

LIPS MODELING THROUGH CONICAL CURVES

Camilo Oscar Girardelli Baptista¹, Carlos Alberto Ynoguti²

Abstract — *Talking Heads are artificial entities that model the human faces behavior. They are used in a wide range of areas like entertainment, man-machine interface, help for handicapped people, creation of avatars in virtual worlds, etc. This article reports some initial results of our investigation in this exciting field of knowledge. Instead of building sophisticated 3D models with muscle simulations, rendering effects, etc., a simpler 2D model was chosen. This simpler model is convenient because it allow us to keep most of the parameters under control. The main goal of this research is to animate a 2D face synchronized with an artificial text-to-speech synthesis system. In this article some initial results regarding lips parametric modeling by conical curves are presented.*

Index Terms — *Talking heads, lips modeling, multimodal user interfaces.*

INTRODUCTION

The world is experimenting important transformations in the communications area. At the beginning of the last century, the radio allowed the news to be given almost at the same time they were taking place. The advent of the television brought, together with the auditive information, also the visual information. With the Internet and the cellular telephony, the world of communications entered the multimedia age, with the simultaneous transmission of several kinds of media: text, images, audio, films, etc.

However, these great technological advances require the users to have knowledge of how to manipulate them and, mainly for the 3rd world countries, contributes to make the social segregation worse in our world, a fact that is called technological segregation. One way to minimize this harmful effect could be the creation of friendly user interfaces for these high technology devices.

In the man-machine interfaces, audio, text, pictures and video are integrated in order to make it available a more natural interface between the human user and the high technology devices. One way this could be done is by the use of an artificial character with the ability to receive commands from the user and respond in an natural and intelligent way [2]. This character could be implemented as an animated talking head with capability to perform text-to-speech conversion with synchronized face motion.

Other applications for talking heads are tools for hearing

aid impaired person communication. These people make use of visual information (lip reading) to communicate to other people. Because of this, the characters should reproduce the facial movements (lips, tongue, jaw, etc.).

Normal people may also derive some benefits from this technology, mainly in noisy environments. The ideal would be the use of real people for this task, but it's not a practical solution. Although not with the same performance, artificial faces are proven to improve the intelligibility of voice communication for either natural or synthetic speech.

Another application for this technology is the creation and development of avatars. Avatars are entities that represent people in virtual communities. In fact, there are nowadays virtual communities in the Internet that reproduce all the aspects of the real life, even with their own currency, that must be acquired via credit cards in the real world. In these communities you can buy a dog, a TV or whatever you want, date with other avatars, make friends, marry and so on.

Artificial characters are becoming famous in a wide range of areas such as TV, cinema, toys, games and so on. The success of Lara Croft, Buzz Lightyear and many others are the proof of this phenomena. The realism of these simulations highly depends on the incorporation of physiological knowledge in the computer models [3].

THE MCGURK EFFECT

In face-to-face communication, the speech perception is both visual and auditive. The visual part is particularly effective when the audio is degraded by noise, limited bandwidth or auditive impairment. However, conflicting visual and auditive information can lead to a distortion in speech perception, called McGurk effect.

"The most striking demonstration of the combined (bimodal) nature of speech understanding appeared by accident. Harry McGurk, a senior developmental psychologist at the University of Surrey in England, and his research assistant John MacDonald were studying how infants perceive speech during different periods of development. For example, they placed a videotape of a mother talking in one location while the sound of her voice played in another. For some reason, they asked their recording technician to create a videotape with the audio syllable "ba" dubbed onto a visual "ga." When they played the tape, McGurk and McDonald perceived "da." Confusion

¹ Camilo Oscar Girardelli Baptista, INATEL- Instituto Nacional de Telecomunicações, Av. João de Camargo, 510,37540-000, Santa Rita do Sapucaí, MG, Brazil, camilo-oscar@inatel.br

² Carlos Alberto Ynoguti, INATEL- Instituto Nacional de Telecomunicações, Av. João de Camargo, 510,37540-000, Santa Rita do Sapucaí, MG, Brazil, ynoguti@inatel.br

reigned until they realized that "da" resulted from a quirk in human perception, not an error on the technician's part. After testing children and adults with the dubbed tape, the psychologists reported this phenomenon in a 1976 paper humorously titled "Hearing Lips and Seeing Voices," a landmark in the field of human sensory integration. This audio-visual illusion has become known as the McGurk effect or McGurk illusion.[1].

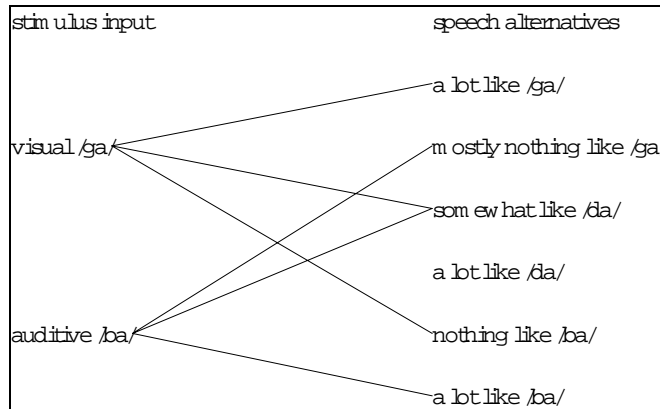


FIGURE 1
ILLUSTRATION OF THE MCGURK ILLUSION (AFTER [9]).

FACIAL MODELS

The main methods used for computer based facial animation are [6]:

1. **Concatenative Model:** a library of faces is created, storing a picture for each situation, and the animation is created by the interpolation between the images.
2. **Parametric Model:** consists in a parametric representation of some points of the face, so the animation is achieved by the temporal variation of these parameters. The advantage of this approach is the simplicity, requiring low memory sizes for the parameter storage.
3. **Muscular Model:** the skin and muscles action properties are simulated using an elastic mesh and forces, and therefore this model have a great potential to realistic reproduce the face movements. However, the high computational cost and the difficulty to manage all the parameters involved make this model very hard to implement.

WISEMES

“A viseme is a generic facial image that can be used to describe a particular sound. A viseme is the visual equivalent of a *phonemes* or unit of sound in spoken language. Using visemes, the hearing-impaired can view sounds visually - effectively, "lip-reading" the entire human face. ” [8]. As shown in Table 1 and Table 2, there can exist a many to one

mapping from phonemes to visemes, i. e., for some cases, a single viseme is used to represent more than one phoneme.

TABLE 1
CONSONANT PHONEMES AND VISEMES IN ENGLISH LANGUAGE (AFTER [9]).

Phone Classes	Phonemes	Example words	Visemes
Stops	/b/	bee	B
	/d/	day	D
	/g/	gay	G
	/k/	key	G
	/p/	pea	B
Affricates	/t/	tea	D
	/ch/	choke	CH
	/jh/	joke	CH
Fricatives	/dh/	then	D
	/f/	fin	F
	/s/	sea	D
	/sh/	she	CH
	/th/	thin	D
	/v/	van	F
	/z/	zone	D
	/zh/	azure	CH
Nasals	/m/	mom	B
	/n/	noon	G
	/ng/	sing	G
Glides	/l/	lay	G
	/r/	ray	R
	/w/	way	R
	/y/	yacht	G
Whisper	/hh/	hay	G

TABLE 2
VOWEL PHONEMES AND VISEMES IN ENGLISH LANGUAGE (AFTER [9]).

Phone Classes	Phonemes	Example words	Visemes
Front	/a/	bat	AE
	/e/	bet	AE
	/ih/	bit	IH
	/i/	beet	IH
Mid	/aa/	bott	AA
	/ah/	but	AH
	/o/	bought	AO
	/er/	bird	ER
Back	/uh/	book	UH
	/u/	boot	UH
Diphthongs	/aw/	bout	AA,UH
	/ay/	bite	AA,IH
	/ey/	bait	AE,IH
	/ow/	boat	AO,UH
	/ou/	boy	AO,IH
Silence	/h#/		#

AUDIOVISUAL SPEECH SYNTHESIS

Audiovisual speech synthesis is a synchronized synthesis of an acoustic speech signal and a visual face animation in order to achieve a realist model. There are several methods to generate audiovisual speech synthesis, but in this work we will focus on the concatenative model, whose block diagram is shown in Figure 2.

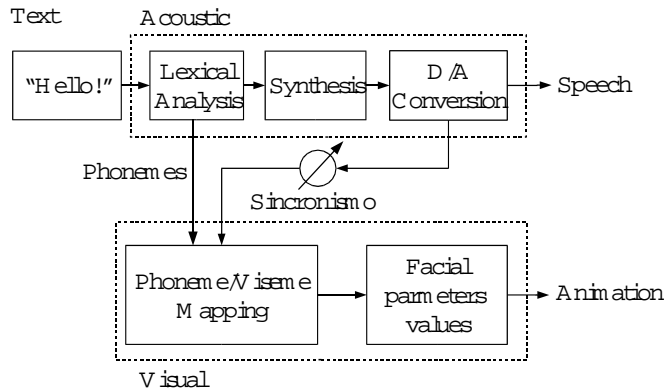


FIGURE 2

BLOCK DIAGRAM FOR AN AUDIOVISUAL SPEECH SYNTHESIS SYSTEM.

Based on this block diagram, we can see that the process of audiovisual speech synthesis can be divided into two parts: acoustic and visual synthesis. Next each part will be explained separately.

Acoustic Synthesis

- Initially, a written text is converted into a sequence of phonemes through a orthographic-phonetic transcriptor (lexical analysis), that can be either manual or automatic.
- Based on the phonetic transcription, a speech synthesizer generates an acoustic waveform that corresponds to the written text. At this stage, prosodic information can also be added to make the final speech sound more naturally.
- The final D/A conversion transforms the digital waveform into an analog signal to be reproduced in a loudspeaker.

Visual Synthesis

- The phoneme sequence generated by the lexical analyzer is converted into the corresponding viseme sequence, according to the rules established in Table 1 and Table 2.
- The face parameters are modified according to the viseme sequence, to generated the animated character.

The fidelity of the visual reproduction of each phone as

well as the synchronism between audio and video are fundamental for this technique to be successful. The McGurk effect is one of such example of bad things that can happen if these aspects are not properly treated.

PARAMETRIC REPRESENTATION OF THE VISEMES

Motivations

The facial structures responsible to speech understanding are mainly the lips, tongue, teeth and jaw. Among them, it's easy to see that the lips movement is the responsible for most of the information, considering only the visual part. For this reason, lips modeling is the natural candidate for a first implementation.

After this reasoning, we decided to initially implement parametric models for 5 vowels: /a/, /e/, /i/, /o/ and /u/. For this work, the visemes generated in [7] were taken as reference. Observing the visemes, we concluded that a good strategy could be to model a mouth with conical curves (parabolas), as shown in Figure 3.

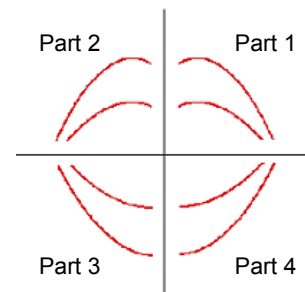


FIGURE 3

AS QUATRO PARTES QUE COMPÕEM O LÁBIO.

Another thing that has to be noted is that due to the lips symmetry, we need only to implement the parabolas of Part 1 and Part 4. The remaining ones can be derived from these ones just by mirroring the original ones. Then for each part, two parabolas were implemented corresponding to the upper and lower limit of each lip, as shown in Figure 4.

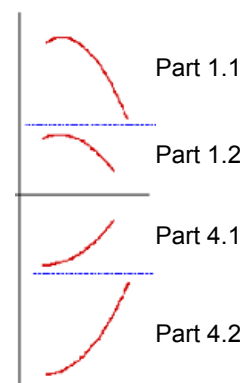


FIGURE 4

UPPER AND LOWER BOUNDS FOR EACH PART OF THE LIPS MODEL.

Before the explanation of the model, it's convenient to make a brief brake and turn into some mathematical review of conical curves.

The Parabola

The parabola is defined as the geometric place of the points that are equidistant from a reference line and a point called *focus* of the parabola. There two other parameters (g, h), corresponding to the x and y coordinates of the minimum/maximum point of the parabola. These quantities are shown in Figure 5:

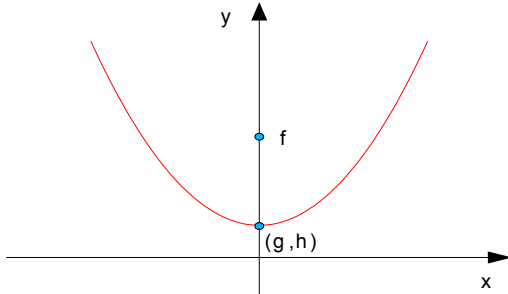


FIGURE 5
PARABOLA AND ITS PARAMETERS.

Based on these parameters, we can express the parabola equation as:

$$y = \frac{x^2}{4 * f} - \frac{g * x}{2 * f} + \frac{g^2 + 4 * f * h}{4 * f} \quad (1)$$

Considering a parabola that is symmetric about the point $x = 0$, the focus is related to the other parameters as follows:

$$f = -\frac{(x + g)^2}{4 * h} \quad (2)$$

where x is the positive root of the parabola. In the next section, the lips models were generated in terms of these three parameters: g , h and f .

Generated Models

Based on (1) and (2) the viseme models were generated using the values shown in Tables 3 to 6. These values were manually set in an interactive trial and error procedure, in order to generate lips models that look like the original ones.

A common coordinate system was used for all 4 parabolas to facilitate the parameters update when generating the animation.

TABLE 3

PARABOLA PARAMETERS FOR PART 1.1 (OUTER LINE OF UPPER LIPS)

vowel	x	g	h
/a/	3,24	1,15	4,8
/e/	3,5	2,1	4
/i/	2,8	2,1	2,6
/o/	2,1	1,1	4,4
/u/	2,6	1,1	4,5
silence	3,8	1,5	3,9

TABLE 4

PARABOLA PARAMETERS FOR PART 1.2 (INNER LINE OF UPPER LIPS)

vowel	x	g	h
/a/	2,6	1,15	2,2
/e/	1,6	2,4	1,4
/i/	0,7	2,4	0,5
/o/	0,3	0,5	1,5
/u/	1,4	0	0,8
silence	5	0	1

TABLE 5

PARABOLA PARAMETERS FOR PART 4.2 (OUTER LINE OF LOWER LIPS)

vowel	x	g	h
/a/	4,9	0	2,55
/e/	6,4	0	1,6
/i/	5,5	0	1,1
/o/	0,9	0,2	0,7
/u/	1,4	0	0,2
silence	5	0	0,4

TABLE 6

PARABOLA PARAMETERS FOR PART 4.1 (INNER LINE OF LOWER LIPS)

vowel	x	g	h
/a/	5,54	0	5,35
/e/	7,9	0	5
/i/	7	0	4
/o/	4,3	0	4
/u/	4,7	0	4,2
silence	6,8	0	3,4

With these values, the following visemes were obtained:

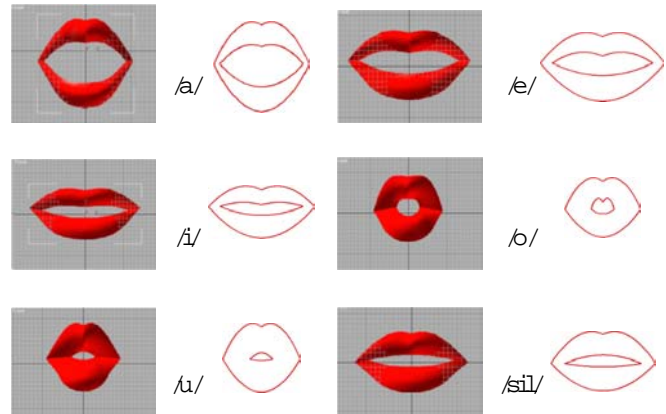


FIGURE 6

COMPARISON BETWEEN THE GENERATED VISEMES (AT RIGHT) AND THOSE REPORTED IN [7] (AT LEFT).

CONCLUSIONS AND FUTURE WORK

In this paper a lip model based on conical curves were presented. In spite of it's mathematical simplicity, the generated models are very realistic and, in fact are very close to the much more complex 3D models presented in [7].

A brief theoretical introduction to the subject was also provided, with the most important issues related to this subject, including some interesting aspects, like the McGurk effect.

For the future, the plans are to animate these visemes synchronized with an input audio in order to study how this transition should be performed. For example, this translation can be linear? And in the case of stops, where the transitions are abrupt?

Another thing that comes to mind is that it's not enough to perform transitions between the parameter values because it would lead to discontinuities in the generated images. Maybe another kind of parameterization should be considered.

REFERENCES

[1] Dominic W. Massaro & David G. Stork, "Speech Recognition and Sensory Integration", *American Scientist*, 1998, vol. 86, p. 236-244.

[2] Lawrence S.Chen, Jörn Ostermann, "Animated Talking Head with Personalized 3D Head Model", *Proceedings of 1997 Workshop of Multimedia Signal Processing*, June 23, 1997, pp: 274-279.

[3] <http://www.haskins.yale.edu/haskins/heads.html> (31/03/2003)

[4] David Lindsay, "Talking Head", *Invention & Technology*, Summer 1997, 57-63.

[5] <http://www.speech.cs.cmu.edu/comp.speech/> (31/03/2003)

[6] J. Kulju, M. Sams and K. Kaski, "A Finnish Talking Head", *Proc. of Finnish Fonetiks Symposium*, August 1998.

[7] http://face.ee.nus.edu.sg/Hari/Masters_home_page/lips_database.htm

[8] http://whatis.techtarget.com/definition/0,,sid9_gci213308,00.html

[9] SoonKyu Lee and Dongsuk Yook, "Viseme Recognition Experiment Using Context Dependent Hidden Markov Models", *Intelligent Data Engineering and Automated Learning*, Springer-Verlag, August 2002, pp. 557-561

[10] <http://mambo.ucsc.edu/psl/dwmdir/da.html>