# 3D Facial Animation for Mobile Devices

José Mario De Martino and Tatiane Silvia Leite
Department of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
University of Campinas
P.O. Box 6101 - ZIP 13083-970 Campinas - SP - Brazil
martino@dca.fee.unicamp.br, tatiane.leite@gmail.com

**ABSTRACT**

This article presents the implementation of a 3D facial animation system for mobile devices. Due to the large processing and memory requirements for this type of application, its use on mobile devices was not possible until recently. Currently, however, with the increasing development of powerful hardware and with the spread of cellular telephony, 3D applications for these devices have become even more promising. The system presented in this article was implemented based on a previous one designed for desktop platforms. Both systems are driven by two kinds of input. The first one is an audio file containing the speech to be presented, and the second is a timed phonetic transcription specifying the phonemes of the utterance and their associated time spans. With the input data, the system converts phonemes into phonetic context dependent visemes that handle perseveratory and anticipatory coarticulation effects. The visemes are then used to modify the 3D polygonal mesh of the fa ce keeping the synchronization with the audio. The initial evaluation of the systems indicates that, despite the limitations of mobile platform, the implementation of video telephony using 3D facial animation is viable in modern mobile devices.

**Keywords:** facial animation, mobile devices, visemes.

## 1 INTRODUCTION

Facial animation is a crucial technology for the development of embodied virtual characters capable of reproducing the communication style, based on speech, facial expressions and gesture, which humans are well familiar with. This technology refers to the techniques for specifying and controlling a synthetic face to reproduce facial expressions and the visual articulatory movements associated with the production of the speech. Unlike the cartoon-like animations, which do not represent all the movements of the face during a speech, admit exaggeration, and usually are not concerned with reality, computer facial animation usually strives for realism.

Depending on the application, the virtual characters created by the computer facial animation can play such diverse roles as instructors, assistants, avatars, presenters or sale representatives. Not so long ago this kind of software was only possible to be implemented in rendering servers or workstation with powerful processor-memory architecture. The large requirements of real-time processing were impediments for implementing this technology in other platforms, specially on mobile ones. But thanks to the advances in microelectronics and the spread of mobile telecommunication, modern mobile devices have the powerful hardware required for this type of application. Additionally the features demanded by users currently have changed significantly, which stimulate the development of innovative applications such as facial animation for mobile devices. In particular, the deployment of facial animation on mobile phones is a promising option for mobile video telephony due to the poten tial reduction of data transmission. In this scenario, instead of video transmission, it is only necessary to transmit a few parameters specifying the speech and associated facial expressions.

The system implemented seeks to perform the animation of the human face in close agreement with the acoustic signal of speech. For the generation of facial animation in sync and in harmony with the speech, it is imperative to reproduce the articulatory movements associated with all phonemes of the language. For that, in addition to identifying the postures of the articulatory gestures associated with the phonemes, it is also necessary to have the representation of the transitions between these postures considering the coarticulation effects.

The coarticulation effects are expressed by the changing of the articulatory pattern of a particular phoneme according to the influence of the articulation of an adjacent phoneme or a close one in the chain of sound production. The effects of coarticulation make, for instance, the "b" of the word "bat" visually different from the "b" of the word "boot". In the latter case, the articulatory movement necessary for the production of the sound associated with the "oo" significantly influences the visible aspects of the articulation of the "b". The coarticulation can be classified as perseveratory or anticipatory. In perseveratory coarticulation, the articulation of a speech segment is influenced by the articulatory movement of a predecessor segment in the chain that composes the phonetic speech. In anticipatory coarticulation, the articulatory movement of the segment is influenced by the articulation of a successor segment.

One of the most important issues of any speech synchronized facial animation system is the establishment of proper parameters for the visible articulatory representation, which is also known as visemes. The system proposed in this article is based on a set of visemes that contemplates the effects of adjacent perseveratory and anticipatory coarticulation for the Brazilian Portuguese language [1].

This paper is organized as follows: Section 2 shows a review of the state of the art in facial animation for mobile devices; Section 3 presents the system architecture; Section 4 provides an overview of the set of visemes; Section 5 explains the facial movements changes occurred during the speech; Section 6 discusses the influence of the movement in the vertexes around the mouth; Section 7 presents the strategy to model de visible articulatory movements based on the articulatory targets of the visemes; Section 8 shows a preliminary evaluation of the implemented system; and finally in Section 9 the conclusion is presented.

## 2 RELATED WORKS

Many different approaches for facial animation designed to run on desktop computers have been proposed in the literature. However, there are fewer published works that refer exclusively to facial animation for mobile platforms.

The main articles in this area have used the MPEG-4 standard [4], also characterized as geometric parametrization, to perform facial animation. Next, it is summarized the most important works based on this standard.

The system presented in [8] describes an approach to enable facial animation framework applications for the Web and mobile platforms. In the work [6], the facial animation is developed to achieve the implementation of a system called "LiveMail: Personalized Avatars for Mobile Entertainment". In this system, the speech and the appropriate facial animation are created automatically by speech synthesis. The facial animation created can be sent over the network as 3D animated messages or as short videos in MMS. The work in [7] describes a parametric 3D facial animation capable of representing and animating faces in a photo-realistic manner.

All mentioned research efforts are based on the MPEG-4 standard for facial animation. In this standard, a face is defined by 84 FPs (Feature Points), which are controlled by 68 FAPs (Facial Action Parameters). Among all 68 FAPs, 66 are considered low level. These FAPs are based on the study of minimal facial actions and are related to muscle actions. They represent a complete set of basic facial actions, and therefore allow the representation of facial expressions more naturally.Basically each low-level FAP is used to define the displacement of an FP in a particular direction. The FPs (and their associated FAPs) are arranged in ten different groups. The FAPs can be efficiently compressed and included in a Face and Body Animation (FBA) low-bit rate streams. Any MPEG-4 compliant facial animation system can decode and interpret FBA bitstreams, and a synthetic animated face can be seen.

The realistic articulatory facial animation in the MPEG4 standard presents difficulties. The low-level FAPs, despite their names suggest an articulatory interpretation, are purely geometric parameters. For example, the FAP *Open-Jaw* operates only in the FP *Bottom of the chin* not influencing, for example, the FPs *Bottom of the lower teeth* and *Top of the lower teeth*, which should move with the jaw in an articulatory interpretation. Additionally, the low-level FAPs do not take into account articulatory gestures as, among others, the mouth opening and the height of the mouth and lip protrusion. The mapping of these gestures to a set of FAPs is not obvious, and is not defined in the standard. To overcome these difficulties, the standard defines a set of high-level FAPs to represent visemes. Unfortunately the 14 high-level visemes that has been defined have the serious limitation of only contemplating, if so, the English language. Furthermore, the standard leaves to the developer the decision of how these visemes deform the face. And finally, it is possible, although not recommended, to change the face using the high-level and low-level parameters simultaneously, which can lead to unpredictable results.

The work [5] presents the system called InCA, which is a distributed personal assistant conversational agent. The front-end runs on a PDA and uses facial animation and natural speech to interact with the user and provide services such as e-mail, calendar, weather reports, among other services, using the network to overcome the PDA computational limitations. Its facial animation considers input data very similar to what it used in work presented in this paper. InCA also needs a playlist with phonemes and the corresponding timing information and the audio speech. However, InCA uses only 18 mouth positions which are generated before the speech production. This way, each phoneme produced is mapped into one of these mouth positions.

Differently from the above mentioned works, the approach presented in this paper uses phonetic context dependent visemes that cope with perseverative and anticipatory coarticulation. Additionally, the movements of the temporomandibular joint and the lip tissue are modeled from a set of visemes established by the analysis of a Brazilian Portuguese linguistic corpus. Although the corpus is restricted to Brazilian Portuguese, the methodology is general enough to be applied to other languages.

## 3 SYSTEM ARCHITECTURE

The implemented system can be represented by the schematic diagram in Figure 1. The core of the system, shown in the darker area, consists of two modules: Phoneme-Viseme Converter and Facial Animation. The Phoneme-Viseme module is responsible for converting the timed phonetic description in a sequence of visemes. The Facial Animation module uses the sequence of visemes generated by the Phoneme-Viseme converter to control the movements of the virtual face. The main task of the facial animation system is to reproduce the articulatory movements described by the visemes and, through the required calculations, to generate the sequence of images that make up the animation.

For each viseme a sequence of images is generated and shown on the screen of the mobile device. As soon the images are generated from Facial Animation module, they are reproduced on the screen synchronized with the audio, using the functionality of the Mobile Media API for this management [10].

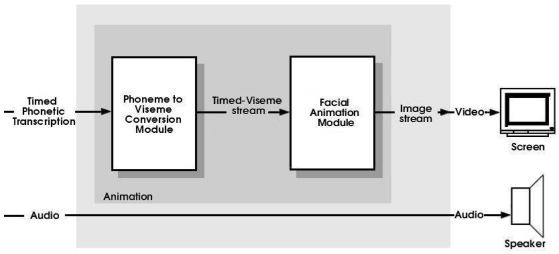



Figure 1: Diagram of animation system implemented.

The first step of the system is the input of the timed phonetic transcription and the corresponding conversion of the phonetic segments into visemes. This timed phonetic transcription contains the sequence of phonetic segments that compose a particular utterance, as well as the duration of each segment. From that description, the phonetic segments that compose the speech are separated into groups of visemes, and at the end of the process a sequence of visemes, identified and established for the Brazilian Portuguese [1] is obtained. Such visemes will give the necessary information for the creation of different images of the virtual head through its parameters.

After the identification of the visemes in the speech, the system makes the changes corresponding to these geometric visemes, modifying the polygonal mesh of the 3D virtual face.

## 4 PHONEME TO VISEME CONVERSION

For the characterization of a set of visemes for the Brazilian Portuguese, the article [1] identifies the visible patterns of articulatory movement during the production of the phonetic segments of the language in different phonetic contexts.

The identified visemes are expressed by a parametric model that incorporates the articulatory targets established for the segments and the instant of reaching these targets within the range of sound production, still contemplating the effects of coarticulation between adjacent segments in the chain of speech production.

The context-based visemes expressed with the symbols of the International Phonetic Alphabet [3] are shown in Tables 1 and 2.

| Viseme | Context |
|---|---|
| <$p_1$> | [pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ] |
| <$p_2$> | [pu] [upɪ] [upʊ] |
| <$f_1$> | [fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ] |
| <$f_2$> | [fu] [afʊ] [ufɪ] [ufɐ] [ufʊ] |
| <$t_1$> | [ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ] |
| <$t_2$> | [ta] [atɐ] |
| <$s_1$> | [si] [sa] [isɪ] [isɐ] [asɪ] [asɐ] |
| <$s_2$> | [su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ] |
| <$l_1$> | [li] [ilɪ] [alʊ] [ulɪ] [ulɐ] |
| <$l_2$> | [la] [ilɐ] [alɪ] [alɐ] |
| <$l_3$> | [lu] |
| <$l_4$> | [ilʊ] [ulʊ] |
| <$ʃ_1$> | [ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ] |
| <$ʃ_2$> | [ʃu] [uʃʊ] |
| <$\lambda_1$> | [λi] [λa] [iλɪ] [iλɐ] [aλɪ] [aλɐ] |
| <$\lambda_2$> | [λu] [uλɪ] [uλɐ] |
| <$\lambda_3$> | [iλʊ] [aλʊ] [uλʊ] |
| <$k_1$> | [ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ] |
| <$k_2$> | [ka] [akɐ] |
| <$k_3$> | [ku] [ikʊ] [akʊ] [ukʊ] |
| <$ɾ_1$> | [ɣi] [ɣa] [irɪ] irɐ] [arɪ] [arɐ] [urɐ] |
| <$ɾ_2$> | [ɣʊ] [irʊ [arʊ] [urɪ] [urʊ] |

Table 1: Consonantal visemes and their phonetic contexts.

| Viseme | Context |
|---|---|
| <$i_1$> | all contexts but [tit] and [ʃiʃ] |
| <$i_2$> | [tit] [ʃiʃ] |
| <a> | all contexts |
| <u> | all contexts |
| <ɪ> | all contexts |
| <ɐ> | all contexts |
| <ʊ> | all contexts |

Table 2: Vocalic visemes and their phonetic contexts.

It is important to remember that although the system has been developed for the parameters determined for the Brazilian Portuguese, it can be adapted to other languages.

# 5 FACIAL MOVEMENTS

The geometric transformations applied to the 3D head geometric model are of two kinds: rigid body transformations and deforming transformations. Rigid body transformations are associated with the articulatory movement that allows the protrusion or retraction (moving forward and backward) and elevation or depression (moving up and down) of the jaw, working only with the lower lip vertexes, and with the face below the lower lip, as shown in Figure 2.

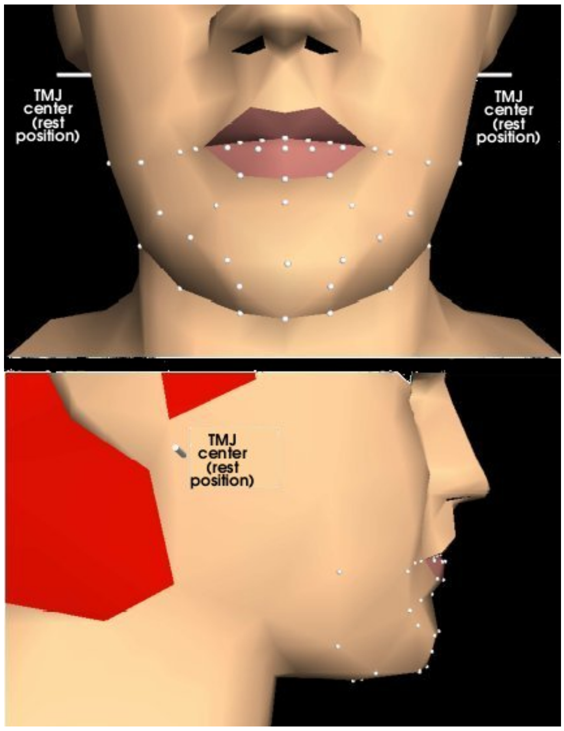



Figure 2: Vertexes transformed because of the articulatory movements.

The deforming transformations are associated with the voluntary movements of the face tissue, which includes the region of the lips and the region around them that characterize the production of the speech segments.

The facial deformations are represented by the trajectories of the fiduciary points $P1$, $P2$ and $P3$, located in the upper lip, in the corner of the mouth and in the lower lip, respectively, as seen in Figure 3. However, only the $P3$ point suffers significant influence from the jaw movement, showing that this is the point that suffers greater influence considering both movements: voluntary movement of the lip and the movements of the jaw. The $P1$ and $P2$ points don't suffer significant influence from the movements of the jaw and their behavior are defined only by the voluntary movements of the lip. The trajectories of the fiduciary points were captured from a real speaker as reported in [1].

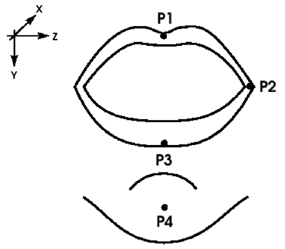



Figure 3: Fiduciary Points.

The rigid body transformations assume that, during speech production, the TMJ (temporomandibular joint) executes both rotation movement and translation movement in sagittal plane). The rotation movement is performed around the center of the TMJ, but this center can also suffer translation due to the sliding of the articular disc, shown in Figure 4. Thus, the trajectory of the fiduciary P4 point located in the chin of the speaker, represents the motion of the jaw and allows to estimate the components of TMJ rotation and translation.

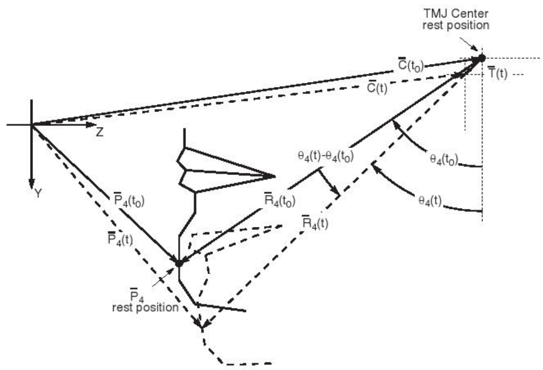



Figure 4: Temporomandibular joint behavior.

## 5.1 Temporomandibular joint behavior

The rotation movement of the jaw is done around the TMJ center that undergoes translation due to movement of the articular disc. Assuming that the trajectory from the fiduciary P4 point, in Figure 4, reflects the movement of the jaw, it is possible to estimate the components of rotation and TMJ translation.

The angle of rotation can be given by the equation 1.

$$\theta_4(t) = arctan\left(\frac{z_4(t) - z_C(t_0)}{y_4(t) - y_C(t_0)}\right) \quad (1)$$

where $y_4(t)$ and $z_4(t)$ are the coordinates $y$ and $z$ of the P4 trajectory at instant $t$, and $y_C(t_0)$ and $z_C(t_0)$ are the coordinates $x$ and $y$ of the TMJ center at instant $t_0$.

It is also possible to estimate the $r_4$ radius, considering the position of rest, presented in equation 2.

$$r_4 = \sqrt{[y_4(t_0) - y_C(t_0)]^2 + [z_4(t)0) - z_C(t_0)]^2} \quad (2)$$

And the components of TMJ translation are given by the equations 3 and 4.

$$y_T(t) = y_4(t) - y_C(t_0) - r_4 cos(\theta_4(t)) \quad (3)$$

$$z_T(t) = z_4(t) - z_C(t_0) - r_4 sin(\theta_4(t)) \quad (4)$$

## 5.2 Lower lip behavior

The fiduciary $P3$ point, in Figure 3, represents the movement of the lower lip. Its trajectory may be separated into two components. The first is referent to the jaw movement that affects the opening of the mouth, thus changing the position of the lower lip. Therefore, the angle of $P3$ rotation is given by the equation 5.

$$\theta_3(t0) = arctan\left(\frac{z_3(t0) - z_C(t_0)}{y_4(t) - y_C(t_0)}\right) \quad (5)$$

And the $r_3$ radius is given by equation 6.

$$r_3 = \sqrt{[y_3(t_0) - y_C(t_0)]^2 + [z_3(t)0) - z_C(t_0)]^2} \quad (6)$$

The second component is referent to the voluntary movement of the lip tissue. This movement is necessary to produce some specific articulatory postures, as protrusion and lip extension. The equations 7 and 8 represent the behavior of the lower lip.

$$y_L(t) = y_3(t) - y_4(t) - r_3 cos(\theta_3(t_0) + \theta_4(t) - \theta_4(t_0)) - r_4 cos(\theta_4(t)) \quad (7)$$

$$z_L(t) = z_3(t) - z_4(t) - r_3 sin(\theta_3(t_0) + \theta_4(t) - \theta_4(t_0)) - r_4 sin(\theta_4(t)) \quad (8)$$

## 5.3 Upper lip behavior

It is assumed that the behavior of the upper lip does not have significant influence over the handling of the temporomandibular joint, so their movements are defined exclusively by the voluntary movement of the upper lip tissue in order to produce the sounds of speech. Therefore, the fiduciary $P1$ point is responsible for defining the behavior of the upper lip.

# 6 REGIONS OF INFLUENCE

To make the geometric changes concerning the voluntary movement of facial tissue around the mouth, the original system provides a model to express the visible movement of this tissue during the talk.

This model provides a spheroidal region of influence around the mouth, inspired in the elliptical formation of the orbicular muscle of the mouth [9]. The region of influence was split into two parts:

- Upper Region: influenced by the behavior of the upper lip;
- Lower Region: influenced by the behavior of the lower lip.

In the system developed for PC, each one of these regions are two concentric spheroids, illustrated in Figure 5, which presents the region that influences the upper lip behavior. And considering these spheroids, the distance is calculated from each vertex to the $P1$ and $P2$ points and from these calculations the movement factor is set, which determines how much each vertex is influenced by the changes.

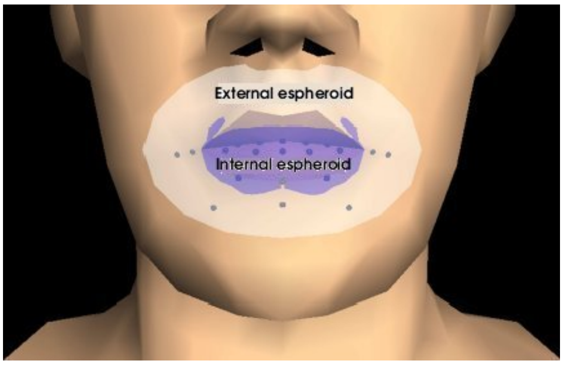



Figure 5: Upper region of influence and associated vertices.

But in the developed system, the values of these factors were pre-established, so they are not calculated during the execution of the program. It was designed to avoid an overburdened processing. And from the values of these factors are calculated the new positions of the vertexes, and in sequence performed the geometric transformations.

# 7 INTERPOLATION

The interpolation curve adopted for the representation of the movement was the cubic parametric curve of Hermite [2]. This choice was taken because it is important that the curve of geometric interpolation preserves the continuity between the various segments of a chain of visemes / phonemes and guarantees that the time derivative equals to zero in the instants when the articulatory target is reached.

Thus, knowing the values of displacements in $x$, $y$ and $z$ directions associated with the articulatory targets of a viseme and the instants of reaching these targets, it is possible to approximate the trajectory of the fiduciary points in a chain of phonemes by the interpolation of the articulatory target displacements.

Equation 9 presents the parametric model adopted.

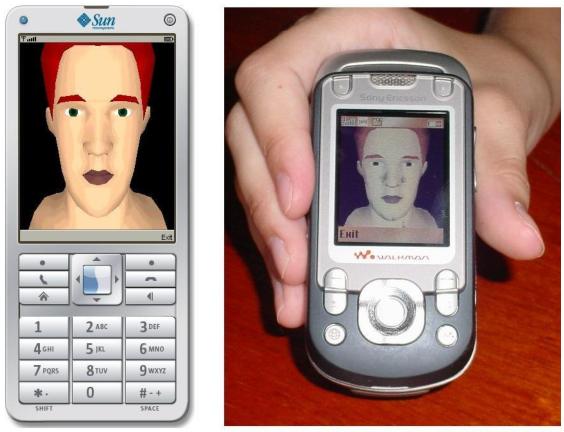
$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} I_x & F_x \\ I_y & F_y \\ I_z & F_z \end{bmatrix} \begin{bmatrix} 2 & -3 & 1 \\ -2 & 3 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ 1 \end{bmatrix} 0 \leq t \leq 1 \quad (9)$$

The parameters of the equation define:

- $x(t)$, $y(t)$ and $z(t)$: coordinates of the fiduciary point;
- $I_x$, $I_y$ and $I_z$ are the shifts in $x$, $y$, and $z$ directions from the fiduciray point in the articulatory target of the initial viseme;
- $F_x$, $F_y$ and $F_z$ are the shifts in $x$, $y$, and $z$ directions from the fiduciary point in the articulatory target of the final viseme;
- and $t$ is an independent variable, normalized in relation to the time interval between two viseme targets.

## 8 PRELIMINARY EVALUATION

Figure 6 shows the snapshots of the implemented 3D facial animation system running on simulator and on a mobile phone.




(a) (b)

Figure 6: (a) in the WTK (Wireless Toolkit) simulator (b) in the device W600i Sony Ericson, which supports the 3D API.

The implemented system was evaluated and had some of its characteristics analyzed, such as the size of the resulting JAR file of the generated code, the necessary memory for the execution and a survey of the most relevant Java methods to the performance of the system through the time of execution and execution frequency of these methods. All these characteristics were raised using the well-known WirelessToolkit 2.5 simulator.

The first feature evaluated was the size of the JAR package, which when is generated contains all the files needed to run the application on the mobile device. The JAR file contains the M3G file with the virtual head, the audio file (WAV) with the speeches, the phonetic transcription timed for the speeches and the bytecodes (Java files) needed to execute the application in the Java Virtual Machine. At the end of the development, the JAR file had approximately 300 KBytes, which is therefore a viable size even for mobile phones with limited memory capacity.

The second feature evaluated was the size of dynamic memory (heap) needed for the execution. This plays an even more important role, because if we do not have enough memory to create the objects, the application will not be executed properly. The size of heap memory used in this application is approximately 2 MB, indicating that the current mobile devices will not have problems on executing the application, because they usually have a significantly greater heap memory.

Finally, the execution frequency of methods and the duration of these executions were analyzed. The results obtained met the expected results; approximately 78% of the processing time refers to the core image painting module (TalkingHeadCanvas.paint). This module contains the methods that are most called during the execution of the application. These methods refer to the updating schedule, calculations of rotation and translation and updating of images on the screen. The other 22% of the execution time is related to the other methods, such as M3G file loading, and also the WAV file loading, among others.

To test the implementation in real devices, it was downloaded into a Sony Ericsson mobile phone, model W600i, which is shown in Figure 6. Not only this Sony Ericsson model supports this type of application, but several models of the largest mobile phone manufacturers also support this type of application, such as Motorola, Nokia, LG and others. These newer models are able to run the implemented application, as well as have a reasonable processing power, an appropriate screen size, dynamic memory equal to or above 2 MB and support the Mobile 3D Graphics API (JSR-184).

## 9 CONCLUSIONS

Even with all the memory and processing limitations in a mobile device, 3D applications have shown to be increasingly promising for embedded systems. The system presented in this paper shows a solution of a specific facial animation for these devices. The main concern is beyond the processing power of the device; it refers primarily to the small display that probably could prevent the visualization of the speech of the talking head. In this work it was possible to see that despite the small size of the screen, the 3D presentation could be done in a satisfactory way, becoming a major attraction since the facial animation is a technology that can help to enable video telephony services at low rates of transmission, which is a common situation in cellular networks. The facial animation loaded on a mobile device needs only the transmission of phonetic timing information, or equivalent information for the visual presentation of the interlocutor, thereby preventing the transmission of a large volume of data associated with the video transmission.

Among the various aspects that make the production of a realistic facial animation a complex achievement, the work has focused on the treatment of visible articulatory movements, taking into account the concept of visemes as well as the effects of articulation which are applied to them. that they suffer. In spite of the fact that the system was developed to Brazilian Portuguese speeches, it can easily be adapted for other languages.

# REFERENCES


[1] Jose Mario De Martino, Leo Pini Magalhães, and Fabio Violaro. Facial animation based on context-dependent visemes. *Computers & Graphics*, 30(6):971–980, 2006.

[2] J. D. Foley, A. van Dam, S. K. Feiner, and Hughes J. F. *Computer Graphics Principles and Practice*. Addison-Wesley Publishing, 2 edition, 1990.

[3] IPA. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. International Phonetic Association, International Phonetic Association - Cambridge, 1999.

[4] ISO/IEC. *Information technology - Coding of audio-visual objects - Parte 2: Visual*. ISO/IEC 14496-2:2001(E), Geneva, Switzerland, December 2001. International Standard.

[5] M. W. Kadous and C. Sammut. *InCA: A Mobile Conversational Agent*. Springer, 2004.

[6] Miran Mosmondor, Tomislav Kosutic, and Igor S. Pandzic. Livemail: personalized avatars for mobile entertainment. In *MobiSys '05: Proceedings of the 3rd international conference on Mobile systems, applications, and services*, pages 15–23, New York, NY, USA, 2005. ACM.

[7] P. Omedas, F. Berrizbeitia, G. Szijártó, B. Kiss, and B. Takács. Model-based facial animation for mobile communication. *1st Ibero - American Symposium on Computer Graphics (SIACG)*, 2002.

[8] Igor S. Pandzic. Facial animation framework for the web and mobile platforms. In *Web3D '02: Proceedings of the seventh international conference on 3D Web technology*, pages 27–34, New York, NY, USA, 2002. ACM.

[9] F. I. Parke and K. Waters. *Computer Facial Animation*. AK Peters, 1996.

[10] Sun Microsystem. *Java Specification Request 135: Mobile Media API*. `http://jcp.org/en/jsr/detail?id=135` (Last access: 05/14/2008).