Speech Synchronized Image-Based Facial Animation

Paula Dornhofer Paro Costa  
Dept. of Comp. Eng. and Industrial Automation  
School of Electrical and Computer Eng.  
University of Campinas  
P.O. Box 6101 - 13083-970  
Campinas - SP - Brazil  
paula@dca.fee.unicamp.br

José Mario De Martino  
Dept. of Comp. Eng. and Industrial Automation  
School of Electrical and Computer Eng.  
University of Campinas  
P.O. Box 6101 - 13083-970  
Campinas - SP - Brazil  
martino@fee.unicamp.br

Edson José Nagle  
CPqD - Telecommunications Research and Development Center  
13086-902  
Campinas - SP - Brazil  
nagle@cpqd.com.br

Abstract—The arise of a new reality where information is accessed ubiquitously from a variety of different devices creates new application opportunities and the need of development of more efficient and intuitive interfaces. This can be specially shown through the seamlessly growth of wireless communications and the use of small and portable devices to communicate and have access to multimedia content. This paper describes a speech synchronized 2D facial animation system that explores the computer facial animation as an enabling technology for the development of more natural and efficient human-machine interfaces and applications. The photorealistic results provided by the system makes it suitable for very low bit rate video telephony applications through the implementation of a model based facial video coding. The designed system is a cross-platform solution capable of delivering videorealistic animations on both limited capacity devices (such as mobile phones and personal digital assistants) and high performance desktop computers with full processing power and memory capacity. The key concepts behind the solution are an image database that can be continuously scaled and an image-based animation synthesis strategy that can be adapted according to the characteristics of the executing platform and/or applications.

Index Terms—facial animation, model-based video coding, image-based facial animation, text to visual speech.

I. INTRODUCTION

The recent advances and worldwide spreading of mobile communications, followed by the increased availability of data networks, made possible the ubiquitous access to information. Moreover, thanks to the evolution of microprocessors, microcontrollers and memory devices, we presence today a large number of portable devices that, among a variety of other functionalities, provide access to data networks and have good performance to handle multimedia content. This results in an increased number of users having completely different ways to access the information media without necessarily being familiarized with regular WIMP (Windows, Icons and Pointing Devices) personal computer interfaces.

On the other hand, although broadband access is being popularized, a large amount of Internet users and mobile subscribers still have low bandwidth data access, resulting in low QoS (Quality of Service) for video streaming applications.

In this context, this work explores the computer facial animation technology combined with a text-to-speech synthesizer (TTS) as an enabling technology to the development of more intuitive and efficient interfaces, taking advantage of face-to-face communication mechanisms we are well trained and aware of. This approach makes possible the development of virtual talking heads that find applications in areas such as: entertainment, personal communications, navigation aid, newscasting, electronic commerce and education [1], [2].

Additionally, computer facial animation is a model based video-coding approach that is appropriate for applications that require visual interaction capacity over a very low bit rate channel and advantageous to the user in situations where he pays for the volume of transmitted data [3].

Reliable and convincing synthetic talking-heads have to deliver satisfactory levels of photorealism and videorealism. In this context, photorealism can be understood as a measure of how close the images generated/displayed by the system can be mistaken by photographs. The photographic quality includes the reproduction of fine details like skin texture, facial wrinkles and hair. A videorealistic animation, besides photorealism, presents accurate lip synchronization, carefully reproducing the speech articulatory movements, and is also able to reproduce nonverbal communication gestures (like head nodding) and movements not directly related to communication (such as eyes blinking and eyebrows movements).

The reproduction of speech articulatory movements involves the modeling of the visible movements of vocal tract displayed on the speaker’s face during speech production, specially in the region around the mouth. To determine such a model it is necessary to consider the characteristic configurations of vocal tract associated to the various phonemes of the language and the related coarticulation effects that arise when the typical articulation pattern of a phoneme is modified by the interaction with nearby speech segments. In this work, each visually contrastive configuration of vocal tract is associated to a viseme, a shorthand for visual phoneme, that corresponds to a static mouth shape associated to the acoustic realization of a phoneme.

Another key factor to determine the level of videorealism of facial animation is the adopted strategy to model the human face, generally divided in two main approaches: the use of
3D geometric models or an image-based approach. Advanced and modern 3D face modeling techniques are successful in synthesizing natural looking faces and high quality images of rigid movements of the head. However, the use of polygonal meshes or other geometric models to reproduce plastic deformations, like the mouth dynamics during speech or the details of human face visual appearance, requires sophisticated models and animation control strategies, implemented at high computational costs [4].

On the other hand, image-based, or 2D, facial animation manipulates photographic pictures from an image database that are captured from a real face, synthesizing the final animation through their appropriate processing, sequencing, concatenation and presentation. This approach inherently generates photorealistic animations due to photographic nature of manipulated images. Most efforts of this type of system are then concentrated in obtaining videorealism through the correct reproduction of visual dynamics of face during speech.

This paper describes the implementation of a 2D facial animation system integrated to a text-to-speech synthesizer for Brazilian Portuguese, that generates the speech audio from a textual input and provides the corresponding timed phonetic transcription that drives the facial animation system.

The designed system encapsulates two synthesis approaches under two operation modes, making it scalable and flexible, and turning possible its deployment on a wide range of device profiles. In a reduced mode, the system is coupled to a small image database and animation synthesis applies a key framing approach based on morphing between visemes. The reduced image database and the low complexity of image morphing algorithm allow the generation of videorealistic facial animation by small devices such as smartphones, digital TV set-top boxes and PDAs (personal digital assistants).

In an extended mode, the morphing between visemes approach is combined with the concatenation of small video fragments characterized by sequences of frames stored in the image database. This combination aims to deliver higher levels of videorealism through an image database that can be continuously scaled. Starting from a minimum image database size that characterizes the reduced mode, the image database can grow continuously resulting in greater videorealism. The additional necessary memory and processing capacities make the extended mode well suited to desktop platforms.

This paper focus on the implementation aspects of reduced mode and its contributions. In order to model the coarticulation effects, our system applies an efficient and straightforward approach through the utilization of context-dependent visemes identified for Brazilian Portuguese [7]. Such modeling allowed a completely new approach to the morphing between visemes animation synthesis strategy, resulting in videorealistic animations from a reduced number of visemes. Typically, speech animation systems based on morphing between visemes synthesis strategy are characterized by a small image database and a low level of videorealism since the coarticulation effects are not properly modeled [5, 6]. Our approach, however, offers a better solution as it copes with the coarticulation effects, improving animation videorealism while keeping the image database small.

The main contributions of this work are: the implementation of a videorealistic 2D speech synchronized facial animation system for Brazilian Portuguese, the design of a cross-platform solution that can be adapted to characteristics of the executing platform and the use of context-dependent visemes as a straightforward approach to model coarticulation effects.

Despite the fact that this system was developed for Brazilian Portuguese, its underlying principles and the whole system implementation can also be applied to other languages.

II. RELATED WORK

The implementation of a 2D facial animation system starts with the definition an audiovisual corpus to be recorded, from which frames are extracted, analyzed and processed in order to build an image database. Figure 1 shows the typical approach to the speech synchronized facial animation synthesis process.

It begins with the definition of a new content to be uttered by the talking-head. The corresponding audio is generated by a TTS. The phonetic transcription of the audio, containing timing information of the occurrence of each phonetic unit, is the key input to the system. Based on the timed phonetic transcription, the system generates the corresponding animation frames sequence, selecting and retrieving key pose images from the image database. The synchronization between audio and facial animation is guaranteed by the generation of a sequence of frames in accordance with the phonetic and timing information previously obtained from synthesized audio.

Different implementations of 2D facial animation systems may differ on the nature of captured corpus, the coarticulation modeling approach and the synthesis strategy to obtain a videorealistic animation.

In [8], for example, the audiovisual corpus is characterized by an existing video excerpt of a subject where no restriction concerning the content of the uttered sentences or the background scenario. To handle coarticulation, small video fragments corresponding to three sequential speech segments, or triphones, are extracted from the corpus, labeled and stored in a database of video fragments. The synthesis of a new video is constructed by concatenating and stitching the appropriate triphone sequences from the database together.
Videorealism is achieved by using triphone contexts to handle coarticulation effects but the quality of final generated video can be limited by the number of triphones contexts existing in the original corpus. The storage of video fragments for a significant number of triphones contexts is memory expensive and contains redundant information due to the storage of fragments in similar contexts.

Another approach is the morphing between visemes synthesis strategy implemented by [5]. In this system, images are extracted from a recorded corpus of a subject uttering predefined words in a well controlled environment. The image database is built from 16 viseme images that are associated to groups of phonemes that are visually indistinguishable, or homophone groups. In order to construct a new visual speech stream, visemes from image database are selected according to the information provided by timed phonetic transcription. Those visemes characterizes the key poses of final animation. Transitioning between two subsequent visemes is made through image morphing technique. The main aspects of this implementation are: the image database has a reduced size; no coarticulation modeling is applied; and the transition function that controls the morphing process is linear (far from the reality of speech dynamics, where transition between two articulatory targets follows complex and non-linear behaviour).

A different synthesis technique is applied in [9]. In this work, the image database contains a large number of images corresponding to visemes produced under a variety of different phonetic contexts. The images are extracted from an audiovisual corpus recorded in well controlled conditions and with predefined sentences. Following this, the images are analyzed and organized according to their visual features and the context they are captured. At synthesis time, the system seeks to identify similar phonetic context and duration characteristics between the new content to be uttered and the samples stored in database. The system tries to use in the final animation the maximum number of frames captured in sequence during corpus recording. Otherwise, the system uses the information about visual aspects of each image that makes possible to choose the best candidate transition frames to produce videorealistic results.

Animation synthesis implemented in [9] is capable of delivering high levels of videorealism thanks to a large amount of pre-processed images stored in a database and a resourceful algorithm to search and select frames for the final animation. On the other hand, systems based on morphing between visemes synthesis technique are successful in animating speech through a small image database but the final animation typically lacks videorealism due to the simplification of modeling of visible articulatory movements reproduced by the synthetic transition between visemes.

The reduced mode of the system described in this paper applies a context-dependent viseme coarticulation model that can be directly applied to the morphing between visemes strategy, combining it with a non-linear transition function to drive the image morphing process. This results in a compact 2D facial animation system, capable of delivering videorealistic animations, based on an image database of just 34 visemes.

### III. Our Approach

#### A. Audiovisual Corpus

In order to build a viseme image database, an audiovisual corpus was built by recording the face of a female subject uttering predefined sentences in a well defined and controlled environment.

The uttered sentences were divided in a set of non-sense words and a set of phrases that contains occurrences of all phonemes of Brazilian Portuguese. The set of non-sense words were defined based on context-dependent visemes for Brazilian Portuguese identified in [7]. This study identified groups of homophones in different contexts that can be visually represented by a unique viseme. Tables I and II show consonantal and vocalic context-dependent visemes used as reference in this work, where phonemes are expressed with the symbols of the International Phonetic Alphabet [12].

The groups of phrases uttered by the subject were used to enrich the database with sequences of frames that captures the speech dynamics on more natural contexts and also to

<table>
<thead>
<tr>
<th>Homophone Group</th>
<th>Visemes</th>
<th>Phonetic Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p,b,m]</td>
<td>&lt; p₁ &gt;</td>
<td>[p] [b] [m]</td>
</tr>
<tr>
<td></td>
<td>&lt; p₂ &gt;</td>
<td>[p] [b] [m]</td>
</tr>
<tr>
<td>[f,v]</td>
<td>&lt; f₁ &gt;</td>
<td>[f] [v]</td>
</tr>
<tr>
<td></td>
<td>&lt; f₂ &gt;</td>
<td>[f] [v]</td>
</tr>
<tr>
<td>[t,d,n]</td>
<td>&lt; t₁ &gt;</td>
<td>[t] [d] [n]</td>
</tr>
<tr>
<td></td>
<td>&lt; t₂ &gt;</td>
<td>[t] [d] [n]</td>
</tr>
<tr>
<td>[s,z]</td>
<td>&lt; s₁ &gt;</td>
<td>[s] [z]</td>
</tr>
<tr>
<td></td>
<td>&lt; s₂ &gt;</td>
<td>[s] [z]</td>
</tr>
<tr>
<td>[l]</td>
<td>&lt; l₁ &gt;</td>
<td>[l] [l]</td>
</tr>
<tr>
<td></td>
<td>&lt; l₂ &gt;</td>
<td>[l] [l]</td>
</tr>
<tr>
<td>[r]</td>
<td>&lt; r₁ &gt;</td>
<td>[r] [r]</td>
</tr>
<tr>
<td></td>
<td>&lt; r₂ &gt;</td>
<td>[r] [r]</td>
</tr>
<tr>
<td>[k,g]</td>
<td>&lt; k₁ &gt;</td>
<td>[k] [k]</td>
</tr>
<tr>
<td></td>
<td>&lt; k₂ &gt;</td>
<td>[k] [k]</td>
</tr>
<tr>
<td></td>
<td>&lt; k₃ &gt;</td>
<td>[k] [k]</td>
</tr>
<tr>
<td>[f,y]</td>
<td>&lt; f₁ &gt;</td>
<td>[f] [y]</td>
</tr>
<tr>
<td></td>
<td>&lt; f₂ &gt;</td>
<td>[f] [y]</td>
</tr>
</tbody>
</table>

#### Table I

**Consonantal context-dependent visemes** (adapted from [7]).
TABLE II
Vocalic context-dependent visemes (adapted from [7]).

<table>
<thead>
<tr>
<th>Homophone Group</th>
<th>Visemes</th>
<th>Phonetic Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i,i]</td>
<td>&lt; i₁ &gt;</td>
<td>All contexts except [tit] and [jij].</td>
</tr>
<tr>
<td></td>
<td>&lt; i₂ &gt;</td>
<td>[tit] e [jij].</td>
</tr>
<tr>
<td>[e,ɛ]</td>
<td>&lt; e &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>&lt; e &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[a,ʌ]</td>
<td>&lt; a &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[o,ɔ]</td>
<td>&lt; o &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[a,u]</td>
<td>&lt; u &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[t]</td>
<td>&lt; t &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[ɔ]</td>
<td>&lt; o &gt;</td>
<td>All contexts.</td>
</tr>
</tbody>
</table>

provide redundant samples of visemes. From the digitized content of recorded corpus it was possible to extract audio and video tracks for each uttered element by the subject. From the video tracks, a collection of images corresponding to the frames of video were extracted. The corpus was recorded under NTSC video standard which generates video at 29.97 frames per second and resolution 720x486 pixels. Images were saved in Microsoft Windows BMP file format, without compression. The audio tracks were digitized as PCM (Pulse-Code Modulation) audio files, sampled at 48 kHz and using 16 bits/sample. Each audio track was manually analyzed in order to get its phonetic transcription, determining in timescale the frontier of each uttered phoneme. The information provided by timed phonetic transcription makes possible the association of phonemes and their corresponding visual realization from the extracted frames of the video.

B. Image Database Building

Figure 2 shows the process by which the extracted images from the raw corpus are pre-processed to constitute the image database.

In order to implement the reduced mode, 34 images were extracted from the audiovisual corpus, corresponding to: 22 visemes representing the consonantal visemes of Table I, 11 visemes representing vocalic visemes of Table II and one viseme corresponding to the silence posture. The images were selected through the visual inspection of audiovisual corpus and considering the extreme articulation point for an analyzed speech segment.

The following sections describe the steps of image database building.

C. Image Registration

The images extracted from corpus need to be aligned in order to make the face position and size uniform along the image database. This operation is necessary to correct natural movements made by the subject during the recording process.

A reference image is selected from the universe of available samples based on desirable characteristics such as a centralized position of the face in the frame, visual quality and absence of distortions due to fast movements.

In the next step, each image extracted from corpus and the reference image were manually processed in order to get coordinates information of the reference points shown in Figure 3: internal corners of eyes, nostrils and junctions of the ears with the head. These points are chosen because they suffer small influence from mouth and jaw movements during speech production.

From these information, it is possible to establish the correspondence mapping between images extracted from corpus and the reference image. The alignment is performed through a registration process where geometrical transformations (translation, scale and rotation) are applied to the images of database in order to align them with the reference image.

D. Extracting Region of Interest

After the alignment of the images, the system extracts from the images the region of interest, which is visually affected by the speech production and involves lips, jaw and the region around them. This operation reduces the size of images to be stored in the image database and is based on the proposed synthesis approach, where visemes are stitched to a base-face.

This process is done through the creation of a transparency mask that helps the extraction of the region around lips and
jaw, allowing the later superposition and fusion of the detached region into the base-face (Figure 4).

The base-face is a reference image from which the mouth and lips regions are discarded. In its place the synthesized images of final animation are stitched.

The mask has the two contours shown in Figure 4(a). Pixels inside the inner contour assume value one (full opacity), pixels that are outside the outer contour have value zero (full transparency), and the region between two contours have a gradient of pixel values ranging from zero to one (Figure 4(b)). This strategy is used to smooth the edges of the extracted region, improving the result of viseme and base-face combination. Figures 4(c) and (d) shows a sample image and the obtained result of mask application respectively.

E. Extracting information from visemes

In order to get additional information about the visual aspect of the captured visemes, a set of feature points is defined and manually measured (Figure 5). Each image of the database is accordingly labeled with such information.

This data is important not just to guide the positioning of visemes on base-faces but also as a tool to achieve smooth transitions during synthesis time.

F. Database

The implementation of an image database is based on the labeling of each stored sample that provides relevant information for the synthesis process.

The database is built by the analysis and consolidation of information derived from the phonetic transcription, the data extracted from images and the original video sequence (Figure 2).

Each viseme on the database is associated to the following information:
- viseme group classification;
- video frame identification;
- phonetic context of captured viseme;
- feature points information obtained during pre-processing step;
- pointer to image file.

G. Synthesis of Facial Animation

The implemented system is designed to be integrated to a text-to-speech synthesizer for Brazilian Portuguese, capable of providing to the facial animation system a timed phonetic transcription of the new content to be uttered. In this implementation, the system is integrated to the commercial TTS “CPqD Texto Fala”. This system includes an audio synthesis module that generates a synthetic speech signal from a sequence of phonetic units derived from the text to be uttered. The synthesizer selects the appropriate acoustic units from a natural speech database coupled to the system, producing a high quality synthetic speech.

The timed phonetic transcription provided as input to the 2D facial animation system is an intermediate step result of “CPqD Texto Fala” text processing and synthesis modules and is composed by the sequence of phonetic units used during the generation of speech signal and their corresponding durations.

The first step of animation synthesis process is to convert the sequence of phonemes determined by the phonetic transcription into a corresponding sequence of target visemes of Tables I and II.

In order to map a phoneme to a context-dependent viseme, it is necessary to analyze the phonetic context of each phoneme, specially for the consonantal visemes, taking into consideration the adjacent phonemes. As shown on Table I, the consonantal visemes are defined for phonetic contexts of type \( V_1CV_2 \), where \( V \) indicates a vowel and \( C \) a consonant. We also observe that \( V_1 \) is optional, that means the consonant can be preceded by silence. Consonantal phonemes present in the phonetic transcription that does not fit this type of context are analyzed and mapped to \( V_1CV_2 \) contexts based on the similarity of speech sounds and a list of exceptions. This is the case, for example, of consonantal clusters such as the one observed in the word “Brazil”. A similar approach is adopted for the first line of Table II.

Each speech phoneme has its duration expressed in the timed phonetic transcription. Therefore, it is possible to de-
fine the time instant associated to each target visemes and to determine the number of transition frames between two subsequent target visemes. The animation synthesis is then performed applying a warping between target visemes in order to obtain smooth transitions between them.

\[ H. \text{ Morphing between Visemes} \]

The application of image morphing technique between key poses of animation is a mechanism to generate a smooth and realistic transition.

Morphing between two images begins by establishing the correspondence map of feature primitives between both images. The correspondence map is then used to compute warping functions that define the spatial relationship between all points in both images. In this work, the feature primitives are the feature points detected during pre-processing step and shown in Figure 5.

Considering the small number of selected feature points, the determination of an warping function can be stated as a scattered data interpolation problem. In this work, radial basis functions (RBF) were adopted, since they are proven to be an effective tool in multivariate interpolation problems of scattered data [14, 15].

The determination of warping function begins with the definition of a correspondence map between source viseme feature points \( p_i \) and target viseme points \( q_i \), where \( i = 1, ..., n \) and \( n = 5 \) (Figure 5). In this notation \( p_i \) and \( q_i \) are position vectors of image pixels with coordinates \([p_{ix}, p_{iy}]\) and \([q_{ix}, q_{iy}]\), respectively. From the correspondence map, the following step is to determine the warping function, which formulation can be stated in terms of input and output:

- **Input:** \( n \) pairs \((p_i, q_i)\) of feature-points, where \( p_i \) and \( q_i \) are position vectors \( \in \mathbb{R}^2 \), \( i = 1, ..., n \) and \( n = 5 \).
- **Output:** an at-least-continuous function \( f : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \) with \( f(p_i) = q_i, i = 1, ..., n \).

The adopted interpolation function \( f \) is a linear combination of radial functions, which values depend only on the distance between each image point \( p \) and a feature point \( p_i \). This distance is the Euclidean distance between two points defined as

\[ d(p, p_i) = \sqrt{(p_{ix} - p_{ix})^2 + (p_{iy} - p_{iy})^2}. \]

\[ f(p) = b_m(p) + \sum_{i=1}^{n} \alpha_i g_i(d(p, p_i)) \tag{1} \]

In the equation 1, \( b_m(p) \) is a \( m \)-degree polynomial added to the formulation in order to reproduce affine linear transformations not naturally obtained by pure radial functions. In this implementation, \( m = 1 \) and the adopted polynomial is

\[ b_1(p_i) = \alpha_1 + \alpha_2 \cdot p_{ix} + \alpha_3 \cdot p_{iy}. \]

The \( \alpha \) coefficients are determined by solving the system of linear equations resulting from \( f(p_i) = q_i, i = 1, ..., n \) and the polynomial precision constraints [14].

The system determines the warping function using multiquadrics radial basis functions as pointed by [15] as an effective and time-efficient radial basis functions:

\[ g(d) = (d^2 + r^2)^\mu, r > 0, \mu \neq 0 \tag{2} \]

According to the analysis made at [15], \( \mu \) was set to 0.5 and an individual value of \( r_i \) was used for each \( p_i \), computed from the distance to the nearest neighbor:

\[ r_i = \min(d(p_i, p_j)), i = 1, ..., n; i \neq j \tag{3} \]

Given two visemes \( A \) and \( B \), the implemented algorithm determines the function that warps image \( A \) to \( B \) in forward direction and the function that warps \( B \) to \( A \) in the backward direction. Taking the number of frames to morph \( A \) to \( B \), two sequences of the corresponding number of intermediate images are generated performing forward and backward transformations. Once both images have been warped into intermediate feature positions, they are cross-dissolved to generate intermediate frames according to a transition control function [13].

\[ I. \text{ Transition control of morphing} \]

The morphing process implemented by the system was controlled by a smooth non-linear interpolation time function. This function controls the feature points trajectory between two key visemes. The adopted approach represents a better modeling of speech dynamics when compared to the linear transition approach.

The interpolation curve adopted was the Hermite parametric curve [16] with coefficients set to preserve geometric continuity and present derivatives equal to zero at the instant of realization of the articulatory targets represented by the visemes from image database.

The adopted interpolation curve is obtained through the equations:

\[
\begin{bmatrix}
  x(t) \\
  y(t)
\end{bmatrix} = \begin{bmatrix}
  p_{ix} & q_{ix} & 2 & -3 & 0 & 1 \\
  p_{iy} & q_{iy} & -2 & 3 & 0 & 0
\end{bmatrix} \begin{bmatrix}
  t^3 \\
  t^2 \\
  t \\
  1
\end{bmatrix} \quad 0 \leq t \leq 1
\]

where \( x(t) \) and \( y(t) \) are the coordinates of feature point; \( p_{ix} \) and \( p_{iy} \) are the articulatory target coordinates of the first phoneme; \( q_{ix} \) and \( q_{iy} \) are the articulatory target coordinates of the second phoneme; and \( 0 \leq t \leq 1 \) is the parametric variable.

\[ J. \text{ Extended Mode} \]

The described synthesis implementation based on morphing between visemes characterizes the reduced mode of the implemented 2D facial animation synthesis, capable of delivering videorealistic facial animations from a database of just 34 visemes.

The implemented system also foresees that the image database can continuously grow and that the synthesis system should be able to adapt itself to deliver higher levels of videorealism through the combination of morphing between visemes approach and the concatenation of small video fragments originally captured from audiovisual corpus.

The database can grow by adding new images that are preprocessed and analyzed according to the steps described on Sections III-C, III-D and III-E. The new images added to the database are organized as additional viseme samples of the
visemes categories shown on tables I and II and also keeps the information of their original frame identification on the audiovisual corpus.

If an extended image database is available, the system is designed to, for each viseme transition, search the database in order to find a similar transition captured in the original corpus. The optimal situation will happen when such transition occurs in the corpus in exactly the same phonetic context and with the same duration in number of frames. In this case, the sequence of animation frames selected from database are simply concatenated to the final animation sequence. Otherwise, the target phonetic context of the final animation, characterized by a triphone, is analyzed against their pairs on the same homophone groups. In this case, the target speech segment “apa” can be visually animated by a sequence of frames corresponding to the speech segment “aba” if they are present in image database (see first line of table I, that shows that phonemes [p] and [b] are in the same homophone group). Following the context-dependent visemes mapping, the extended mode always tries to minimize the number of viseme transitions that requires morphing. With this approach the system is designed to deliver higher levels of videorealism and can be applied on systems with higher processing and memory capacity, such as high end desktops.

IV. CONCLUSION

This paper presented a videorealistic 2D speech synchronized facial animation system for Brazilian Portuguese designed to be a cross-platform solution that can be adapted to characteristics of the executing platform through two different synthesis modes: a reduced mode and an extended mode.

The paper focused on the description of reduced mode implementation, in which the synthesis is exclusively based on morphing between visemes approach and the image database is composed by 34 visemes. Our implementation presents two significant contributions to the morphing between visemes synthesis approach, that result in a more accurate modeling of visual speech. First, the system adopts a straightforward modeling of coarticulation effects through the use of context-dependent visemes. Secondly, the system applies a simple nonlinear transition function that drives the morphing process. The reduced image database and the non-complex models used to implement the synthesis process turn possible the embedding of such system on small or portable devices with limited processing and memory capacities (Figure 6).

Preliminary speech intelligibility test results made with animations generated under reduced mode show that the level of videorealism obtained by the system is fair enough to greatly improve the comprehension of information when the audio is heavily degenerated by noise. This work is still under development and the next activities related to the project are: the full implementation of synthesis under extended image database, the implementation of automatic image processing algorithms to process image database and the conclusion of speech intelligibility tests with the system under extended mode.

The authors would like to thank CPqD for the financial and technical support and Carolina Franciscañeglis for her contribution on audiovisual corpus recording process.

REFERENCES