Realistic Speech Animation of Synthetic Faces

Barış Uz Uğur Güdükbay Bülent Özgüç
Bilkent University
Department of Computer Engineering and Information Science
Bilkent, 06533 Ankara, Turkey
E-mail: {baris, gudukbay}@cs.bilkent.edu.tr, ozguc@bilkent.edu.tr

Abstract

In this study, we combined physically-based modeling and parameterization to generate realistic speech animation on synthetic faces. We used physically-based modeling for muscles. Muscles are modeled as forces deforming the mesh of polygons. Parameterization technique is used for generating mouth shapes for speech animation. Each meaningful part of a text, which is a letter in our case, corresponds to a specific mouth shape and the mouth shape is generated by setting a set of parameters used for representing the muscles and jaw rotation. We also developed a mechanism to generate and synchronize facial expressions while speaking. Some tags specifying the facial expressions are inserted into the input text together with the degree of the expression. In this way, the facial expression with the specified degree is generated and synchronized with speech animation.

Key words: facial animation, speech animation, muscle-based, physically-based, facial expression.

1. Introduction

Facial animation has attracted many researchers during the last decade since the face is very important for identifying people and has a very complex structure. Researchers worked extensively on the realistic animation of faces.

Animating a face is generally understood as modeling speech as well as the facial expressions, such as fear, anger, surprise, disgust, happiness and sadness. It is very difficult to define and construct a model capable of performing realistic face motions. It is even more complicated to construct a model which is both realistic and efficient enough to run at interactive rates. This difficulty is mainly caused by the complexity of the facial anatomy.

In this study, we developed a system for realistic animation of speech on a synthetic face according to a given text. For this purpose, we modified the facial animation software developed by Waters [9]. He modeled the face muscles as the forces deforming a group of vertices of the face model's polygon mesh. We added some *pseudomuscles* to emulate the basic muscle actions for the mouth [9]. These actions are essentially due to the contractions of the orbicularis oris. To increase realism, we added teeth and eyes to the model.

We developed a mechanism to generate facial expressions while speaking. Tags specifying the facial expressions related to a word or sentence are inserted to the input text together with the degree of the expression. In this way, the facial expression with specified degree is generated and synchronized with speech animation.

2. Previous Work

Since speech animation is a part of facial animation, the techniques used for facial modeling and animation provide a basis for speech animation.

2.1. Facial Animation

Previous studies for facial modeling and animation started with the seminal work of Parke [6]. He used keyframing techniques to animate the face. Since each key frame must be completely specified to animate the face, simple keyframing cannot be easily used for three-dimensional (3D) facial animation. Parametric systems have emerged as a result of this [8]. A parametric facial animation system defines a set of parameters for the face. These are mainly the expression parameters for different parts of the face, such as mouth and eyes and the conformation parameters that apply globally to the whole face. The major parameters for the mouth are jaw rotation (for mouth opening), width of the mouth, etc. The major parameters for the eyes are pupil dilation, eyelid opening, eyebrow position and shape. The conformation parameters are skin color, aspect ratio of the face, etc. Each expression parameter effects a set of vertices of the face model. In this way, key frames of an animation can be defined easily. A "facial state" can be created by

altering the expression parameters that move the vertices to desired new positions.

The disadvantage of the parametric systems is that they cannot easily blend facial expressions since each parameter effects a disjoint set of vertices in the face model. This limitation led to the development of structure based facial models which are based on the anatomy of the face [11]. However, all these models do not take into account the fact that the face is not only a geometric model but a complex biomechanical system. Physically-based face models have emerged to overcome these limitations. Terzopoulos and Waters model the face in a layered fashion and incorporate an anatomically based muscle model with a physically based layered tissue model. They used a trilayer facial tissue concept which is modeled as a trilayer mesh of vertices connected by springs [12]. The trilayer mesh propagates the tissue deformations from the innermost layer (muscle actuators inserted into the innermost later) to the face skin with the help of springs connecting the vertices. Anatomically based muscle models are first used by Waters [13] to animate major facial expressions.

2.2. Speech Animation

Speech animation has two parts: first, the animation of a synthetic face according to a given text, and second, synchronizing speech with facial animation.

Generating Speaking Face Models

To generate a convincing speaking face model one has to model a variety of mouth and lip postures and interpolate these postures in a realistic way. The first attempt to generate speech animation of synthetic faces is by Parke [7]. Pearce et al. [10] used a parametric approach to animate speech. There are also image-based approaches for speech animation. Watson et al. [16] developed a morphing algorithm for interpolating phoneme images to simulate speech.

Waters and Frisbie [14] described a coordinated muscle model to model complex muscle interactions around the mouth. In this way, they produce natural-looking speech on a facial image but their muscle model is two-dimensional.

Basu [1] developed a 3D model of human lips and a framework for training it from real data. Although his work is mainly for reconstruction of lip shapes from real data, it can also be used for lip shape synthesis for speech animation.

Synchronizing Speech with Facial Animation

Animating a synthetic face synchronized with a given audio requires lip synchronization. Keyframed, parametric and muscle-based systems are used with non-automated techniques to achieve lip-synchronization. This process requires the specification of keyframes, essentially jaw and lip positions together with the timing information. Parke used a parametric approach to demonstrate this synchronization [7]. Pearce et al. [10] used a rule-based speech synthesizer to incorporate speech parameters for a 3D facial model. They record the speech sequence generated to the audio channel and the frames of a moving face model generated using the speech parameters to the video channel. They play the sequence to animate the face. However, the non-automated process of lip synchronization is not flexible since changing the audio requires the whole process to be repeated [15].

"DecFace" is an automatic lip-synchronization algorithm for synthetic faces [15]. DecFace has the ability to generate speech and graphics at real-time rates, where the audio and the graphics are tightly coupled to generate expressive facial characters. It synchronizes the synthesized speech samples generated from a given text and the motion of the face model. To achieve this, it computes a mouth shape for each phoneme and interpolates the mouth shapes using cosine interpolation techniques. The phonemes are determined by querying the audio server for each phoneme so that the correct mouth shape is computed synchronously for each phoneme.

Another important problem is the generation of facial expressions due to emotions. Kalra et al. [4] developed a facial animation system based on layered abstractions. They decompose the problem into five layers. The higher layers are more abstract and specify "what to do" and the lower layers describe "how to do it". The highest layer allows abstract manipulation of the animated entities. During this process, speech is synchronized with the eye motion and emotions by using a general and extensible synchronization mechanism provided by a high-level language they developed. At the lowest level, they use abstract muscle action procedures [5], which is a *pseudomuscle-based* technique that develops models with a few control parameters emulating the basic facial muscle actions.

Cassell et al. [2] also developed a system for automatically generating and animating conversations between multiple human-like agents with appropriate and synchronized speech, intonation, facial expressions and hand gestures. They derive facial expression, head and eye motion from spoken input automatically.

3. Overview of Facial Muscle Anatomy

There are three types of muscles in face [13]:

1. *Sphincter*, like the one around the mouth (squeezing).
2. *Linear*, like the one used while smiling (pulling).
3. *Sheet*, like the one used when raising the eyebrow.

The musculature of the lower face differs from the musculature of the other parts of the human body. Most of the

other muscles of the human body originate and insert into the bone, but muscles in the lower face originate and insert into other muscles. The principal muscles around the mouth are the orbicularis oris, the buccinator, the levator labii superioris alaeque nasi, the levator labii superioris, the zygomaticus major and minor, the levator anguli oris, the anguli oris, the depressor labii inferioris, the risorius, and the mentalis. The muscles closing the lips are the orbicularis oris and the incisive muscles. The muscles opening the lips are known as radial muscles, which are divided into the radial muscles of upper and lower lips, superficial and deep.

The shape of the lips is mostly determined by the orbicularis oris and affected by some of the other muscles around the mouth. The orbicularis oris consists in part of muscle fibers derived from the other facial muscles that converge to the mouth. For example, the buccinator forms the deep layer of the orbicularis oris. A detailed explanation of the facial muscles and their actions can be found in [9].

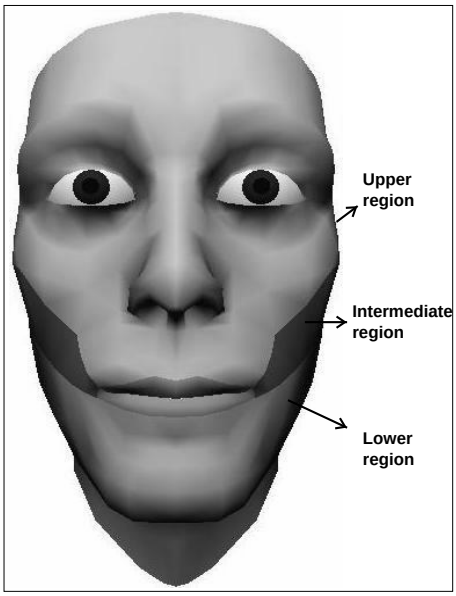

Figure 1. Regions of face.

4. Modeling

Our synthetic face is based on the model developed by Waters [9]. We modified this face model to animate speech in a realistic manner. The modifications are explained below.

4.1. Face Model

The face model consists of 888 triangles and is symmetric about the vertical plane cutting through the center of the face (sagittal plane). The vertices of the upper lip are duplicated to distinguish lower lip vertices from upper lip vertices.

The model consists of only one layer which is enough for speech animation. Since basic aim is to create a realistic speech animation, we did not elaborate on realistic modeling of face skin and tissue, such as modeling wrinkles around the mouth, nose and above the nose and eye. Our muscles and skin are in the same layer, although some of the muscles are going towards the outside of the skin.

During speech animation, we need an abstraction on the face. Speech mainly occurs in the lower face. Our model contains 240 polygons in the lower face and 610 polygons in the upper face. Remaining polygons belong to the intermediate region that connects upper and lower regions as seen in Figure 1. As a result, when the lower face is deformed, or the jaw is rotated, these polygons provide the continuity of the face. Rather than classifying polygons as lower or upper, we classified their vertices as lower and upper. Upper and lower face are affected by muscle actions. In addition, lower face is affected by jaw rotation.

Our muscles may have the tags UPPER, LOWER or BOTH, which means that the muscle action affects the upper face vertices, lower face vertices, or both, respectively. We also have an additional motion type JAWROT, which affects lower face and lower teeth. Each vertex also has a tag specifying the muscle actions and motions affecting the vertex. If this tag is NONE, the vertex is not affected by any muscle action and motion (for the upper teeth). The relationship between muscles (and jaw rotation) and the face vertices is given in Table 1.

Motion or Muscle Tag	Vertex Tag				
	UPPER	LOWER	BOTH	NONE	JAWROT
UPPER	+	-	+	-	-
LOWER	-	+	+	-	-
BOTH	+	+	+	-	-
JAWROT	-	+	+	-	+

Table 1. Relationship between a muscle (motion) and a vertex.

4.2. Facial Muscles

Types of Facial Muscles

In our model, there are 34 muscles. We used four linear muscles to represent the orbicularis oris and the other muscles are represented as pairs of muscles which have left and right components. The facial muscle structure is shown in Figure 2. The major muscles that affect the speech are explained in the sequel.

1. *Orbicularis oris:* This muscle has the most significant role in composing the shape of the mouth during

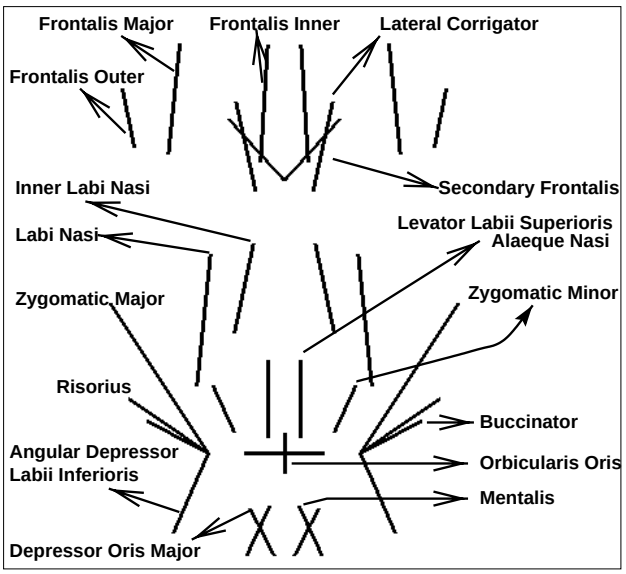

Figure 2. Facial muscles in our model.

speech. Especially, when we say "O" or "a(h)", the orbicularis oris determines the shape of the mouth. It has also very important position since the orbicularis oris is where all of the other muscles around the mouth merge into.

2. *Mentalis, buccinator, depressor anguli oris major, depressor labii inferioris:* These muscles are placed in the lower face. They control lower lip and lower face. These muscles play a great role in speech together with the orbicularis oris and jaw rotation.
3. *Zygomatic minor, levator labii superioris alaeque nasi, levator labii superioris:* These muscles are placed in the upper face region. They are rarely used and activated during speech.
4. *Risorius, zygomatic major:* These muscles are located around the cheeks. They play an important role in simulating expressions, rather than speech.

Modeling of Facial Muscles

The skin is interpreted as an elastic material. Muscles are modeled as forces deforming the skin. The skin is similar to other elastic materials in that it is deformed under a force. However, it is not perfectly elastic since after a tension its visco elastic behavior prevents the skin from deforming any more. The skin is thought as a mesh of springs and it is deformed under a tension of a muscle. There are two main types of muscles in our face model.

1. *Linear*, like the one used while smiling (pulling).
2. *Sphincter*, like the one around the mouth (squeezing).

A linear muscle deforms the mesh like a force. We can model a linear muscle with the following parameters:

- *Influence zone:* Each muscle has an influence zone in which the vertices are mostly affected. This varies from one muscle to the other and it is typically between 35 and 65 degrees.
- *Influence start (fall start):* Each muscle's influence will appear after a tension.
- *Influence end (fall end):* Each muscle has a limit to be tensioned. After this limit, skin resists deformation.
- *Contraction value:* Muscle tension.

Changing these parameters alter the effect of muscle action. We used the formulation in [13] to model muscles. Since a linear muscle is modeled as a force, its direction is also important and is defined by the direction of muscle vector. Starting point of the vector is never repositioned and it is the originating point of the muscle. A muscle pulls or pushes the vertices along this muscle vector (Figure 3).

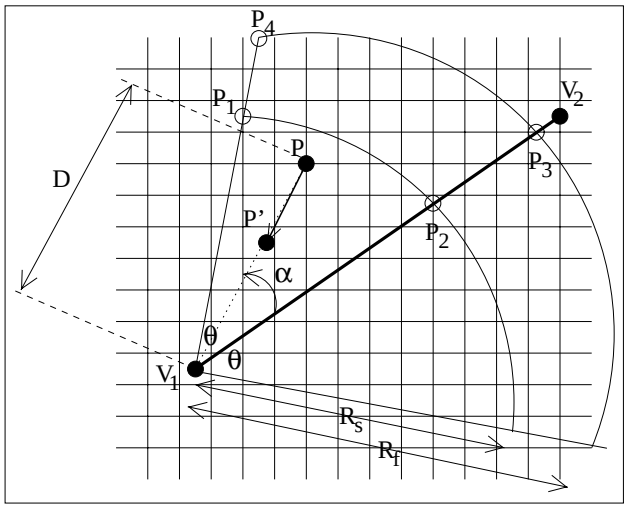

Figure 3. Parameters of a muscle.

In Figure 3,

- P is a point in the mesh,
- P' is its new position after the muscle is pulled along the V_1V_2,
- R_s and R_f represents muscle fall start and fall finish radii, respectively,
- θ represents the maximum zone of influence, typically between 35 and 65 degrees,
- D is the distance of P from muscle head and
- α is the angular displacement.

Note that V_1 and V_2 are not necessarily placed as a node of mesh because muscles are thought as forces which can be anywhere in the space. This muscle representation is for 2D but it can easily be adapted to 3D by applying the same rules to the third dimension.

If P is in the region of $V_1P_3P_4$, new position $P' = P + k\ a\ r\ \frac{PV_1}{\|PV_1\|}$, where

k is the muscle spring constant,

$a = cos(\alpha)$ and

$$r = \begin{cases} cos((\frac{1-D}{R_s})\frac{\pi}{2}) & \text{if } P \text{ in } (V_1P_1P_2) \\ cos((\frac{D-R_s}{R_f-R_s})\frac{\pi}{2}) & \text{if } P \text{ in } (P_1P_2P_3P_4) \end{cases}$$

This holds for a linear muscle but not for a sphincter muscle. However, the orbicularis oris is a sphincter muscle that is elliptical. We represented the orbicularis oris as a group of 4 linear muscles. Vertical parts have 140 degrees of influence zone where horizontal ones have 40 degrees (Figure 4). This creates the desired elliptic effect together with the other facial muscles and is very practical to implement.

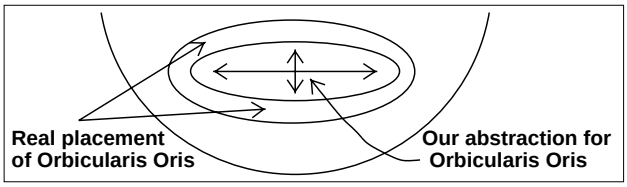

Figure 4. Abstraction of orbicularis oris.

4.3. Jaw Rotation

In facial modeling and animation, it is very important to rotate jaw, which makes the animation convincing. Human jaw is composed of two parts: upper and lower jaw. The movable part is the lower jaw. This motion is essentially a rotation around an axis connecting the two ends of the jaw bones [3].

In our model, the jaw contains the vertices with tag LOWER and JAWROT (lower teeth). Since the lower teeth are also attached to the lower jaw, they will be affected by jaw rotation.

4.4. Eyes end Teeth

The eyes and teeth of our model are also defined as meshes of polygons. An eye can rotate about $x - axis$, $y - axis$ or both. It is not possible for an eye to rotate about $z - axis$. One eye cannot rotate independent from the other. Two eyes cannot be looking at different directions.

Teeth are pure polygons and affected only by jaw rotation. Upper teeth are not affected by an action. However, lower teeth move with the lower jaw.

5. Speech Animation

Turkish is a syllable based language. The vowels in a word determine the lip motions and lip shape changes. In Turkish, the vowels are classified as shown in Table 2.

	Low	High
Non-round	a e	ı i
Round	o ö	u ü

Table 2. Classification of vowels in Turkish.

Although Turkish is a syllable-based language, we modeled speech based on letters. The mouth and lip shapes are determined according to letters since each letter is associated with a sound. To do this, each letter is associated with a mouth shape by defining the parameter values for muscles around the mouth and jaw rotation. For example, “de” and “do” syllables have different lip shapes, due to the characteristics of letters “e” and “o”. For “do”, mouth is round, but for “de”, mouth is flat. So, it is meaningful to define “d”, “e” and “o” *sounds* individually and then we can compose a syllable, like “de”, easily. By using the classification in Table 2, the vowels can be grouped into four since the pairs “a” and “e”, “o” and “ö”, “ı” and “i”, and “u” and “ü” have similar mouth shapes. In addition, some consonants have similar mouth shape characteristics. We can, for example, use the same mouth shape for letters “b”, “m” and “p” consonants. Consequently, we can define thousands of words by using approximately 20 *sound* definitions. As a result, we use sounds to compose syllables and syllables to make words, as in Turkish.

5.1. Keyframing Based on Parameterization

To generate animation of a speaking face model, we use keyframing based on parameters of the muscles around the mouth and jaw rotation. Each keyframe of an animation sequence includes a properly positioned mouth and face shape according to the current settings of the expression and the letter to be spoken. In Turkish, words are pronounced by strict rules. The written form dictates the pronunciation and there are no exceptions. Hence, the database for mouth shapes can easily be based on letters. Each entry in the database will contain the following:

1. *Letter* whose parameters will be defined. This is the *key* field of an entry in the database.
2. *Muscle contraction values* to determine which muscles are active while pronouncing that letter.
3. *Jaw rotation angle*, necessary for some letters.

The system accepts the text to be spoken as input. Then, it creates necessary keyframes for each letter. It uses cosine

interpolation scheme to generate in-betweens. The framework for the animation system is shown in Figure 5.

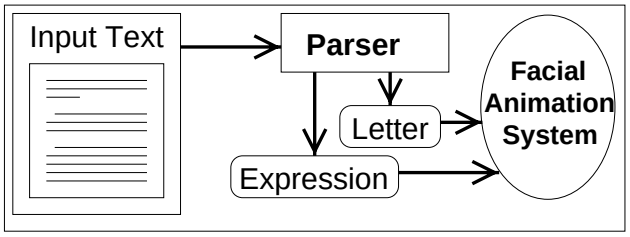

Figure 5. The animation system.

5.2. Cosine Interpolation

Cosine interpolation is suitable for approximating the viscoelastic behavior of the skin [9]. A cosine function is used to calculate the time of displaying an inbetween. This scheme is also known as acceleration and fits well to the facial animation. The display time for an inbetween frame i is given by the formula

$$tB_i \;=\; t_1 + \Delta t(1 - cos(\theta))$$

where

$$\theta = \frac{i\pi}{2(n+1)}, \;\; 0 < \theta < \frac{\pi}{2}, \;\; \text{and} \;\; i = 1, 2, 3, \ldots, n.$$

6 Synchronizing Speech with Expressions

Since we are interested in creating realistic speech animation, we should be able to generate a sequence of expressions produced during speech and synchronize them with speech. Facial expressions caused by the emotions can be determined from the text in two ways:

Guessing from the Text

By using punctuation, postfixes and keywords, we could guess some information about the expressions. However, this never gives a unique result as in the following example. Both sentences finish with "!" but feelings are different:

`Come here !` *anger*

`Oh, it's very nice to see you !` *surprise*

Similarly, keywords do not give any clear idea about the sentence:

`When will he come ?`
A question, voice will raise

`I don't know when he will come.`
A negative sentence, with a question word

It is even more difficult, if not impossible, to determine the different meanings of a single word which may cause different expressions to be generated. Hence, it is better to insert the information necessary to determine the facial expressions into the text manually.

By Inserting Tags into the Text

To guide the generation of facial expressions during speech, some tags could be inserted into the input text specifying the expressions, which is what we did. Expression intensities and the duration of an expression are also specified by adding the necessary parameters to the tags. Since the intensities of the facial expressions are not constant during speech, we define minimum and maximum intensity levels for each facial expression and linearly interpolate between these two values for the intermediate intensity levels. The tags for the facial expressions have the following format:

`\b{expr level}` starts an expression `expr` of degree `level`. If this expression is set before, `level` is used to increase the degree of the expression.

`\e{expr level}` ends or decreases the degree of an expression by `level`. If `level` is -1, expression is removed from the face.

We can also blend facial expressions. For example, assume that input text contains tags of two expressions for a sentence, one for raising eyebrows and one for smiling. These two expressions are combined together to give the combined facial expression of smiling and surprise. The following input text generates this effect. The Turkish letter "ı" is denoted by "I".

```
\b{SMILE 3} merhaba, nasIlsIn? \e{SMILE 1}
\b{EYEBROW 4} yeni araban nasIl? \e{SMILE -1}
```

In this example, there are 3 parts:

1. `SMILE 3`: Sets the face using expression `SMILE` with degree 3. "merhaba, nasIlsIn" is said with this expression.
2. `SMILE 1`: Decreases the degree of `SMILE` by 1. `EYEBROW 4`: Sets the face using expression `EYEBROW` with degree 4. Now, the face is both smiling and the eyebrow is raised to a degree of 4. "yeni araban nasIl" is said with this expression.
3. `SMILE -1`: Removes the `SMILE` expression.

At the end of the talk, eyebrows are left raised. `SMILE` expression is completely removed from the face. The algorithm for speech animation is given in Figure 6. Still frames from an animation sequence and the timings of the facial expressions are shown in Figure 7.

```
While not all of the text is processed {
  Read a character
  If a tag is beginning {
  /* "\" is read */
    Read tag
    /* name and degree of expression */
    If degree is -1,
       Remove expression from the face
    else
       Set face according to expression
           with specified degree
  }
  If a valid character {
  /* a letter or a punctuation mark */
    If this is the first character to say
       Set face using current expression
           and letter settings
       Display face
    else
       for each in-between
           Calculate vertex coordinates
               using cosine interpolation
           Display face
    Store vertex coords for future reference
  }
}
```

Figure 6. The algorithm for speech animation.

7. Conclusions and Future Work

In this work, we focused on realistic animation of speech on a synthetic face according to a given text. Our work is based on Turkish. In Turkish, we speak what we wrote. Each letter corresponds to a specific mouth posture. So, we need a mouth posture for each of the 29 letters in Turkish. This is not the case in English. English is based on small structures, called *phoneme*s. In English, we need 45 phonemes to pronounce all of the words. There are 18 visually distinct mouth postures for English and some of them are not used for Turkish, like the one for phoneme "th". We also developed a mechanism for generating facial expressions caused by the emotions during speech animation. This is done by inserting some tags specifying the types and degrees of the facial expressions into the input text.

This work can be extended by adding the following:

i. Texture mapping could be implemented to increase realism of animations.

ii. Tongue could be added to the face model. It has an important role when saying letters 'd', 'l', 'n' and 't'.

iii. Hair could be added by using texture mapping.

Acknowledgment

We thank Keith Waters for giving permission to use his facial software without which this research cannot be done.

References

[1] Basu, S, "A Three Dimensional Model of Human Lip Motion", M.Sc. Thesis, Dept. of Electrical Eng. and Computer Science, Massachusetts Institute of Technology, 1997.

[2] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M., "Animated Conversation: Rule-Based Generation of Facial Expression Gesture and Spoken Intonation for Multiple Conversational Agents", *ACM Computer Graphics (Proc. SIGGRAPH)*, pp. 413-420, July 1994.

[3] Güdükbay, U., "A Movable Jaw Model for the Human Face", *Computers & Graphics*, Vol. 21, No. 5, pp. 549-554, 1997.

[4] Kalra, P., Mangili, A., Magnenat-Thalmann, N., Thalmann, D., "SMILE: A Multilayered Facial Animation System", *IFIP WG 5.10*, pp. 189-198, Tokyo, 1991.

[5] Magnenat-Thalmann, N., Primeau, N. E., Thalmann, D., "Abstract Muscle Action Procedures for Human Face Animation", *Visual Computer*, Vol. 3, No. 5, 290-297, 1998.

[6] Parke, F.I., "Computer Generated Animation of Faces", *Proc. of ACM National Conference*, Vol. 1, pp. 451-457, 1972.

[7] Parke, F.I., "A Model for Human Faces that Allows Speech Synchronized Animation", *Computers & Graphics*, Vol. 1, No. 1, pp. 1-4, 1975.

[8] Parke, F.I., "Parameterized Models for Facial Animation", *IEEE CG & A*, Vol. 2, No. 9, pp. 61-70, November 1982.

[9] Parke, F.I., and Waters, K., *Computer Facial Animation*, A. K. Peters, Wellesley, MA, 1997.

[10] Pearce, A., Wyvill, B., Wyvill, G., Hill, D., "Speech and Expression: A Computer Solution to Face Animation", *Proc. Graphics Interface '86*, pp. 136-140, 1986.

[11] Platt, S.M., "A Structural Model of the Human Face", Ph.D. Thesis, University of Pennsylvania, Dept. of Computer and Information Science, 1985.

[12] Terzopoulos, D., Waters, K., "Physically-based Facial Modeling, Analysis, and Animation", *The Journal of Visualization and Computer Animation*, Vol. 1, pp. 73-80, 1990.

[13] Waters, K., "A Muscle Model for Animating Three-Dimensional Facial Expression", *ACM Computer Graphics (Proc. SIGGRAPH)*, Vol. 21, no. 4, pp. 17-24, July 1987.

[14] Waters, K., Frisbie, J., "A Coordinated Muscle Model for Speech Animation", *Proc. Graphics Interface '95*, pp. 163-170, May 1995.

[15] Waters, K., Levergood, T.M., "DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces", Tech. Report, CRL 93/4, DEC *Cambridge Research Lab.*, 1993.

[16] Watson, S.H., Wright, J.R., Scott, K.C., Kagels, D.S., Freda D., Hussey K. J., "An Advanced Morphing Algorithm for Interpolating Phoneme Images to Simulate Speech", Tech. Report, Jet Propulsion Lab., California Institute of Technology, 1996.

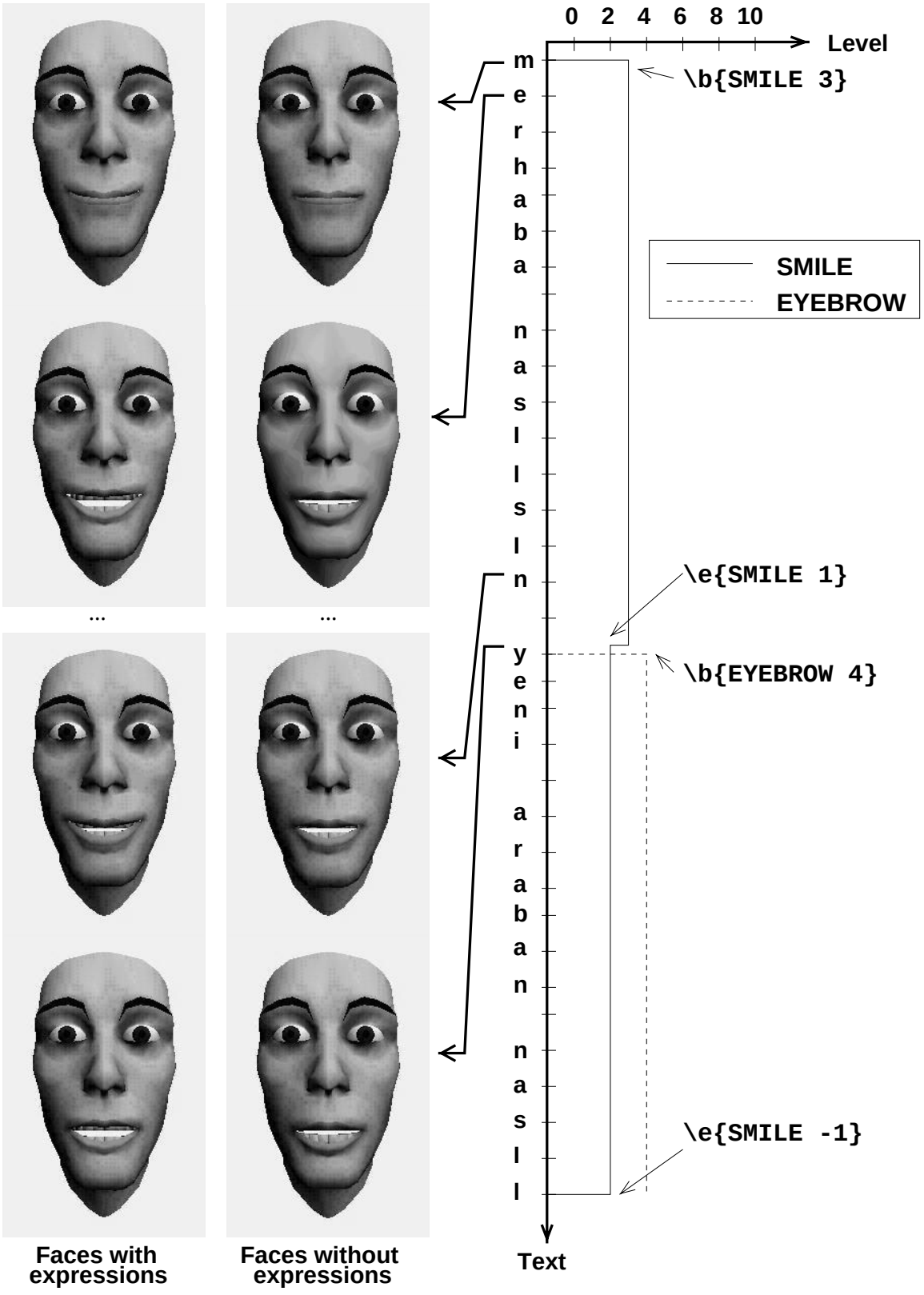

Figure 7. Still frames from an animation and the timings of expressions.