NANYANG TECHNOLOGICAL UNIVERSITY

Undergraduate Research Experience on CAmpus
AY 2012-13

Talking Head: Literature Review

Authors:

Prerna Chikersal and Pooja Sukheja

Supervisor:

Prof. Chng Eng Siong

Contents

1 Introduction 2

2 Applications 3
2.1 Current Applications 3
2.2 Future Applications 3

3 Facial Modeling and Animation 4
3.1 Face Modeling 4
3.1.1 Generic Model Individualization 4
3.1.2 Example based facial modeling 6
3.2 Facial Animation 6
3.2.1 Geometry Manipulation Techniques 6
3.2.2 Image Manipulation Techniques 7

4 Facial Action Coding System (FACS) 9

5 Phonemes and Visemes 12

6 Some Examples of Talking heads 14

7 Conclusion 17
7.1 Linguistic Issues and Challenges 17
7.2 Our Approach 17

1 Introduction

In recent years, the development in the field of computer graphics and animation has made it easier to interact with computers. For example, education and learning has become an easy and interactive process with improvement in computer animation. Another essential instance is the creation of the talking head that has contributed to the improvement of Human Computer interaction to a great extent. A talking head is an individualized 2D or 3D model of a human head with computer based facial expressions and emotions. When synchronized with a speech synthesizer the talking head is capable of generating phonemes. The talking head represents a VTTS(Visual text to speech) interface in which the TTS produces sound and the face renderer generates appropriate visemes that creates the impression of a talking head[1]. Animated talking heads provide a natural interaction between humans and computers. They provide a user friendly interface for applications used for educational purposes[2] and various websites and blogs[3]. An application becomes even more eye-catching for the customers with the presence of a natural and photo-realistic life-like talking agent. The findings of a study examining the use of talking heads in online advertisements revealed that human faces attracted the most attention. For a talking head to be believable, it is essential to make it realistic, natural and personalized. The construction of a talking head has been a flourishing research topic in the field of computer graphics for more than 25 years .There are many methods and software available in the market to make personalized and photorealistic 2D and 3D models of a human faces which also incorporate ways to manipulate the facial attributes [5]. Although much work has been done in this field, more studies need to be conducted to improve the realism of the talking head. This paper focuses on examining the different existing tools used for creating a 2D or 3D model of a human face and evaluating each of the available software. This study is identified as important as it forms the basis of our research that is to be done to allow the easy creation of a photo-realistic human face which is capable of generating visemes and synchronizing those visemes to phonemes in order to represent a talking head.

2 Applications

Human computer interaction has been improved greatly with the use of animated agents that speak, gesture and convey emotions. Talking agents like these provide users with a more natural interaction with the computer. The possibility for using talking heads in software applications seems endless. In the following sections, some of the existing and future applications of talking heads have been reviewed.

2.1 Current Applications

Talking heads are currently a part of a variety of applications that empower online communication. One application of talking heads is to assist online users in the form of a virtual being .Conversational agents are used in web presentations to present information by verbally commenting on the graphics, video clips and text content embedded in the dynamic presentation. Additionally, they have the ability to signal emotions[4]. Use of facial animation combined with text-to speech synthesis has proved to be successful in supporting E care and E commerce[5]. Animated automated guides provide a friendly and helpful user interface to assist the users in web navigation[3]. Secondly, talking heads play an important role in the field of education by enhancing the effectiveness of teaching. Character animation makes the learning environment more interactive through animated pedagogical agents. These are intelligent talking heads which make the communication between students and computers more engaging and enjoyable[2].

2.2 Future Applications

Though a large number of applications of the talking head exist presently, it has the potential to be further developed. Facial expressions are a significant part of verbal communication since the face is a carrier of the phonetic content of human speech. Facial expressions play a key role in conveying emotions and revealing observable aspects of prosodic features of the voice. Expressive and animated talking faces can increase the precision of the speech especially during acoustic degraded conditions despite of whether the auditory degradations is due to external sound or hearing impairment[6]. Thus, computer applications with animated talking heads are capable of helping hearing impaired people in the future. Recent research shows that emotionally expressive avatars facilitate learning in people with autism. The facial expressions of the animated characters are easily identifiable and well distinctive which makes it easier for people with autism to understand[7]. A prospective application of avatars and talking heads can be to enhance the learning in people with autism spectrum disorders. Furthermore, photo realistic and personalized talking heads can play a crucial role in this system. Talking heads can also be used to supplement the mobile phones voice control feedback like Apple's SIRI. In addition, due to the popularity of talking heads in the field of research, it is believed that it has the potential to gain wider acceptance in the fields of the media and advertisement industry, virtual reality applications , online web applications and customer service industry[8].

3 Facial Modeling and Animation

Facial modeling and animation refers to the technique of visualising a human face on a computer and animating it such that the lip movements, expressions and other animations appear natural like that of a real person. The anatomy of a human head is very complex with six components, namely the skull, facial muscles, skin, eyes, teeth and tongue. The more precisely these components are simulated, the more realistic the animation will be. Unfortunately, even with recent advances in technology, simulating the anatomy accurately is still a challenge. Moreover, the facial features of people vary, due to difference in the shapes and sizes of their bones and muscles.

With recent advancements in this field, it has become possible to create very life-like human heads. Comparison between some readily available software tools that animate human faces or heads is shown in section 6. However, according to our survey, the applications which generate a 3D human head either do not allow the user to dynamically change the face being envisioned or require the face to be manually changed. This serves as a major limitation and as stated by Noh and Neumann[9], it highlights the goal for facial modeling and animation, which is "a system that 1) creates realistic animation, 2) operates in real time, 3) is automated as much as possible, and 4) adapts easily individual faces".

In this section, we will discuss the various techniques which have been and are being used for facial modeling and animation in past and current times.

3.1 Face Modeling

Facial modeling Facial modeling aims to create a high quality realistic facial model. Facial modeling and its various techniques have been an on-going research topic for the past few years. The diverse techniques of facial modeling can be categorized into two main approaches[10] 1. Generic Model Individualization 2. Example based facial modeling The following approaches have been discussed in detail below.

3.1.1 Generic Model Individualization

This is a generalized approach in which a generic model, which constitutes a standard generic mesh is adapted to the facial model of a specific person by using captured data. The captured data refers to the details of the positions of several facial features of the person like eyes and nose positions, mouth corners, which facilitate the adaption of the generic mesh to the individualized mesh. Hence, this method is also called model adaption. There are various ways of detecting the facial feature points of a person for the creation of the individualized model. The detected feature points are the basic inputs for this approach. These methods have been discussed in detail below.

- *Individualization from multi-view images* :
  he facial feature points are manually observed in a series of photographs taken from different angles. The generic model is then adapted to the feature points portrayed by the multiple images by using a scattered data interpolation technique for the deformation of the generic model. This method was improved by Lee et al.[11] who performed model adaption by using only 2 photographs, a frontal and side view image. Individualization of the generic model takes place in two steps, firstly a global matching is done using captured data to focus on the location of the facial features. Secondly, a detailed matching is performed to identify the shape of each facial feature. The feature points from the above

input methods are combined into 3d points which enable the deformation of the generalized model.

In the following three methods, anthropometric measurements are used to describe the facial features.

- *Individualization from anthropometric data* :
  Anthropometry is the science of measuring the human individual. In this technique, anthropometric data constitute the vital elements in the formation of face models to match different individuals. This approach of facial deformation was proposed by DeCarlo[12]. Anthropometric measurements describe the facial features of the human individual, consisting of feature points and their comparative distances. With these measurements of the human face, a static model of a human face is generated by variational modeling. Variational modeling is a methodology to optimize the 3D model of a face constrained by the features defined by anthropometric data.

- *Model adaption with a function based generic face* :
  The purpose of this process is to map a functional generalized face model to a individualized face model. It requires a set of anthropometric landmarks on the generic and specific face model. Projection mapping is carried out to obtain the anthropometric landmarks from the 3D model. Subsequently, global adaption is carried out for the adaption of the alignment and position of the generic model towards the specific model, depending on the recovered anthropometric 3D landmarks. Following that, a local adaption is done to make the generic model fit to the vertices of the specific scanned model.

- *Model adaption with a multi-layer model* :
  This process consists of a generic model with several layers which includes skin, mass-spring , skull , muscle and further separated constituents like eyes, mouth etc. Anthropometric data is collected from the models with the assistance of anthropometric landmark points used to define the facial features. The marked anthropometric landmark points facilitate the modeling of the generic model to the specific scanned 3D model.

- *Individualization using topographic representation* :
  In this method, the face features are detected by topographic analysis of the input, front view image. A topographic analysis of the input face image is performed to mark each pixel of the image as a topographic label. The data measured by these labels is used to optimize the individualization process by adapting the generalized model to the individualized model.

- *Individualization from video streaming* :
  Facial models are created by detecting features and capturing data through video cameras via stereo matching. This technique has highly specific demands of six synchronized videos working at 60 frames per second. The videos are taken from the left-side and the right-side with three videos having a left-view of the human face and the remaining three videos having a right view of the human face. Afterwards, a fitting and tracking process is employed to fit the face model to depth maps in the head frame and then perform tracking for all the subsequent frames. Depth maps are acquired by using a stereo algorithm which is employed during feature detection and data capturing. The products of the fitting process are good quality 3D meshes portraying the facial geometry and color of the individuals face.

3.1.2 Example based facial modeling

This procedure defines a method which is supported by an example set of existing face models to create the facial model of the individual with the required features.

- *Morphable face modeling*:
  Morphable modeling presented by Blanz and Vetter[13] has 2 necessities for the creation of the 3D face model. Firstly, a minimal input of a single photo by the user for feature detection. Example based face modeling requires the least number of input images as compared to any of the techniques of Generic model individualization that use photographs for detecting facial features however the method of individualization using topographic representation is an exception since it also requires a single image. Secondly, a collection of available 3D models. The shapes and textures of these example face models are transformed into a vector space representation. However, implementation is not feasible due to time constraints. Also, it is complicated as we need to establish exact correspondence between the models. Furthermore,Blanz et al.[14] proposed a face exchange technique which was based on the concept of morphable face modeling.

- *Multi linear modeling*:
  Multi-linear modeling is an alternative to morphable face modeling. Likewise, it requires pre-processing of the collected examples in order to achieve full correspondence between them. The process takes place as quoted herein. Then, these examples are organized in the form of a data tensor, which encodes model variations in terms of different attributes, such as identity, expression and viseme. This allows independent variation of each of these attributes. By using the organized data tensor, an arbitrary face model with desired facial expression can be modeled as a linear combination of these examples[10]. Face transfer uses face models which are made using multi-linear face modeling.

3.2 Facial Animation

Facial animation techniques can be broadly divided into two categories, that is, those based on geometry manipulation and those based on image manipulations[9]. 3-dimensional models of the face contain several vertices in space which together form polygons. Geometry manipulation techniques are used to animate such 3D models by manipulating these vertices. While, image manipulations are used to animate 2-dimensional photographs of the face by morphing from one photograph to another.

We will now discuss the two techniques in detail.

3.2.1 Geometry Manipulation Techniques

Some common geometry-based methods are described below:

- *Key-framing with interpolations*:
  Key-frame refers to the morph targets or the starting and ending points of any transition. In facial animation, each key-frame usually corresponds to an expression or a viseme. Morphing from one key-frame to another involves interpolating or moving the vertices of the face by small distances, frame by frame. Interpolation schemes are functions which describe movement of these vertices with time. Although, linear interpolation schemes may be used, expressions of a real person's face do not change linearly with time. So, cosine functions or other variations usually provide better results as demonstrated by the pioneering works of Parke[15].

- *Direct parameterization*:
  Even though, this method is based of the key-framing and interpolations approach, it is considerably superior as it allows explicit control of specific facial configurations. Combination of parameters provide a large range of facial expressions with relatively low computational costs. However, tedious manual tuning is required to set parameter values and the animation rarely looks natural[9].

- *Using Physics-based muscles*:
  There are three main techniques for facial animation using physics-based muscles: spring mesh muscle, vector muscle and layered spring mesh muscles. Platt and Badler[16] created a spring mesh model which generated realistic facial expressions by applying forces to elastic meshes through muscle arcs. Waters[17] created a very successful vector muscle model. He animated expressions like anger, fear, surprise, disgust, joy and happiness by implements the Facial Action Coding System (discussed in section **??**). Terzopoulos and Waters[18] proposed a layered spring mesh muscle model containing three layers, representing the cutaneous tissue, subcutaneous tissue, and muscle layer of the human face. Their approach created very realistic facial expressions, however, it required extensive computation.

- *Using pseudo-muscles*:
  This is a far more superior method than key-framing and interpolations and direct parameterisations. It is also less computationally extensive than the physics-based muscles approach. The physics-based muscles approach simulates the exact muscles long with its underlying anatomy. However, this approach animates the facial mesh in a muscle-like manner while ignoring the underlying anatomy of the human face.

3.2.2 Image Manipulation Techniques

These techniques are commonly used in motion pictures, such as "The Matrix Reloaded"[19]. Some common image-based methods are described below:

- *Image morphing*:
  This method involves taking a series of static photographs and interpolating between them. This is also called “blend-shape animation”. The static photographs serve as blend shapes or key facial expressions and are used to define a linear space of facial expressions. This model disregards the anatomy or mechanics of a human face and instead considers every expression to be a linear combination of these blend shapes. Joshi et al.[20] defined the blend shape model for n blend shapes as follows:
  $$\text{V}=\sum_{i=1}^{n}\alpha_i V_i$$
  where V is a vertex in the model and the scalars α_i are the blending weights and V_i is the location of the vertex in blend shape i.
  $$\alpha_i \geq 0 \text{ for all i and } \sum_{i=1}^{n}\alpha_i = 1$$
  FaceGen[21] modeller and SDK follow a similar approach using 3D blend shape models.

- *Vascular expressions*:
  When we express emotions, our skin color also changes. Kalra et al.[22] proposed a model in which they redefined emotion as a function of two signals in time, one for its intensity for spatial changes and the other for the color. At any instance of time, the state of

emotion et can be defined as:
et = k ft(s,c)
where k is a constant, s is the parameter for spatial intensity and c is the parameter for the color signal.

- *Texture manipulation*:
  Textures are often used to create realistic facial animation because they provide surface detail to each pixel which is absent from the surface geometry. Pighin et al.[23] created such facial animation models using multiple photographs.

4 Facial Action Coding System (FACS)

Facial Expression plays an important role in revealing the emotional state of a person. Expressions depict the intensions and emotions of a person sooner than they speak or even understand their feelings. During our everyday life, emotions come across through subtle changes in facial features, for instance raised inner eyebrows and lowered lip corners indicate sadness. Over the years, people have worked to develop a tool to study Facial Behavior. Researchers in the past have analysed information that observers are able to understand from the face, instead of measuring facial activity itself. Some of the contributors to current research on facial measurement are Birwhistel[24], Ekman, Friesen, and Tomkins[25] and Paul Ekman and W.V Friesen [26].

Paul Ekman and W.V Friesen developed the Facial Action Coding System for the measurement of facial movement[26].This system uses an anatomical approach where muscle movements describe facial expressions. The Facial Action code defines Action units (AUs) to describe expressions of the human face. The action of an isolated face muscle or group of face muscles is represented by an Action Unit(AU). Table 1 describes some of the single action units and their muscular base as shown in the Facial Action Code[27].

Table 1: Table listing Single Action Units as shown in Facial Action Code[27]

AU Number	FAC Name	—Muscular Basis
1.	Inner Brow Raiser	—Frontalis, Pars Medialis
2.	Outer Brow Raiser	—Frontalis, Pars Lateralis
4.	Brow Lowerer	—Depressor Glabellae; Depressor Supercilli; Corrugator
5.	Upper Lid Raiser	—Levator Palpebrae Superioris and Frontalis, Pars Medialis
6.	Upper Lid Raiser	—Levator Palpebrae Superioris
7.	Lid Tightener	—Orbicularis Oculi, Pars Palebralis
9.	Nose Wrinkler	—Levator Labii Superioris, Alaeque Nasi
10.	Upper Lid Raiser	—Levator Labii Superioris, Caput Infraorbitalis
11.	Nasolabial Fold Deepener	—Zygomatic Minor
12.	Lip Corner Puller	—Zygomatic Major
13.	Cheek Puffer	—Caninus
14.	Dimpler	—Buccinnator
14.	Dimpler	—Buccinnator
14.	Dimpler	—Buccinnator
15.	Lip Corner	—Depressor Triangularis
16.	Lower Lip Depressor	—Depressor Labii

17.	Chin Raiser	—Mentalis
18.	Lip Puckerer	—Incisivii Labii Superioris; Incisive Labii Inferioris
20.	Lip Stretcher	—Risorius
22.	Lip Funneler	—Orbicularis Oris
23.	Lip Tightner	—Orbicularis Oris
24.	Lip Pressor	—Orbicularis Oris
25.	Lips Part	—Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris
26.	Jaw Drop	—Masetter; Temporal and Internal Pterygoid Relaxed
27.	Mouth Stretch	—Pterygoids; Digastric
28.	Lip Suck	—Orbicularis Oris

Facial Expressions can be mapped to a single AU or a combination of AUs. When AUs occur in combination, they might be additive or non-additive. An additive combination has no visible effect on the constituent AUs. A non-additive combination has visible effect on the constituent AUs[28].Even though the total number of Action Units is small, there exist numerous combinations of Action units[29]. Below is an example of how an action Unit is described in the FAC Manual[27].

An Example of description of an Action Unit as given in the FAC for each Action Unit[27].

ACTION UNIT 15-Lip Corner Depressor

The muscle underlying AU 15 emerges from the side of the chin and runs upwards attaching to a point near the corner of the lip. In AU 15 the corners of the lips are pulled down. Study the anatomical drawings which show the location of the muscle underlying this AU.

(1) Pulls the corners of the lips down.

(2) Changes the shape of the lips so they are angled down at the corner, and usually somewhat stretched horizontally.

(3) Produces some pouching, bagging, or wrinkling of skin below the lips' corners, which may not be apparent unless the action is strong.

(4) May flatten or cause bulges to appear on the chin boss, may produce depression medially under the lower lip.

(5) If the nasolabial furrow* is permanently etched, it will deepen and may appear pulled down or lengthened. The photographs in FAC show both slight and strong versions of this Action Unit. Note that appearance change (3) is most apparent in the stronger versions. The photograph of 6+15 shows how the appearance changes due to 6 can add to those of 15. Study the film of AU 15.

How To Do 15

Pull your lip corners downwards. Be careful not to raise your lower lip at the same time-do not use AU 17. If you are unable to do this, place your fingers above the lip corners and push downwards, noting the changes in appearance. Now, try to hold this appearance when you take your fingers away.

When To Score Slight Versions of 15

Elongating the mouth is irrelevant, as it may be due to AU 20, AU 15, or AU 15+20.

(1) If the lip line is straight or slightly up in neutral face, then the lip corners must be pulled down at least slightly to score 15.

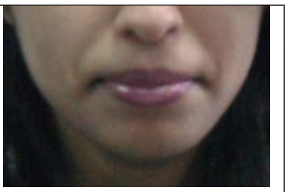
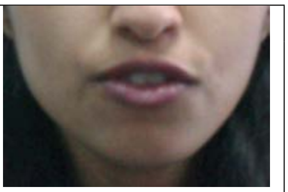
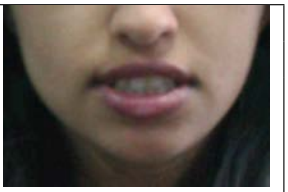
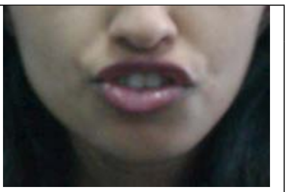
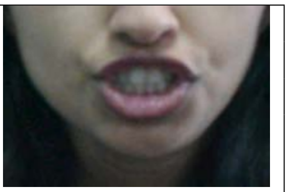
or (2) If lip line is slightly or barely down in neutral face, then the lip corners must be pulled down slightly more than neutral and not the result of AU 17 or AU 20.

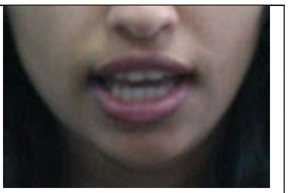
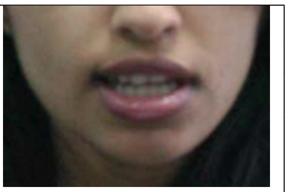
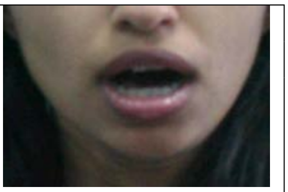
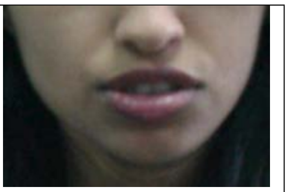
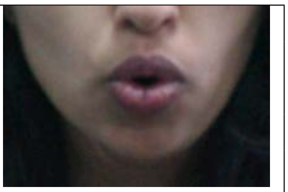
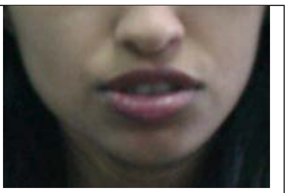
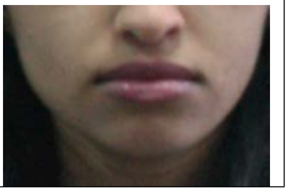
5 Phonemes and Visemes

Phonemes are the basic distinctive units of speech sound by which morphemes, words, and sentences are represented. They are different for every language.

A viseme is a generic facial image that can be used to describe a particular sound. A viseme is the representation of a phoneme in the visual domain. The English language has around 40 phonemes. These phonemes correspond to certain visemes, which can be can be mapped to 12 visemes[30] as shown below:

Table 2: Mapping phonemes to 12 visemes

Phoneme	Alphabetic spelling	Phonetic spelling	Viseme	
P	Pay	P EY	p	
B	Web	W EH B		
M	Lemon	L EH M AH N		
F	Four	F AO R	f	
V	Van	V AE N		
T	Raft	R AE F T	t	
D	Reed	R IY D		
S	Sit	S IH T		
Z	Arms	AA R M Z		
TH	Three	TH R IY		
DH	With	W IHDH		
DX	Forty	F AO R DX IY		
W	Swap	S W AA P	w	
R	Rest	R EH S T		
CH	Chair	CH EY R	ch	
JH	Cage	K EY JH		
SH	Blush	B L AH SH		
ZH	Asian	EY ZH AH N		

EH	Check	CH EH K	ey/iy	
EY	Take	T EY K		
AE	Bat	B AE T		
AW	Cow	K AW		
IY	Team	T IY M		
IH	Little	L IH DX AX L		
K	Coin	K OY N	k	
G	Peg	P EH G		
N	Night	N AY T		
L	Late	L EY T		
NG	Anything	EH N IY TH IH NG		
HH	Halt	HH AO L T		
Y	Yes	Y EH S		
AA	Barn	B AA R N	aa/ah	
AH	What	W AH T		
AX	About	AX B AW T		
AY	Type	T AY P		
ER	Church	CH ER CH	er	
AO	More	M AO R	ao/uh	
OW	Loan	L OW N		
OY	Hoist	H OY S T		
UH	Book	B UH K		
UW	Flew	F L UW		
IX	Action	AE K SH IX N	Silent	
SIL	-	The silence duration is variable but typically about 40 ms/ short pause	Sp	
SP	-	space btw words/characters		

6 Some Examples of Talking heads

We want to create a realistic 3D talking head which can be automated using text-to-speech and customised to resemble a particular person's face. Table 2 shows the comparison between the talking heads we reviewed. The Morpho app does not lip sync animation, while Bing Dictionary's talking head cannot be customised by the user. Sitepal produces decent 2D lip sync animations with a customisable head, but we cannot add any expressions. Using FaceShift, users can animate their 3D head or an animated character in real-time by standing in front of a Kinect sensor. However, FaceShift does not allow automation via text-to-speech.

Microsoft's Photo-real Talking Head[31] seems to be the closest to what we need. However, it is not readily available and adding expressions to it is still a challenge.

Table 3: Table comparing the talking heads we reviewed.

Application	2D/3D	Lip Sync	Expressions	Availability	Customisability	Automation
Morpho (IPhone App)	animated 2D model	Lips simply open and close; they are not in sync	Expressions included are "happy", "sad", "anger", "surprise", "fear" and "disgust". They seem realistic but the realism greatly depends on the accuracy of the location of facial features in the model.	This app is readily available on the Apple App Store.	The user can choose a picture from which the 2D model is created and some manual adjustments often need to be made.	The model creation is automatic but often requires some minor manual adjustments.
Bing Dictionary	2D; morphing across several images to animate	Lip sync seems very accurate	No expressions are explicitly used	It's web-based and reads out sentences to users of the Bing dictionary.	It cannot be customised at all.	The lips are animated by automatically morphing between images. The user cannot customise the face in any way.

SitePal	Animated 2D face model using a single photograph as input.	Lip sync is moderately accurate. Besides the few common visemes, the 2D model cannot portray a large number of other visemes.	Expressions depicted are quite realistic and can be easily inserted by clicking on the required expression. Available expressions are Happy, Sad, Smile, Angry, Disgust, Surprise and Thinking. Twitching of eyebrows and natural movement of the head add to the realistic features.	The animated agent can be integrated with other websites on purchase of the software.	The user can choose a picture from which the 2D model is created or select one of the available photo realistic and animated models. Options to change the models hairstyle, gender, background and accessories are also provided.	Model creation is automatic and uses a text to speech synthesizer to generate the voice.

FaceShift (EPFL)	3D model using Microsoft Kinect sensor	Lip sync is very accurate. However, needs to be controlled in real-time by a person standing in front of the sensor.	Expressions made by the person are mimicked by the model in real-time.	The software can be bought from the developer's website.	The talking head is highly customisable. The user can use his own head or an animated character's head.	The creation process requires the user to create a profile which needs extensive work. The final head is not automated using text-to-speech; it must instead be controlled in real-time by a person standing in front of a Kinect sensor.
CrazyTalk	Animated 2D model using a photograph of the complete human body as input.	Lip sync is not very accurate since this software focuses on the animation of the whole body and not only the face.	Expressions are poorly depicted since this software is used to animate the whole body of the model and the face.	This software has to be purchased to be used.	The user can choose a photo from which a complete model from the head to toe will be created or select one of the existing models with an animated head or with a photorealistic head.	Creation of the model is automatic once landmark points are placed on the selected 2D model or inputted image.
Photo-real Talking Head (Microsoft)	3D Talking head	High-quality lip sync animation has been achieved.	Expressions are still difficult to insert.	It has not been made publicly available yet. It is a research project at Microsoft Research Asia[31].	The user can create a 3D model of his own face. This project can give us a photorealistic talking head.	The talking head takes text as an input and converts it into speech animation photo-realistically.

7 Conclusion

This literature review shows that while facial modelling and animation techniques have greatly improved over the years, there are still several challenges and linguistic issues which impede the creation of a talking head which can both lip sync as well as display expressions. In this section, we will describe some of the issues we face and the approach we plan to follow for our project.

7.1 Linguistic Issues and Challenges

– *Intonation*
  Intonation depends upon three things: (i) how the sentence is uttered (for example, interrogative, declarative), (ii) attitude that the speaker wants to voluntarily show to the listener (for example, rudeness, sarcasm), and (iii) emotions that the speaker involuntarily shows to the listener (called paralanguage). The pitch, loudness, pitch contour, tempo (rate of speech) and pause[32] greatly depend on these three aspects. It is hard to mimic these in facial animation.

– *Coarticulation*
  A word is made up of several phonemes. Coarticulation refers to the changes in the articulation of a phoneme depending on preceding (backward coarticulation) and upcoming segments (forward coarticulation). Cohen and Massaro[33] illustrated that backward articulation can be seen when there is a difference in the articulation of a final consonant in a word depending on the preceding vowel, e.g. boot vs beet; while, an example of forward coarticulation is seen when the lips round up at the beginning of the word ”stew”.

– *Intensity of emotions*
  The intensity of emotions greatly affect the facial movements. For example, an angry or happy person will show more facial motion than a sad person[32].

– *Punctuators*
  Punctuators refers to the pauses made by the speaker while speaking. They are emotion-dependent and affect the rate of speech. For example, a person crying will pause more, while an angry person will just constantly scowl without pause.

– *Variation in blinking rate*
  The blinking of the eyes in synchronised with the articulation. It is also emotion-dependent. For example, during fear or excitement a person blinks more, while during shock and concentrated thought, blinking decreases[32].

7.2 Our Approach

Firstly, we need to create a custom 3D model of a person's face from his photograph. During the survey, we came across the FaceGen SDK[21] which seemed suitable for this task. The "Photo Fit" feature of the SDK will generate a 3D head from a person's photograph provided as input. It'll also generate the various morph targets or blend shapes for us. Each viseme and each expression of emotion corresponds to one blend shape. The intensity or weight of each blend ahape can be controlled using certain parameters.

We will be using the Microsoft Speech SDK to convert the input text into audio and to generate the phoneme segmentation file. The text provided as input is broken up into phonemes, which are then stored in the phoneme segmentation file along with their occurrence time and duration.

Expressions can be added as “tags” in the input text. Once the Speech SDK splits the input text into phoneme segments, we will parse through the phoneme segmentation file and insert the expressions as specified by the tags.

In our research, we will attempt to solve the problem of “coarticulation” and “variation in intensity of emotions”. In order to solve the problem of variation in intensity of emotions, we can either ask the user to explicitly specify the intensity as an attribute within the tags or we can experiment the Microsoft Kinect sensor to directly obtain these parameter from the users “actual” expressions. To deal with coarticulation, we will parse the modified phoneme segmentation file (containing phoneme and expression data) and apply forward and backward coarticulation rules to it.

In some cases, the expressions might overlap with the visemes; for example, a smile requires the lips to be stretched and closed, while the viseme “ao/uh” requires the lips to be rounded. Such cases will lead to problems. In our project, we will also attempt to find methods to resolve conflicts between expressions and visemes.

Hence, the propose problem statement is as follows:

1. To create a custom Talking Head, given a person’s photograph and to lip sync the head according to text provided as input.
2. To allow the user to add in expressions as tags into the input text and to produce these expressions in the Talking Head.
3. To solve the problem of coarticulation, variation in intensity of emotions and the overlap of expressions and visemes.

The proposed low fidelity plan is as follows:

1. Photograph of a person is passed through the FaceGen SDK’s Photo Fit feature in order to generate a 3D model of the head. The head may then be customised using the FaceGen customiser to add hair, spectacles and other superficial components of a human head.
2. The user is asked to input text along with tags for emotions. The Microsoft Speech SDK then converts this text into an audio file and a phoneme segmentation file (while ignoring the tags for expressions).
3. We pass the phoneme segmentation file through an “Emotion processor” which will add information about occurrence time, duration and type of these emotions. The problem pertaining to variation in the intensity of these emotions may be dealt with at this step.
4. Finally, we parse the phoneme segmentation file through a “coarticulation processor” which will apply forward and backward coarticulation rules to help with the problem of coarticulation. The problem of expression and viseme overlap may also be handled at this step.
5. In the end, we shall have a phoneme segmentation file containing information about which blend shape to show at what time and for how long. It would also hold information about the intensity or weight in time for each blend shape.

6. We will need to synchronise the animation according to the information in the phoneme segmentation file.

7. If adding expressions changes the overall timing of phonemes, we will need to alter the audio file as well. However, we will try to avoid that by allowing the expressions to occur either during pauses or along with the visemes. If occurrence time and duration of phonemes stays in sync with the audio, the audio file need not be changed.

The above problem statement and plan is merely a proposal. The final scope of the project will greatly depend upon the time constraints and availability of funds for the software.

References

[1] J. Ostermann, M. Beutnagel, A. Fischer, and Y. Wang, "Integration of talking heads and text-to-speech synthesizers for visual tts," 1998.

[2] W. Johnson, J. Rickel, and J. Lester, "Animated pedagogical agents: Face-to-face interaction in interactive learning environments," *International Journal of Artificial intelligence in education*, vol. 11, no. 1, pp. 47–78, 2000.

[3] G. Blonder and A. Milewski, "System and method for providing structured tours of hypertext files," Jan. 21 1998. EP Patent 0,820,024.

[4] E. André, T. Rist, and J. Müller, "Guiding the user through dynamically generated hypermedia presentations with a life-like character," in *Proceedings of the 3rd international conference on Intelligent user interfaces*, pp. 21–28, ACM, 1998.

[5] J. Ostermann and D. Millen, "Talking heads and synthetic speech: An architecture for supporting electronic commerce," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 1, pp. 71–74, IEEE, 2000.

[6] N. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *Journal of Speech, Language and Hearing Research*, vol. 12, no. 2, p. 423, 1969.

[7] E. Konstantinidis, M. Hitoglou-Antoniadou, A. Luneski, P. Bamidis, and M. Nikolaidou, "Using affective avatars and rich multimedia content for education of children with autism," in *Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments*, p. 58, ACM, 2009.

[8] K. Balci, "Xface: Mpeg-4 based open source toolkit for 3d facial animation," in *Proceedings of the working conference on Advanced visual interfaces*, pp. 399–402, ACM, 2004.

[9] J. Noh and U. Neumann, "A Survey of Facial Modeling and Animation Techniques," USC Technical Report 99-705, University of Southern California, 1998.

[10] N. Ersotelos and F. Dong, "Building highly realistic facial modeling and animation: a survey," *The Visual Computer*, vol. 24, no. 1, pp. 13–30, 2008.

[11] N. Magnenat-Thalmann, D. Thalmann, and J. Wiley, *Handbook of virtual humans*. Wiley, 2004.

[12] D. DeCarlo, D. Metaxas, and M. Stone, "An anthropometric face model using variational techniques," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pp. 67–74, ACM, 1998.

[13] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, ACM Press/Addison-Wesley Publishing Co., 1999.

[14] V. Blanz, K. Scherbaum, T. Vetter, and H. Seidel, "Exchanging faces in images," in *Computer Graphics Forum*, vol. 23, pp. 669–676, Wiley Online Library, 2004.

[15] F. I. Parke, "Computer generated animation of faces," in *Proceedings of the ACM annual conference - Volume 1*, ACM '72, (New York, NY, USA), pp. 451–457, ACM, 1972.

[16] S. M. Platt and N. I. Badler, “Animating facial expressions,” *SIGGRAPH Comput. Graph.*, vol. 15, pp. 245–252, Aug. 1981.

[17] K. Waters, “A muscle model for animation three-dimensional facial expression,” *SIGGRAPH Comput. Graph.*, vol. 21, pp. 17–24, Aug. 1987.

[18] D. Terzopoulos and K. Waters, “Physically-based facial modelling, analysis, and animation,” *The Journal of Visualization and Computer Animation*, vol. 1, no. 2, pp. 73–80, 1990.

[19] G. Borshukov and J. Lewis, “Realistic human face rendering for the matrix reloaded,” in *ACM Siggraph 2005 Courses*, p. 13, ACM, 2005.

[20] P. Joshi, W. Tien, M. Desbrun, and F. Pighin, “Learning controls for blend shape based realistic facial animation,” in *ACM SIGGRAPH 2005 Courses*, p. 8, ACM, 2005.

[21] S. Inversions, “Facegen.”

[22] P. Kalra and N. Magnenat-Thalmann, “Modeling of vascular expressions in facial animation,” in *Computer Animation'94., Proceedings of*, pp. 50–58, IEEE, 1994.

[23] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin, “Synthesizing realistic facial expressions from photographs,” in *ACM SIGGRAPH 2006 Courses*, p. 19, ACM, 2006.

[24] R. Birdwhistell, *Kinesics and context: Essays on body motion communication*, vol. 2. University of Pennsylvania press, 1970.

[25] P. Ekman, W. Friesen, and S. Tomkins, “Facial affect scoring technique: A first validity study,” *Semiotica*, vol. 3, no. 1, pp. 37–58, 1971.

[26] P. Ekman and W. Friesen, “Facial action coding system: A technique for the measurement of facial movement. palo alto,” *CA: Consulting Psychologists Press. Ellsworth, PC, & Smith, CA (1988). From appraisal to emotion: Differences among unpleasant feelings. Motivation and Emotion*, vol. 12, pp. 271–302, 1978.

[27] P. Ekman and W. Friesen, “Measuring facial movement,” *Journal of Nonverbal Behavior*, vol. 1, no. 1, pp. 56–75, 1976.

[28] Y. Tian, T. Kanade, and J. Cohn, “Recognizing action units for facial expression analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 97–115, 2001.

[29] K. Scherer and P. Ekman, “Methodological issues in studying nonverbal behavior,” *Handbook of methods in nonverbal behavior research*, pp. 1–44, 1982.

[30] Annosoft, “Phoneme mapping with 12 visemes.” http://www.annosoft.com/docs/Visemes12.html, 2008.

[31] L. Wang, W. Han, and F. Soong, “High quality lip-sync animation for 3d photo-realistic talking head,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4529–4532, IEEE, 2012.

[32] C. Pelachaud, M. Steedman, and N. Badler, “Linguistic issues in facial animation,” *Center for Human Modeling and Simulation*, p. 69, 1991.

[33] M. Cohen, D. Massaro, *et al.*, “Modeling coarticulation in synthetic visual speech,” *Models and techniques in computer animation*, vol. 92, 1993.