Context-dependent visemes: a new approach to obtain realistic 2D facial animation from a reduced image database

Paula Dornhofer Paro Costa  
Dept. of Comp. Eng. and Industrial Automation  
School of Electrical and Computer Engineering  
University of Campinas  
Campinas - SP - Brazil  
paula@dca.fee.unicamp.br

José Mario De Martino  
Dept. of Comp. Eng. and Industrial Automation  
School of Electrical and Computer Engineering  
University of Campinas  
Campinas - SP - Brazil  
martino@dca.fee.unicamp.br

Abstract—The arise of a new reality where information is accessed ubiquitously from a variety of different devices demands the development of more efficient and intuitive human-machine interfaces. In this context, speech synchronized facial animation research aims the development of interactive talking-heads capable of reproducing the natural and intuitive face-to-face communication mechanisms. This paper describes a novel approach to the image-based morphing between visemes synthesis strategy through the use of context-dependent visemes capable of modeling speech coarticulation effects. The proposed approach was validated through the implementation of a videorealistic 2D speech synchronized facial animation system for Brazilian Portuguese based on an image database of just 34 photographs and integrated to a text-to-speech (TTS) synthesizer. Speech intelligibility tests results show that the level of videorealism obtained by the system is capable of improving message comprehension when audio is degenerated by noise, with results similar to the obtained with a video of a real face. The proposed methodology can be adapted to deliver photorealistic animations through the appropriate processing, sequencing, concatenation and presentation of images captured from a real face.

I. INTRODUCTION

The expansion of mobile communications and the technological evolution of portable devices, allowed the ubiquitous access to information by an increased number of users that are not necessarily familiarized with traditional human-computer interfaces.

Facial animation systems synchronized with speech, or talking-heads, constitute an enabling technology for the development of human-machine interfaces capable of reproducing the natural and intuitive face-to-face communication mechanisms, turning applications more human, involving and attractive, being an interesting alternative to WIMP (Windows, Icons, Menus and Pointing devices) interfaces [1], [2], [3].

On the other hand, 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 is advantageous in situations where the user pays for the volume of transmitted data [4].

In order to awake reliability and trust, synthetic talking heads have to deliver satisfactory levels of videorealism, reproducing the look and fell of the video of a real person. In this context, image-based, or 2D, facial animation approach is considered an alternative to the model-based, or 3D, face modeling. The 2D approach inherently generates photorealistic animations through the appropriate processing, sequencing, concatenation and presentation of images captured from a real face.

This work describes a new approach to the 2D morphing between visemes synthesis strategy. Typically, 2D speech synchronized animation systems based on morphing between visemes are characterized by a small image database and a low level of videorealism since they are not able to properly model the coarticulation effects that arise when the typical articulation pattern of a speech segment is modified by the interaction with nearby segments.

In this work, the modeling of coarticulation effects is done applying an efficient and straightforward approach based on context-dependent visemes identified for Brazilian Portuguese [5]. Such modeling was applied in the implementation of a 2D facial animation system, integrated to a text-to-speech synthesizer for Brazilian Portuguese, making possible to synthesize videorealistic animations from just 34 photographs representing different context-dependent visemes.

The paper also presents the results of speech intelligibility tests using animations generated by this system. The test results show that the obtained facial animation videorealism is capable of greatly improving message comprehension when audio is degenerated by noise and presents results similar to the video of a real face.

The main contributions of this work comprehends: the use of context-dependent visemes in a morphing between visemes synthesis strategy, the implementation of a videorealistic 2D speech synchronized facial animation system for Brazilian Portuguese and the application of speech intelligibility tests to evaluate the level of videorealism obtained by
the animation.

Despite the fact that our system is being developed for Brazilian Portuguese, the underlying concepts can be straightforward applied to other languages.

The remaining sections of this paper present the related work and highlights the underlying principles of the proposed approach describing the implemented talking-head system. A detailed description of this approach and the implemented system can be found in [6].

II. RELATED WORK

The implementation of 2D facial animation systems starts with a recorded audiovisual corpus of a real face, from which is possible to extract a set of images spanning a range of mouth shapes corresponding to visemes. The captured visemes are processed to create an image database from which frames are processed and selected to synthesize the final animation.

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. From the related literature it is possible to identify three different approaches for the synthesis of 2D facial animation.

The first approach is exclusively based on morphing between key pose face images, or visemes. Such systems are characterized by the use of a tiny image database in detriment of the videorealism. In order to reduce the image database size, the modeling of visible speech articulatory movements is overly simplified and, typically, the coarticulation effects are not properly modeled [7], [8], [9].

The second approach characterizes a concatenative visual speech synthesis strategy. This approach is based on appropriate selection and concatenation of video fragments stored in a database built from existing audiovisual corpus. These systems are able to model coarticulation effects capturing the dynamics of speech production for specific contexts present on the audiovisual corpus. A representative work of this approach is Video Rewrite [10].

Finally, is possible to identify a third synthesis strategy where final animation is obtained through the identification of a suitable trajectory of images on a database, combined with the generation of novel images [11], [12]. These systems are based on a refined preprocessing step of a corpus from which images are extracted and analyzed in order to parameterize the image database space. The final animation is obtained through sophisticated algorithms used to determine a trajectory in this space in order to reproduce coarticulation effects and guarantee smooth transitions and videorealistic results.

The system described in this paper implements a new approach to the morphing between visemes synthesis strategy capable of modeling the coarticulation effects. The main contribution of this approach is to improve the final animation videorealism while keeping the image database small.

III. CORPUS AND IMAGE DATABASE BUILDING

A. Defining 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 defined in order to make possible to capture visual samples of occurrences of context-dependent visemes for Brazilian Portuguese as identified in [5]. This study identified homophone groups that are groups of phonemes in different phonetic 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 [13].

<table>
<thead>
<tr>
<th>Homophone Group</th>
<th>Visemes</th>
<th>Phonetic Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p,b,m]</td>
<td>[p1 &gt;</td>
<td>[pi [pa] [pi] [ipe] [ipu]</td>
</tr>
<tr>
<td></td>
<td>[p2 &gt;</td>
<td>[pu [apu] [aup] [apu] [uap]</td>
</tr>
<tr>
<td>[f,v]</td>
<td>[f1 &gt;</td>
<td>[fi [fa] [fi] [ife]</td>
</tr>
<tr>
<td></td>
<td>[f2 &gt;</td>
<td>[fu [afu] [auf] [auf]</td>
</tr>
<tr>
<td>[t,d,n]</td>
<td>[t1 &gt;</td>
<td>[tu [it] [itu] [itu]</td>
</tr>
<tr>
<td></td>
<td>[t2 &gt;</td>
<td>[ta [ait] [ait] [ait]</td>
</tr>
<tr>
<td>[s,z]</td>
<td>[s1 &gt;</td>
<td>[si [isa] [is]</td>
</tr>
<tr>
<td></td>
<td>[s2 &gt;</td>
<td>[su [us] [usu]</td>
</tr>
<tr>
<td>[l]</td>
<td>[l1 &gt;</td>
<td>[li [ili] [il]</td>
</tr>
<tr>
<td></td>
<td>[l2 &gt;</td>
<td>[la [ail] [ail]</td>
</tr>
<tr>
<td>[n]</td>
<td>[n1 &gt;</td>
<td>[ni [in] [in]</td>
</tr>
<tr>
<td></td>
<td>[n2 &gt;</td>
<td>[na [ain] [ain]</td>
</tr>
<tr>
<td>[k,g]</td>
<td>[k1 &gt;</td>
<td>[ki [iki] [ik]</td>
</tr>
<tr>
<td></td>
<td>[k2 &gt;</td>
<td>[ka [aik] [aik]</td>
</tr>
<tr>
<td>[r,y]</td>
<td>[r1 &gt;</td>
<td>[ry [iry] [ir]</td>
</tr>
<tr>
<td></td>
<td>[r2 &gt;</td>
<td>[ry [ir] [ir]</td>
</tr>
</tbody>
</table>

Table I

CONSONANTAL CONTEXT-DEPENDENT VISEMES (ADAPTED FROM [5]).
Table II  
VOCALIC CONTEXT-DEPENDENT VISEMES (ADAPTED FROM [5]).

<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 [ji].</td>
</tr>
<tr>
<td></td>
<td>&lt; i₂ &gt;</td>
<td>[tit] e [ji].</td>
</tr>
<tr>
<td>[e, e]</td>
<td>&lt; e &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[r]</td>
<td>&lt; r &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[a, æ]</td>
<td>&lt; a &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[ɔ]</td>
<td>&lt; o &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[o, o]</td>
<td>&lt; o &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[u]</td>
<td>&lt; u &gt;</td>
<td>All contexts.</td>
</tr>
<tr>
<td>[ʊ]</td>
<td>&lt; ʊ &gt;</td>
<td>All contexts.</td>
</tr>
</tbody>
</table>

From the digitized content of recorded corpus it was possible to extract the 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.

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 the phonetic transcription in time makes possible the association of phonemes and their corresponding visual realization from the extracted frames of the video.

B. Image Database Building

The images extracted from the raw corpus needed to be preprocessed to constitute the image database.

The first step was to align the images with the goal of making the face position and size uniform along the samples of database. This operation is necessary due to natural movements made by the subject during the recording process and it was treated as a registration problem between captured samples and a reference image. The registration was performed considering the regions of face that have small influence of mouth and jaw movements during speech production such as eyes, nose and ears that are used as reference elements to determine the geometrical transformation to perform the registration. A small set of fiducial points were manually defined as an input to the automatic registration algorithm implemented [6].

After the alignment of the images, it was extracted from the images the region of interest that is visually affected by the production of speech and involves lips, jaw and the region around them. This operation results in smaller images being stored in image database and is based on the synthesis approach where visemes are stitched to a base-face.

In this approach, a base-face image has its mouth and jaw region discarded and in its place the visemes from image database are applied. This operation 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 1). The transparency mask has a gradient of pixels values that makes possible to smooth the edges of the extracted region, improving the result of viseme and base-face combination. Figures 1 (c) and (d) shows a sample image and the obtained result of mask application respectively.

![Figure 1](image1.png)

Finally, in order to get additional information about the visual aspect of the captured visemes, a set of feature points is defined and determined on the captured images and the corresponding information is accordingly stored in the database.

From the set of feature points it is possible to determine the distance between mouth corners, the position of superior and inferior lips and position of jaw. 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. Figure 2 illustrates the detected feature points.

![Figure 2](image2.png)
The final implementation of the image database is based on the labeling of each stored sample with the information about its feature points.

IV. Animation Synthesis

A. Audio

In order to implement a talking head, the 2D facial animation system is driven by the phonetic transcription of the content to be uttered. In this work, the system was integrated to the commercial text-to-speech synthesizer for Brazilian Portuguese “CPqD Texto Fala”. This system includes an audio synthesis module that generates a synthetic speech signal from a sequence of phonetic units extracted from a text and from a natural speech database coupled to the system. The synthesizer selects the appropriate acoustic units from its database and process them adding rhythm and intonation, producing a high quality synthetic speech.

The phonetic transcription provided as input to the 2D facial animation system is an intermediate step result of the TTS synthesis module and is composed by the sequence of phonetic units used during the generation of speech signal and their corresponding durations.

B. Selecting Visemes for Final Animation

The animation synthesis process begins with the conversion of the sequence of phonemes determined by the phonetic transcription into a corresponding sequence of target visemes of Tables I and II. According to these tables, the image database is composed by just 34 visemes, 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.

In order to map a phoneme to a context-dependent viseme, it is necessary, specially for the consonantal visemes, the analysis of the phonetic context of each phoneme, taking into consideration which phonemes are adjacent to it. 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.

C. Morphing between Visemes

Each speech phoneme has its duration determined by the phonetic transcription from which is possible to define the time instant to which the key visemes are associated and the number of transition frames between two subsequent key-poses. The animation synthesis is performed applying a morphing between key visemes as a mechanism to obtain smooth and realistic transitions between them.

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 shown in Figure 2.

Considering the small number of selected feature points, the determination of an warping function was stated as a scattered data interpolation problem. In this case, 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 complete mathematical formulation is detailed in \[6\].

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.

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.

V. Speech Intelligibility Tests

Assessing the quality of an animation is a key aspect to compare and evaluate different synthesis approaches and for tracking any progress on a developing system. However, determining the quality of an animation in an objective way is quite difficult because no universally accepted criteria exist for this purpose. The standard approach to assessing naturalness of talking heads are based on subjective tests where human observers express their opinions about aspects such as “synchronicity”, “smoothness” and “precision” of the animations. This approach however, usually requires specific orientation to people to correctly judge these criteria and are also subject to discrepancies caused by individual reactions, since a viewer can be influenced by the face characteristics (a viewer may dislike the face “hairstyle” or consider a face “friendly” or “arrogant”).

In this work we assessed the effective contribution to speechreading of context-dependent visemes approach. The
contribution to speechreading is an objective measurement of speech synchronized facial animation quality. The benchmark procedure used is derived from the oral speech intelligibility test proposed by Sumby and Pollack [17]. Essentially, the intelligibility test involves the presentation of audio signals with, and without, supplementary visual information, conducted under different conditions of acoustic degradation.

The carried out intelligibility test consisted on the presentation of a set of 27 non-sense words or logatomes conveyed in a vehicle phrase uttered by a Brazilian female speaker and previously recorded in the audiovisual corpus. The adopted vehicle phrase had the following structure: “Ela fala <logatome>” (“She says <logatome>”). The vehicle phrase was devised as a mean to get the test subject’s attention prior to the logatome utterance. The recorded audio was contaminated with noise in order to produce three different Signal-to-Noise-Ratio (SNR) conditions: -12dB, -18dB and -24dB.

Recorded video fragments from the audiovisual corpus were used which audio tracks were detached and used as the base material for the generation of new versions with different SNRs, resulting a total of 81 (27 x 3) audio files. The degradation of the acoustic signal was realized through the addition of white noise.

The original audio tracks, not degraded, were manually segmented and the resulting timed phonetic transcriptions were used to drive the 2D facial animation system, which generated frames with dimensions 720x486 (NTSC). Copies of the degraded audio files were then properly (re)synchronized with the corresponding real speaker video and with the facial animation frames.

The resulting 243 (81 x 3) files were organized for presentation in 9 groups, each containing the 27 different logatomes (and the associated vehicle phrase) with the same audio quality. The file contents of these groups were presented to the test subjects in a random order. Furthermore, in each group the presentation order of the 27 logatomes was also randomly varied.

A group of 40 subjects were invited to participate in the intelligibility test. This group of individuals consisted of normal-hearing people with age ranging from 18 to 54 years.

After presentation of a vehicle phrase, each subject was asked to indicate the understood logatome checking one of the 28 available options, consisting of the 27 logatomes and a NDA (none of the alternatives) option. The subjects were strongly encouraged to guess even if they were in doubt between different logatome options, and only to choose the NDA (none of the alternative) option when they had absolutely no clue about the logatome presented or they had “heard” something else not provided as one of the logatome options.

The results of the intelligibility test are summarized in Figure 3. The figure shows the percentage of correct responses to the logatome stimulus.

As could be expected, it can be seen that the visual information provided by both real and virtual speaker increases the speech intelligibility.

In particular, it is possible to state that facial animation can be considered an efficient visual aid to improve speech intelligibility in situations where audio suffers degeneration. This can be clearly seen in the worst audio quality condition tested, at SNR -24 dB, where the intelligibility gain obtained by the facial animation is 23% compared to the audio only situation, against 37% obtained by the real video.

Comparing to the audio only situation, the intelligibility gain provided by the inclusion of the facial animation at SNR -18 dB is 33% and the corresponding gain of the real speaker’s video amounts to 40%. At SNR -24 dB, the intelligibility gain obtained by the facial animation is 23%, against 37% obtained by the real video.

Further analysis were driven from the analysis of variance (ANOVA) between the groups. The analysis of variance for the three groups at the three different values of SNR values results that the percentage of correct responses for the audio only and audio accompanied by visual information situations can be considered statistically discernible with a statistic significance $p < 0.001$.

However, the analysis between facial animation and real video groups at SNR -12 dB and SNR -18 dB cannot be considered statistically distinguishable since their fluctuations can be confused. In other words, considering the correct responses given by the subjects, it is not possible to statistically distinguish the intelligibility levels reached by the facial animation and real video. These fluctuations can be illustrated, for example, for SNR -12 dB level by the boxplots in figure 4. On the boxplots, each box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with
values beyond the ends of the whiskers and are indicated by crosses.

The same situation is not observed at the -24 dB SNR level. In this case, the facial animation generated, presents statistically significant difference in the contribution to the speech intelligibility in noise when compared to the real video.

VI. CONCLUSION

This paper presented a new approach to model coarticulation effects and improve videorealism on 2D facial animation systems based on the traditional morphing between visemes synthesis strategy. The use of context-dependent visemes is the key aspect of the proposed approach. The paper emphasized the description of the methodology to build an image database and also presented an implementation of the proposed synthesis strategy, providing a complete overview of the implemented talking head system.

The described methodology was implemented in a 2D facial animation system for Brazilian Portuguese integrated to a TTS. The main characteristics of the system are the videorealism, through the presentation of visible articulatory movements including the coarticulation effects, and the reduced image database size that makes it also suitable for limited memory and processing platforms.

Speech intelligibility test result showed that, the facial animation generated following the proposed approach is able to improve speech intelligibility in conditions where audio is degenerated by noise, presenting results similar to real video in those signal to noise ratio conditions where residual speech information can be distinguished from noise.

ACKNOWLEDGMENT

The authors would like to thank CPqD–Telecommunications Research and Development Center, for both technical and financial support to the project.

REFERENCES


