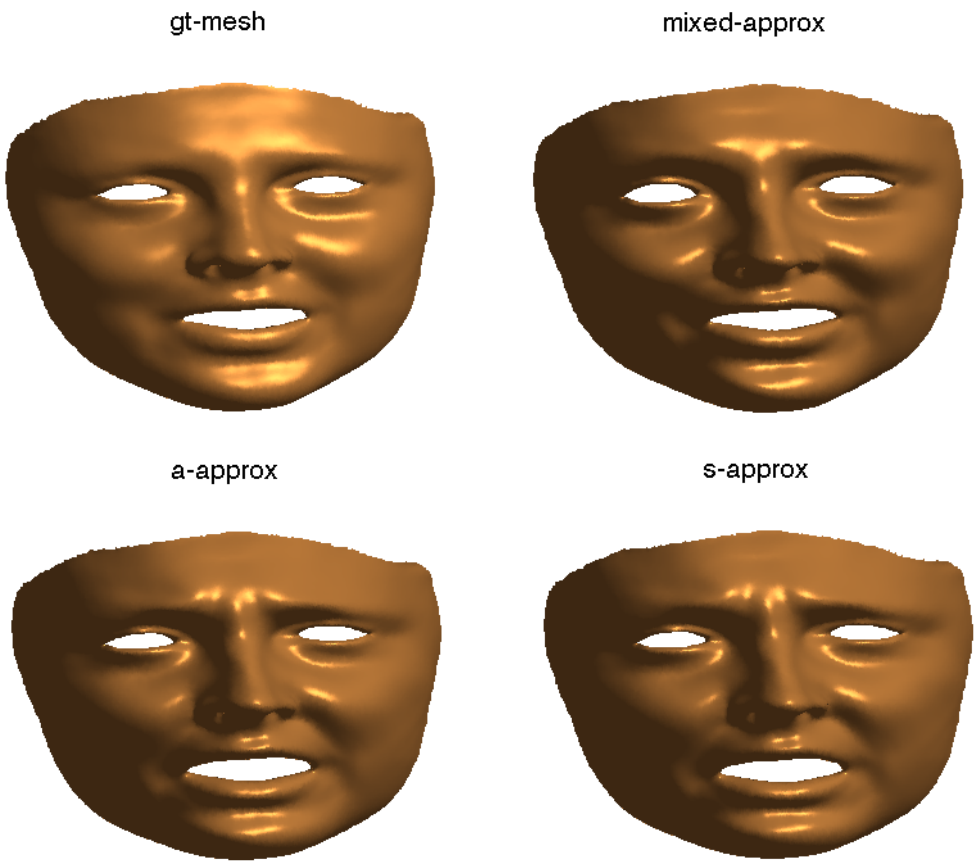


Figure G.4 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, sad model approximation

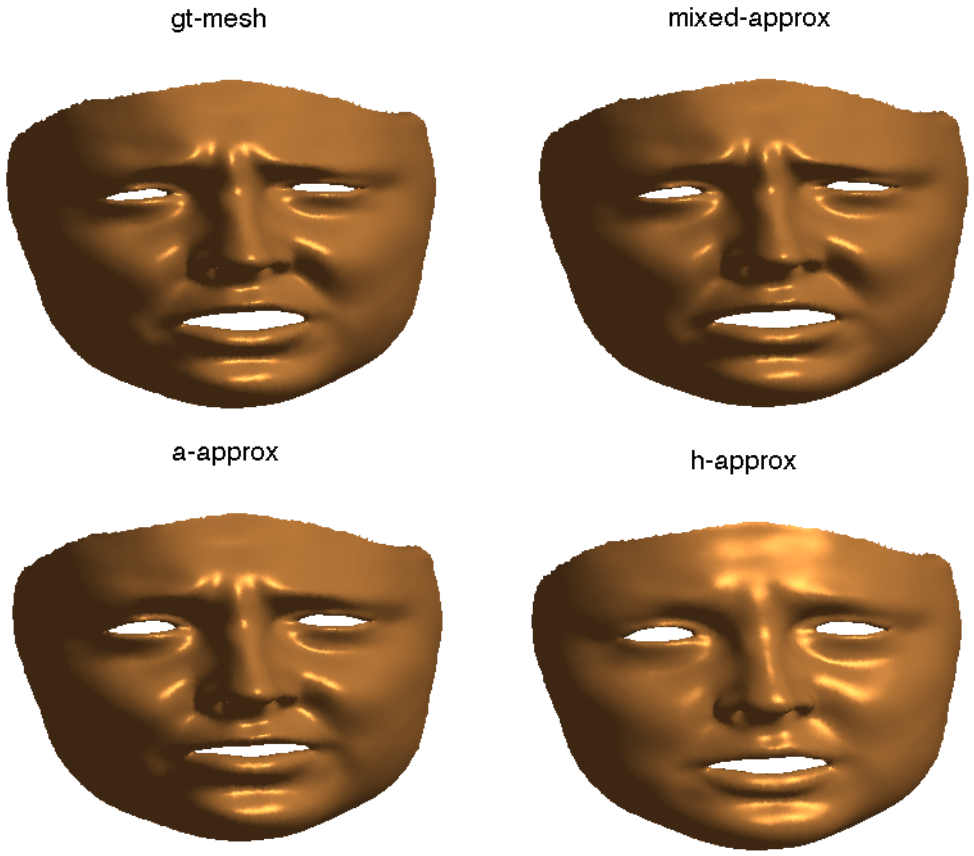


Figure G.5 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, happy model approximation

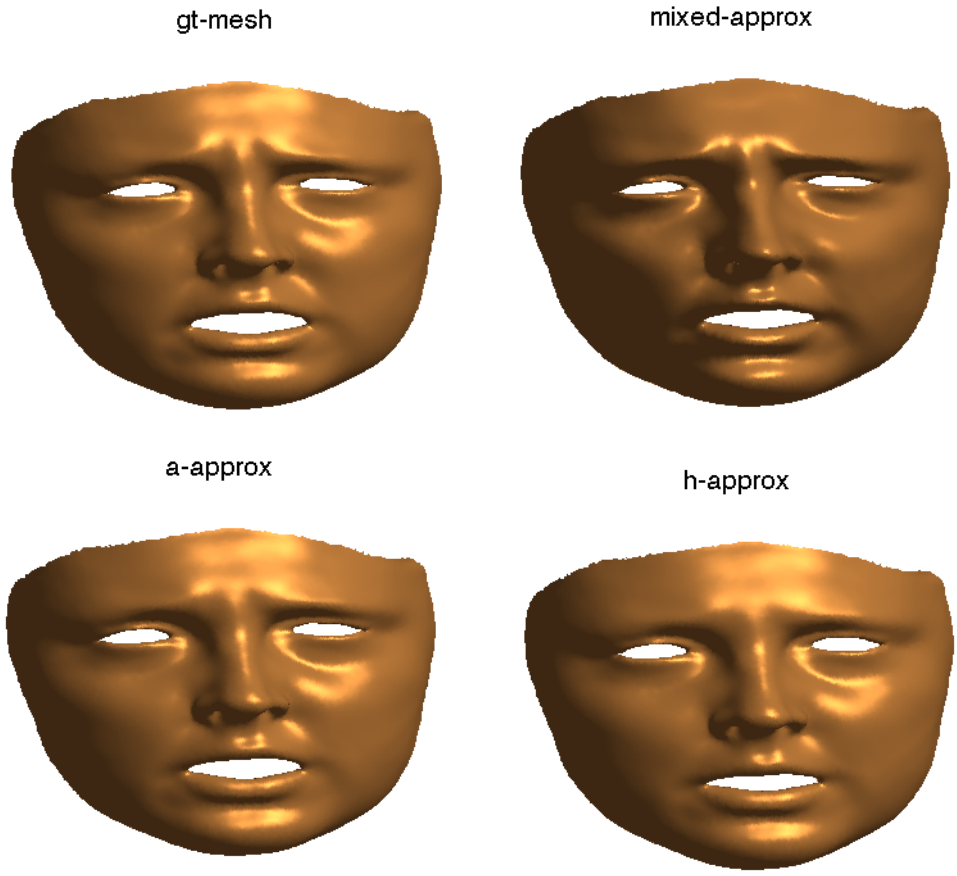


Figure G.6 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, happy model approximation

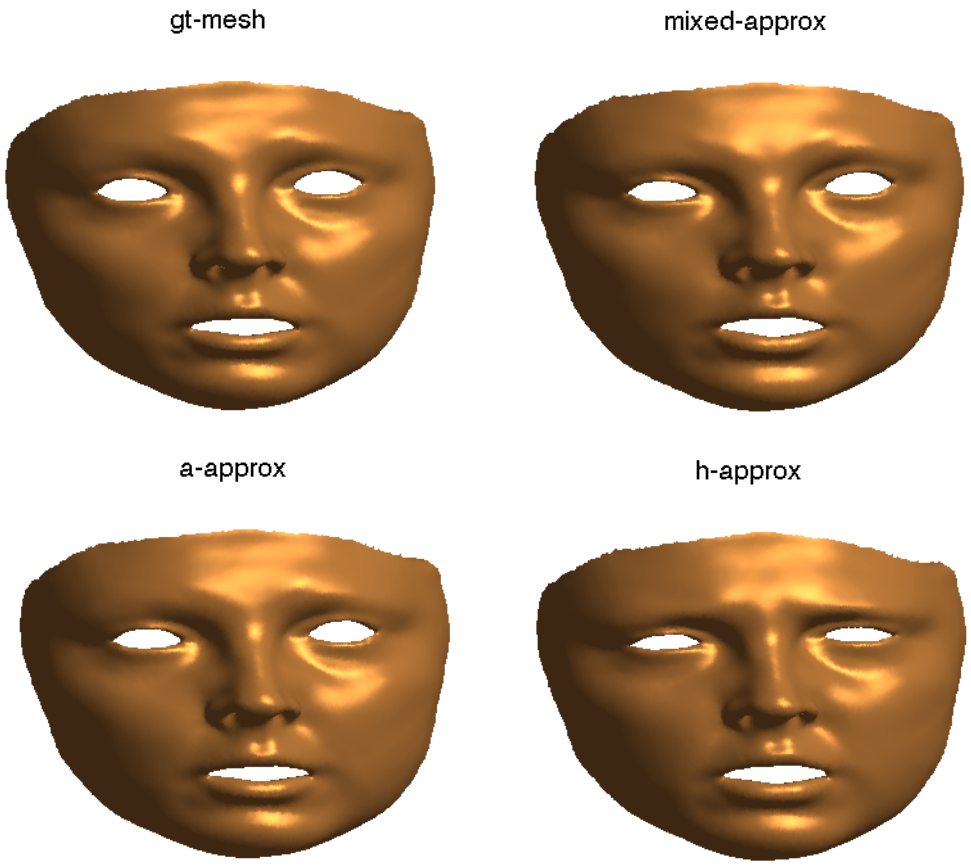


Figure G.7 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, happy model approximation

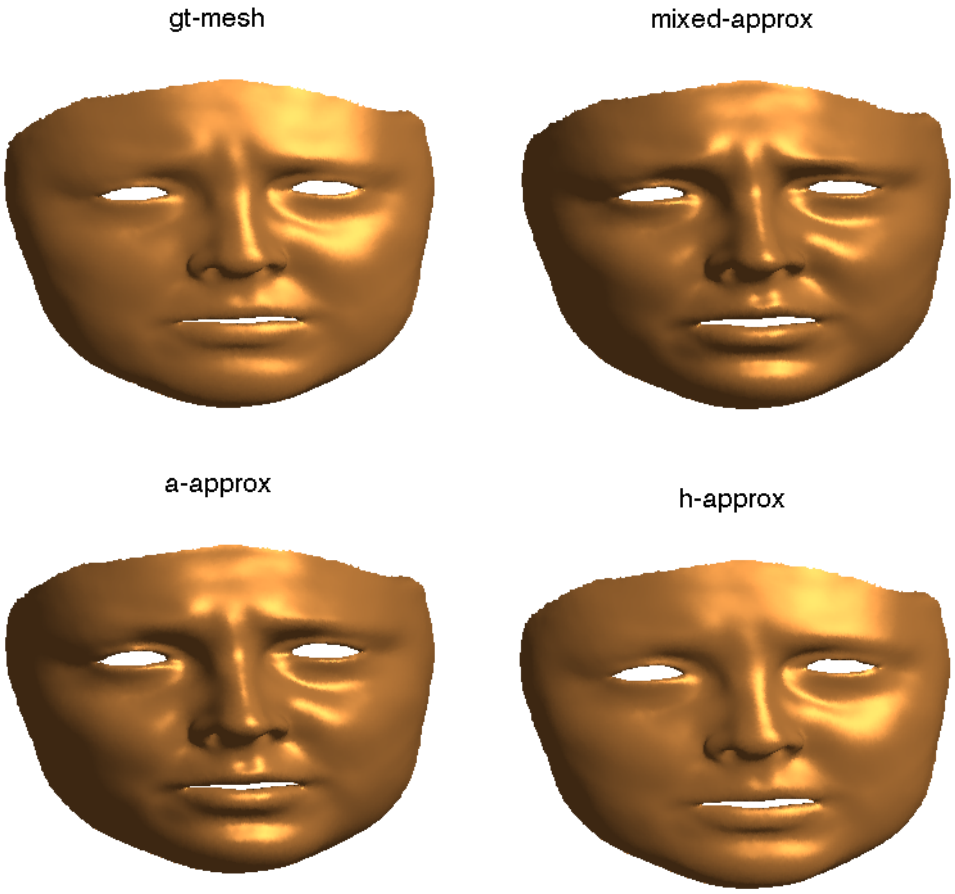


Figure G.8 In clockwise order from top left: ground truth, mixed model approximation, angry model approximation, happy model approximation

Bibliography

Albrecht, I., Schroder, M., Haber, J., and Seidel, H.-P. (2005). Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8(4):201–212.

Anderson, R., Stenger, B., Wan, V., and Cipolla, R. (2013). Expressive Visual Text-To-Speech Using Active Appearance Models. *Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Apple (2014). *Final Cut Pro*.

Arai, K., Kurihara, T., and Anjyo, K.-i. (1996). Bilinear interpolation for facial expression and metamorphosis in real-time animation. *The Visual Computer*, 12(3):105–116.

Barton, C. and Inghelbrecht, P. (1999). *Shazam*.

Bellard, F. (2014). *FFmpeg*.

Bermano, A. H., Bradley, D., Beeler, T., Zund, F., Nowrouzezahrai, D., Baran, I., Sorkine-Hornung, O., Pfister, H., Sumner, R. W., and Bickel, B. (2014). Facial performance enhancement using dynamic shape space analysis. *ACM Transactions on Graphics (TOG)*, 33(2):13.

Beskow, J. and Nordenberg, M. (2005). Data-driven synthesis of expressive visual speech using an MPEG-4 talking head. *Proc. Interspeech - 2005*, pages 793–796.

Bevacqua, E., Mancini, M., and Niewiadomski, R. (2007). An expressive ECA showing complex emotions. In *Proceedings of the ...*.

Bevacqua, E. and Pelachaud, C. (2004). Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 15(34):297–304.

Brand, M. (1999). Voice puppetry. In *the 26th annual conference*, pages 21–28, New York, New York, USA. ACM Press.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360. ACM Press/Addison-Wesley Publishing Co.

Brooke, N. M. and Scott, S. D. (1994). Computer graphics animations of talking faces based on stochastic models. In *ICSIPNN '94. International Conference on Speech, Image Processing and Neural Networks*, pages 73–76. IEEE.

Buisine, S., Abrilian, S., Niewiadomski, R., Martin, J.-C., Devillers, L., and Pelachaud, C. (2006). Perception of blended emotions: From video corpus to expressive agent. In *Intelligent virtual agents*, pages 93–106. Springer.

Burton, J. (2010). Morpheus Facial Rig.

Cao, Y., Faloutsos, P., and Pighin, F. (2003). Unsupervised learning for speech motion editing. In *Proceedings of the 2003 ACM SIGGRAPH ....*

Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. (2005). Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302.

Chan, C. S. and Tsai, F. S. (2010). Computer animation of facial emotions. *Cyberworlds (CW), 2010 International Conference on*, pages 425–429.

Choe, B., Lee, H., and Ko, H. S. (2001a). Performance-driven muscle-based facial animation. *The journal of visualization and computer animation*, 12(2):67–79.

Choe, B., Lee, H., and Ko, H. S. (2001b). Performancedriven musclebased facial animation. *The journal of visualization and computer animation*, 12(2):67–79.

Chuang, E. (2002). Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report*.

Chuang, E. and Bregler, C. (2005). Mood swings: expressive speech animation. *Transactions on Graphics (TOG*, 24(2).

Chuang, E. S., Deshpande, F., and Bregler, C. (2002). Facial expression space learning. In *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*, pages 68–76. IEEE.

Cicconetti, C., Akyildiz, I. F., and Lenzini, L. (2009). FEBA: a bandwidth allocation algorithm for service differentiation in IEEE 802.16 mesh networks. *IEEE/ACM Transactions on Networking (TON*, 17(3).

Clark, R. A. J., richmond, K., and King, S. (2004). Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proc. 5th ISCA workshop on speech synthesis*.

Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. *Models and techniques in computer animation*, pages 139–156.

Cootes, T. (2000). An introduction to active shape models. *Image Processing and Analysis*, pages 223–248.

Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685.

Cosatto, E. and Graf, H. P. (2000a). Photo-realistic talking-heads from image samples. *Multimedia, IEEE Transactions on*, 2(3):152–163.

Cosatto, E. and Graf, H. P. (2000b). Photo-realistic talking-heads from image samples. *Multimedia, IEEE Transactions on*, 2(3):152–163.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schroder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Cox, S. J. and Simons, A. D. (1990). Generation of Mouthshapes for a Synthetic Talking Head. In *Proceesings of the Institute of Accoustics*, pages 475–482.

Curio, C., Breidt, M., Kleiner, M., Vuong, Q. C., Giese, M. A., and Bülthoff, H. H. (2006). Semantic 3D motion retargeting for facial animation. In *APGV '06: Proceedings of the 3rd symposium on Applied perception in graphics and visualization*. ACM Request Permissions.

DeMarco, A. and Cox, S. J. (2013). Native accent classification via i-vectors and speaker compensation fusion. In *Proc. Interspeech 2013*, pages 1472–1476.

Deng, Z., Bulut, M., Neumann, U., and Narayanan, S. (2004). Automatic dynamic expression synthesis for speech animation. *Proc. of IEEE Computer Animation and Social Agents 2004*, pages 267–274.

Deng, Z., Chiang, P.-Y., Fox, P., and Neumann, U. (2006). Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 43–48. ACM.

Deng, Z. and Neumann, U. (2006). eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls. *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 251–260.

Der, K. G., Sumner, R. W., and Popovic, J. (2006). Inverse kinematics for reduced deformable models. *SIGGRAPH '06: SIGGRAPH 2006 Papers*.

Du, Y. and Lin, X. (2003). Emotional facial expression model building. *Pattern recognition letters*, 24(16):2923–2934.

Edge, J. D. and Maddock, S. (2001). Expressive visual speech using geometric muscle functions. *Proc. Eurographics UK. 2001*, pages 11–18.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Ekman, P. and Friesen, W. V. (1977). *Facial Action Coding System*. Consulting Psychologists Press, Stanford University, Palo Alto.

Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., and Van Gool, L. (2010a). A 3-d audio-visual corpus of affective communication. *Multimedia, IEEE Transactions on*, 12(6):591–598.

Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., and Van Gool, L. (2010b). Acquisition of a 3d audio-visual corpus of affective speech. *BIWI technical report no. 270, Computer Vision Lab, ETH Zürich*, (270).

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Forney, Jr, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Gavert, H., Hurri, J., Sarela, J., and Hyvarinen, A. (2005). FastICA. page GNU GPL Version 2.

Geiger, G., Ezzat, T., and Poggio, T. (2003). Perceptual evaluation of video-realistic speech. Technical report.

Hong, P., Wen, Z., and Huang, T. S. (2002). Real-time speech-driven face animation with expressions using neural networks. *Neural Networks, IEEE Transactions on*, 13(4):916–927.

Huang, F. J., Cosatto, E., and Graf, H. P. (2002). Triphone based unit selection for concatenative visual speech synthesis. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, pages II–2037. IEEE.

Huckvale, M. A., Leff, J., and Williams, G. (2013). Avatar Therapy: and audio-visual dialogue system for treating auditory hallucinations. *Interspeech 2013*.

Hyvarinen, A. (1997). A family of fixed-point algorithms for independent component analysis. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, pages 3917–3920.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., and Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, page 201200155.

Jia, J., Zhang, S., Meng, F., Wang, Y., and Cai, L. (2011). Emotional Audio-Visual Speech Synthesis Based on PAD. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):570–582.

Joshi, P., Tien, W. C., Desbrun, M., and Pighin, F. (2005). Learning controls for blend shape based realistic facial animation. *ACM SIGGRAPH 2005 Courses*, page 8.

Kahler, K., Haber, J., and Seidel, H.-P. (2001). Geometry-based muscle modeling for facial animation. In *Graphics Interface*, pages 37–46.

Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386.

Kholgade, N., Matthews, I., and Sheikh, Y. (2011). Content Retargeting Using Parameter-Parallel Facial Layers. In *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2011)*, pages 195–204, New York, New York, USA. ACM, ACM Press.

Kleiser, J. (1989). A fast, efficient, accurate way to represent the human face. *Course Notes on State of the Art in Facial Animation, SigGraph '89*, 22:35–40.

Kshirsagar, S., Molet, T., and Magnenat-Thalmann, N. (2001). Principal components of expressive speech animation. *Computer Graphics International 2001. Proceedings*, pages 38–44.

Lee, Y., Terzopoulos, D., and Waters, K. (1995). Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, New York, New York, USA. ACM.

Leff, J., Williams, G., Arbuthnot, M., and Leff, A. P. (2013). Silencing voices: a proof-of-concept study of computer-assisted therapy for medication-resistant auditory hallucinations. *British Journal of Psychiatry*.

Lewis, J. P., Mooser, J., Deng, Z., and Neumann, U. (2005a). Reducing blendshape interference by selected motion attenuation. *Proceedings of the 2005 symposium on Interactive 3D graphics and games. ACM, 2005.*, pages 25–29.

Lewis, J. P., Mooser, J., Deng, Z., and Neumann, U. (2005b). Reducing blendshape interference by selected motion attenuation. In *Proceedings of the 2005 symposium on Interactive 3D graphics and games*, pages 25–29. ACM.

Lines, J., Bagnall, A., Caiger-Smith, P., and Anderson, S. (2011). Classification of household devices by electricity usage profiles. In *Intelligent Data Engineering and Automated Learning-IDEAL 2011*, pages 403–412. Springer.

Liu, K. and Ostermann, J. (2011). Realistic facial expression synthesis for an image-based talking head. *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6.

Liu, Z., Shan, Y., and Zhang, Z. (2001). Expressive Expression Mapping with Ratio Images . In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, pages 271–276, New York, New York, USA. ACM Press.

Magnenat-Thalmann, N., Primeau, E., and Thalmann, D. (1988). Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297.

Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, J., and Tokuda, K. (1998). Text-to-visual speech synthesis based on parameter generation from hmm. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, pages 3745–3748. IEEE.

Mathworks (2015). Kurtosis Matlab.

Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164.

Mazzei, D., Lazzeri, N., Hanson, D., and De Rossi, D. (2012). HEFES: An Hybrid Engine for Facial Expressions Synthesis to control human-like androids and avatars. In *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*, pages 195–200.

Mlakar, I. and Rojc, M. (2011). Towards ECA's Animation of Expressive Complex Behaviour. *Analysis of Verbal and Nonverbal Communication and Enactment.*, pages 185–198.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The Uncanny Valley [From the Field]. *Robotics & Automation Magazine, IEEE*, 19(2):98–100.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, pages 69–84.

Nelder, J. A. and Mead, R. (1965a). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Nelder, J. A. and Mead, R. (1965b). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Noh, J.-y. and Neumann, U. (2006). Expression cloning. *SIGGRAPH '06: SIGGRAPH 2006 Courses*.

Ostermann, J. (2002). Face animation in mpeg-4. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, pages 17–55.

Parke, F. I. (1972). Computer generated animation of faces. In *Proceedings of the ACM annual conference-Volume 1*, pages 451–457. ACM.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pelachaud, C., Badler, N. I., and Steedman, M. (1991). Linguistic issues in facial animation. In *Computer animation'91*, pages 15–30. Springer.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D. H. (2006). Synthesizing realistic facial expressions from photographs. *ACM SIGGRAPH 2006 ...*, page 19.

Platt, S. M. and Badler, N. I. (1981). Animating facial expressions. In *ACM SIGGRAPH computer graphics*, pages 245–252. ACM.

Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Rabiner, L. R., Rosenberg, A. E., and Levinson, S. E. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *The Journal of the Acoustical Society of America*, 63(S1):S79–S79.

Ratanamahatana, C. A. and Keogh, E. (2004). Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*, pages 22–25.

Rumman, N. A. and Fratarcangeli, M. (2014). Position based skinning of skeleton-driven deformable characters. In *SCCG '14: Proceedings of the 30th Spring Conference on Computer Graphics*. ACM Request Permissions.

Schmidt, M. N. and Olsson, R. K. (2006). Single-channel speech separation using sparse non-negative matrix factorization. *Spoken Language Proceesing*.

Seif El-Nasr, M., Ioerger, T. R., Yen, J., House, D. H., and Parke, F. I. (1999). Emotionally expressive agents. In *Computer Animation, 1999. Proceedings*, pages 48–57. IEEE.

Serra, J., Ribeiro, M., Freitas, J., Orvalho, V., and Dias, M. S. (2012). A proposal for a visual speech animation system for European Portuguese. *Advances in Speech and Language Technologies for Iberian Languages*, pages 267–276.

Sifakis, E., Neverov, I., and Fedkiw, R. (2005). Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM Transactions on Graphics (TOG)*, pages 417–425. ACM.

Sifakis, E., Selle, A., Robinson-Mosher, A., and Fedkiw, R. (2006). Simulating speech with a physics-based facial muscle model. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 261–270. Eurographics Association.

Stamp, M. (2004). A revealing introduction to hidden Markov models. *Department of Computer Science San Jose State University*.

Sumner, R. W. and Popovic, J. (2004). Deformation transfer for triangle meshes. *SIGGRAPH '04: SIGGRAPH 2004 Papers*.

Tamura, M., Masuko, T., Kobayashi, T., and Tokuda, K. (1998). Visual speech synthesis based on parameter generation from HMM: Speech-driven and text-and-speech-driven approaches. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.

Tao, J., Xin, L., and Yin, P. (2009). Realistic visual speech synthesis based on hybrid concatenation method. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):469–477.

Taylor, S. L., Mahler, M., and Theobald, B. J. (2012). Dynamic Units of Visual Speech. *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2012)*.

Terzopoulos, D. and Waters, K. (1990). Physicallybased facial modelling, analysis, and animation. *The journal of visualization and computer animation*, 1(2):73–80.

Terzopoulos, D. and Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(6):569–579.

Theobald, B. (2007). Audiovisual speech synthesis. *International Congress on Phonetic Sciences*.

Theobald, B., Matthews, I., and Cohn, J. (2007a). Real-time expression cloning using appearance models. *Proceedings of the 9th ....*

Theobald, B. J. (2014). *AAM Toolkit*. University of East Anglia.

Theobald, B. J. and Matthews, I. (2012). Relating Objective and Subjective Performance Measures for AAM-Based Visual Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2378–2387.

Theobald, B.-J., Matthews, I. A., Cohn, J. F., and Boker, S. M. (2007b). Real-time expression cloning using appearance models. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*. ACM Request Permissions.

Theobald, B.-J. and Wilkinson, N. (2008). A probabilistic trajectory synthesis system for synthesising visual speech. In *INTERSPEECH*, pages 1857–1860.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, pages 1315–1318.

Vlasic, D., Brand, M., Pfister, H., and Popovic, J. (2006). Face transfer with multilinear models. *SIGGRAPH '06: SIGGRAPH 2006 Courses*.

Waters, K. (1987). A muscle model for animation three-dimensional facial expression. *ACM SIGGRAPH Computer Graphics*, 21(4):17–24.

Weise, T., Leibe, B., and Van Gool, L. (2007). Fast 3d scanning with automatic motion compensation. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.*, pages 1–8.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83.

Young, S. and Woodland, P. (2009). *The Hidden Markov Model Toolkit (HTK)*. University of Cambridge.

Yu, H., Garrod, O., Jack, R., and Schyns, P. (2014). A framework for automatic and perceptually valid facial expression generation. *Multimedia Tools and Applications*.

Yu, H., Garrod, O. G. B., and Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics*, 36(3):152–162.

Yu, H. and Liu, H. (2014). Regression-based facial expression optimization. *IEEE Transactions on Human-Machine Systems*, pages 386–394.