The Development of a Brazilian Talking Head

José Mario De Martino & Fábio Violaro

Abstract— **This paper describes partial results of a research, in progress at the School of Electrical and Computer Engineering of the State University of Campinas, aimed at developing a realistic three-dimensional Brazilian Talking Head. Through an extensive analysis of a video-audio linguistic corpus, a set of 29 phonetic context-dependent visemes (22 consonantal plus 7 vocalic visemes), that accommodates perseverative and antecipatory coarticulation effects, was established for Brazilian Portuguese. Visemes are the recognizable visual motor patterns associated to one or more speech segments. A viseme permits the visual identification of the segment group to which a speech sound belongs, providing the visual cues for lipreading. The proposed visemes were used to drive the visible articulatory movements (temporomandibular joint and lips) of a three-dimensional virtual head. The quality of the resulting talking head was assessed by speech intelligibility tests performed in the presence of high noise levels. The results show a clear improvement of the speech intelligibility using the talking head in comparison with the audio alone presentation. This is a first implementation trial, and new refinements must be provided in a near future in order to improve the naturalness of the speech synchronized facial animation.**

Index Terms— **embodied conversational agents, speech synchronized facial animation, talking head.**

Resumo— **Este artigo descreve resultados parciais de pesquisa em desenvolvimento na Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas, voltada ao desenvolvimento de uma cabeça falante do Português do Brasil. Através de uma extensa análise de corpus lingüístico composto de material audiovisual, um conjunto de 29 visemas dependentes do contexto fonético (22 visemas consonantais e 7 visemas vocálicos), que contempla os efeitos da coarticulação antecipatória e perseveratória, foi estabelecido para o português do Brasil. Visemas são padrões visuais de movimentação articulatória associados a um ou mais segmentos de fala. Um visema permite a identificação visual do grupo ao qual o segmento pertence, fornecendo as pistas visuais necessárias à leitura labial. Os visemas propostos foram usados para controlar os movimentos articulatórios visíveis (junta temporomandibular e lábios) de uma cabeça falante tridimensional. A qualidade da movimentação articulatória da cabeça falante foi avaliada através de testes de inteligibilidade da fala efetuados com a adição de um nível elevado de ruído ao sinal de áudio. Os resultados mostram uma evidente melhora de inteligibilidade da fala ao se apresentar o áudio mais a cabeça falante em comparação com a apresentação apenas do áudio. Esta é uma primeira versão de implementação, e novos refinamentos devem ser efetuados em futuro próximo com o objetivo de melhorar a naturalidade da animação facial produzida.**

Palavras chave— **agentes conversacionais personificados, animação facial sincronizada com a fala, cabeça falante.**

J. M. De Martino (martino@dca.fee.unicamp.br) and F. Violaro (fabio@decom.fee.unicamp.br) are faculty members of the School of Electrical and Computer Engineering of the State University of Campinas - FEEC/Unicamp. Av. Albert Einstein, 400 - Campinas - SP - Brazil - ZIP 13083-970.

I. Introduction

During the last decade, many speech related technologies have evolved and become a reality in the form of text-to-speech conversion systems (TTS), automatic speech recognition (ASR), and speech-based systems. Simultaneously the virtual face animation (talking-head) technology has also experienced improvements and advances. Nowadays all these technologies are being combined in order to forge embodied conversational agents. An embodied conversational agent is a virtual character capable of carrying on conversations with humans by both understanding and producing speech and facial expressions. Such agent can make human-computer interaction more similar to the human-human dialogue, we are all well familiar with. For example, an user formulates questions by voice (speech signal). The speech signal is processed by an ASR and a dialogue system that, directed by the keywords, access a database to retrieve the required text information. The talking head returns the retrieved information by speech, synthesizing the text by means of a TTS and providing articulatory movements synchronized with the synthesized speech.

Many applications can benefit from these emergent technologies. As application examples incorporating the aforementioned technologies, in an isolated or combined way, we can mention:

1) Help to handicapped people. A paraplegic person can command many house operations like opening or locking a door, turning on or off the lights or the television set, dictating a text or e-mail to a microcomputer, accessing a telephone, etc. A hearing impaired person can read a telephone call converted into text, can learn speechreading comparing the text with the synchronized articulatory speech movements of a talking head. A blind person can hear a book read by means of a TTS.
2) For not handicapped people, many applications are already currently available or being introduced: access to bank-services, ticket reservations, reception of an electronic newspaper or e-mail by means of a TTS while driving or making another activity. Also, for specific tasks, many professionals can fill repetitive formularies by using dictation machines (ASR) with a limited vocabulary and a specific grammar.
3) For training and educational purposes we can use a conversational agent to teach a new language, to train articulatory movements and to improve literacy, through story listening.

This paper focuses the implementation of a talking head for the Brazilian Portuguese language synchronized to a speech signal produced by a TTS [1] or by a real speaker. In the future we intend to merge all the above technologies into a Brazilian conversational agent.

II. Speech Synchronized Facial Animation

The architecture of the implemented speech synchronized facial animation system is presented in Figure 1.

The inputs of the system are a speech audio file and its timed phonetic transcription. Currently, the system accepts RIFF Waveform format (.wav) files as audio. The timed phonetic transcription is a text file that specifies the sequence of phones and their durations, associated to the utterance in the audio file. The timed phonetic transcription can be generated by either manual or automatic phonetic segmentation of the utterance or by a TTS synthesis system. The system has been tested with both manually segmented audio and a TTS synthesis system.

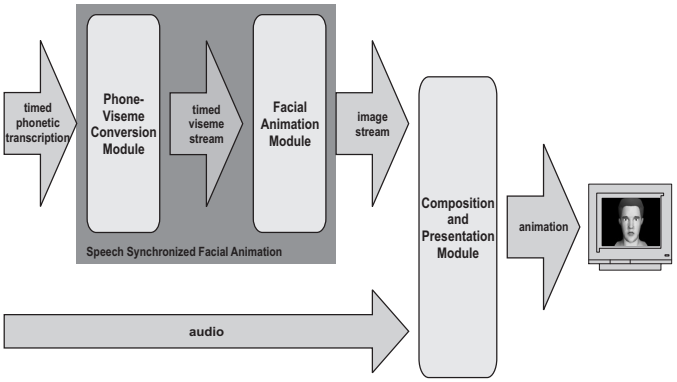

Fig. 1. Speech synchronized facial animation system architecture.

The task of the speech synchronized facial animation system is the generation of a sequence of images that, after being properly synchronized with the audio and presented in an adequate frame rate, gives the illusion of motion. The core of the system, as depicted in Figure 1, is composed by the *Phone-Viseme Conversion Module* and the *Facial Animation Module*. The *Phone-Viseme Conversion Module* translates the sequence of timed phones into a sequence of timed visemes, where visemes are the distinctive visual patterns associated to a group of phonemes (homophenes). The *Facial Animation Module* uses the timed viseme stream to control the virtual face. The main task of this module is to reproduce on the virtual face the articulatory movements expressed by the visemes.

The final processing step of the system entails the composition of the audio with the sequence of images generated by the *Facial Animation Module*.

III. Video and Audio Database

In order to define a set of context-dependent visemes and to measure the articulatory movements associated to the production of these visemes, a video plus audio database was created. A 22 years old student, born and raised in the city of São Paulo, Brazil, was chosen as the reference speaker. Four fiduciary points were marked on the speaker's face in order to measure the articulatory movements: a point on the middle of the upper lip (P_1), a point on the middle of the lower lip (P_3), a point on the left cormer of his mouth (P_2) and a point on the tip of his chin (P_4), as shown in Figure 2. Two synchronized video cameras were positioned 90^0 apart in front of the speaker. A microphone was positioned at nearly one meter from the speaker in a recording studio ambient. The objective was to measure the articulation movements of the mouth, lips and chin associated to the production of the different speech sounds. A special helmet (see Figure 3) was used to define a steady reference system attached to the head, despite of involuntary head movements produced during video recording. The visible articulatory movements were measured from video using stereo photogrammetric techniques [2].

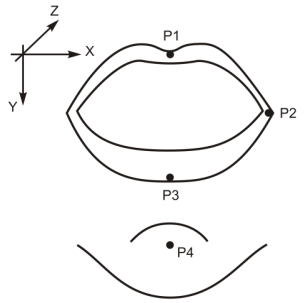

Fig. 2. Coordinate system and fiduciary points.

The audio material consisted of a set of logatomes that is described in the next section. The speech material was presented on the screen of a portable computer positioned in front of the speaker. The recording section lasted 3 hours. During this time the speaker uttered 2100 logatomes of the type $'C_1V_1C_2V_2$, where C stands for consonant, V stands for vowel and the symbol ' indicates that the first syllable is a tonic one. But, as the computational work to determine the articulatory movements is very time consuming, currently only a fraction of the recorded material (102 logatomes) was processed. These 102 logatomes were submitted to a manual segmentation in order to support the calculation of the articulatory targets that will be discussed in the next section. A full description of the database generation can be found in [3].

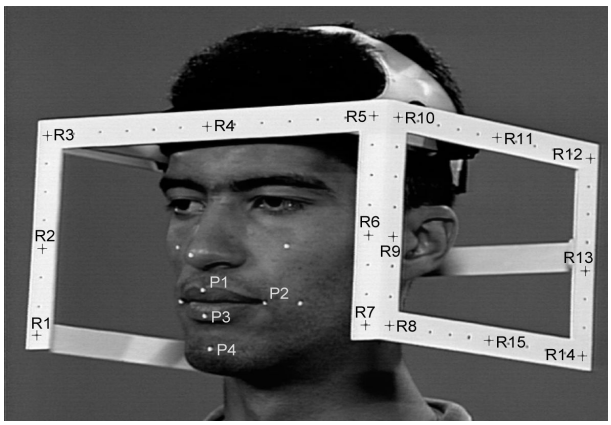

Fig. 3. Sample frame of the left side camera.

After the recording session, the synchronized audio and video signals were digitized. The original video signal consisted of a 29.97 Hz frame rate and was stored on Mini-DV format. The audio material was manually segmented (choice of a frontier between different phones) and labelled.

This segmentation was accomplished by using a spectrogram analysis. Due to the 33 ms between different image frames, a quantization error of 16 ms can be found between the audio and video frontiers.

IV. Context-Dependent Visemes

Visemes are recognizable visual motor patterns usually common to two or more speech sounds [4]. Sounds that appear alike on the face and cannot be distinguished by visual cues alone are called homophenes [5]. The adopted grouping of the Brazilian Portuguese consonants into homophenes was: [p,b,m], [f,v], [t,d,n], [s,z], [l], [ʃ,ʒ], [ʎ,ɲ], [k,g], [ɾ], and [ɣ] using the notation of the International Phonetic Alphabet (IPA). For vowels the following vocalic homophene grouping was considered: [i,ĩ], [e,ẽ], [ɛ], [a, ɐ̃], [ɔ],[o,õ], [u, ũ], [ɪ], [ɐ], [ʊ]. As the visemes are visually distinguishable articulatory movements, each group of phones (homophenes) is represented by a same viseme.

In respect to facial animation, a viseme can be characterized by:

- The geometric description of the articulatory targets that express the vocal tract conformation for the speech segment production;
- The instants of realization of the articulatory targets;
- The transition between articulatory targets taking into account the effects of coarticulation.

In order to characterize a viseme set for Brazilian Portuguese fulfilling the above requirements, we measured from the video sequences the three-dimensional trajectory of the 4 reference points marked on a real speaker face during the production of a corpus composed of a set of logatomes. As each reference point has 3 coordinates (x, y, z), a position vector with dimension 12 is associated to each image frame. A subset of the previous database corpus, composed of 102 logatomes, was used for the context dependent viseme set determination: 81 logatomes of type $'CV_1CV_2$, where C=[p,t,k,f,s,ʃ,l,ʎ,(ɣ)ɾ], V_1=[i,a,u] and V_2=[ɪ,ɐ,ʊ]; and 21 logatomes of type $'V_1V_2$, where V_1=[i,e,ɛ,a,ɔ,o,u] and V_2=[ɪ,ɐ,ʊ].

In the $'CV_1CV_2$ logatomes, as the [ɾ] does not occur at the beginning of a word in Portuguese, just the $'[ɣ]V_1[ɾ]V_2$ logatomes were used for treating the segments [ɾ] and [ɣ]. It is also important to note that in the C and V groups, only one representative of each homophene was used.

The captured data was processed in order to find representative articulatory targets for each viseme and the relative time when the articulatory targets are reached. For this task, the trajectories of the four fiduciary points (P_1 to P_4) of a given context-dependent viseme were calculated, considering all the image frames in a logatome utterance. For each point trajectory, it was calculated the dimension that suffers the largest variation (x, y or z), that is, the dominant direction, and the instant of derivative zero in the trajectory associated to this direction. The point of derivative zero was considered as the instant when the articulatory target was achieved, since after that instant it starts changing in order to reach the subsequent articulatory target. In Figure 4 it is shown the x, y and z coordinate variations of fiduciary point P_4 during the production of logatome [ˈpapɐ]. In this case the y coordinate is the one with the largest variation (dominant direction). The instants of derivative zero in the dominant direction were then associated also to the other dimensions (x and z in this example) and the resulting displacements were considered as the articulatory targets in these dimensions too. Note that these points in the non-dominant directions are also very close to instants of derivative zero.

The vertical lines in Figure 4 delimit the duration of each phone. Note in the figure that, as the phone [p] is a plosive, the instant of lips closure, which characterizes the segment and articulatory target, occurs one frame before the acoustic prodution. The instant of derivative zero associated to each context-dependent viseme was then normalized relative to the viseme duration and a mean value considering all realizations of this context-dependent viseme in all logatomes was calculated. This mean value was considered as the instant when the articulatory target was reached in the production of the considered context-dependent viseme. A complete table with the relative occurrence instant of each context-dependent viseme can be found in [3].

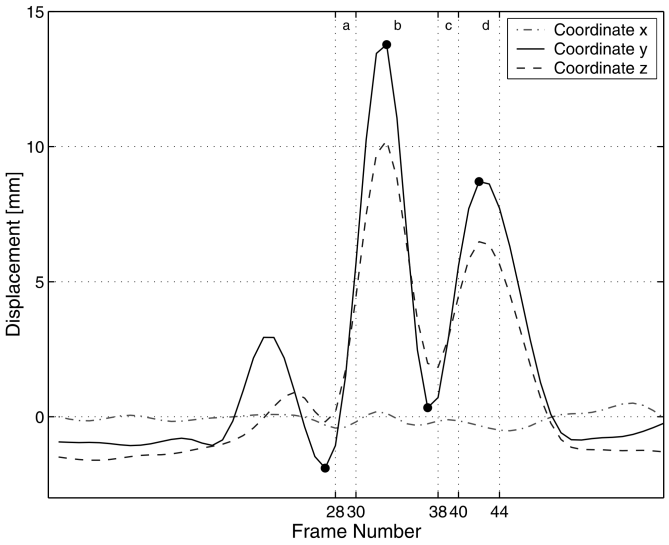

Fig. 4. x, y and z coordinate variation of fiduciary point P_4 during the production of logatome [ˈpapɐ].

After calculating the articulatory targets of each viseme, the x, y and z coordinates of the four fiduciary points associated to this target (vector of dimension 12), together with he information of its predecessor and sucessor visemes, were then clustered in a interactive partitional clustering based on the K-Means algorithm [6]. In this clustering phase, we systematically varied the target number of clusters, run the K-Means algorithm 40 times starting from different initial partitions and selected the clustering result with the smallest Davies-Boulding index. The final number of clusters obtained in the analysis of a given viseme establishes the number of contexts to be considered for this viseme and the cluster centroids represent the articulatory targets. As the analysis and clustering considered not only the phone produced, but also its predecessor and sucessor, the viseme set established is context-

dependent and reflects adjacent coarticulation effects.

Table I and Table II present our set of consonantal and vocalic context-dependent visemes respectively.

TABELA I

CONSONANTAL VISEMES AND THEIR PHONETIC CONTEXTS.

Symbol	Context	Homophene
<p1>	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]	[p, b, m]
<p2>	[pu] [upɪ] [upʊ]	
<f1>	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]	[f, v]
<f2>	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]	
<t1>	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]	[t, d, n]
<t2>	[ta] [atɐ]	
<s1>	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]	[s, z]
<s2>	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]	
<l1>	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]	[l]
<l2>	[la] [ilɐ] [alɪ] [alɐ]	
<l3>	[lu]	
<l4>	[ilʊ] [ulʊ]	
<ʃ1>	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]	[ʃ, ʒ]
<ʃ2>	[ʃu] [uʃʊ]	
<λ1>	[λi] [λa] [iλɪ] [iλɐ] [aλɪ] [aλɐ]	[λ, ɲ]
<λ2>	[λu] [uλɪ] [uλɐ]	
<λ3>	[iλʊ] [aλʊ] [uλʊ]	
<k1>	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]	[k, g]
<k2>	[ka] [akɐ]	
<k3>	[ku] [ikʊ] [akʊ] [ukʊ]	
<ɾ1>	[ɣi] [ɣa] [iɾɪ] [iɾɐ] [aɾɪ] [aɾɐ] [uɾɐ]	[ɣ], [ɾ]
<ɾ2>	[ɣʊ] [iɾʊ] [aɾʊ] [uɾɪ] [uɾʊ]	

TABELA II

VOCALIC VISEMES AND THEIR PHONETIC CONTEXTS.

Symbol	Context	Homophene
<i1>	all contexts but [tit] e [ʃiʃ]	[i, ĩ]
<i2>	[tit] [ʃiʃ]	
<a>	all context	[a, ɛ, ɐ̃]
<u>	all contexts	[u, o, õ, ũ]
<ɪ>	all contexts	[ɪ]
<ɐ>	all contexts	[ɐ, e, ɔ, ẽ]
<ʊ>	all contexts	[ʊ]

Given the articulatory targets of all visemes associated to a given utterance and their relative instants of occurrence, it is possible to provide a smooth interpolation between the adjacent targets positioned at their relative realization instants [7], [8]. This interpolation is provided mantaining a zero derivative at the relative instants when the articulatory target of each fiduciary point is reached. Equation 1 describes the parametric model used for the interpolation between a given target (left target L) and the next one (right target R), where ($x(t)$, $y(t)$, $z(t)$) are the coordinates of a given point of interest (P_1 to P_4), L_x, L_y and L_z are the coordinates of the left target, R_x, R_y and R_z are the coordinates of the right target, and t is the independent time variable, normalized to the time interval between the left and right articulatory targets, $0 \leq t \leq 1$.

$$\begin{bmatrix} x(t) \\ y(t) \\ x(t) \end{bmatrix} = \begin{bmatrix} L_x & R_x \\ L_y & R_y \\ L_z & R_z \end{bmatrix} \begin{bmatrix} 2 & -3 & 1 \\ -2 & 3 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ 1 \end{bmatrix} \quad (1)$$

V. FACIAL ANIMATION

In order to control the dynamic behavior of a 3D synthetic face, the modelled fiduciary points movements were broken down into three components: a rigid body component associated with the mandible movement and two non-rigid components associated with the upper and lower lip movements.

A. Mandible movement

During speech, the mandible rotates and slides forward and backwards due to the temporomandibular joint [9]. The temporomandibular joint, or TMJ, is the joint connecting the mandible to the temporal bones at both sides of the head. Figure 5 presents the typical behavior of the temporomandibular joint within the midsagittal plane during speech. The TMJ behavior can be modelled by the composition of a rotation around the center of the TMJ in rest position, at initial time t_0, when the mouth is closed, followed by the translation of this center.

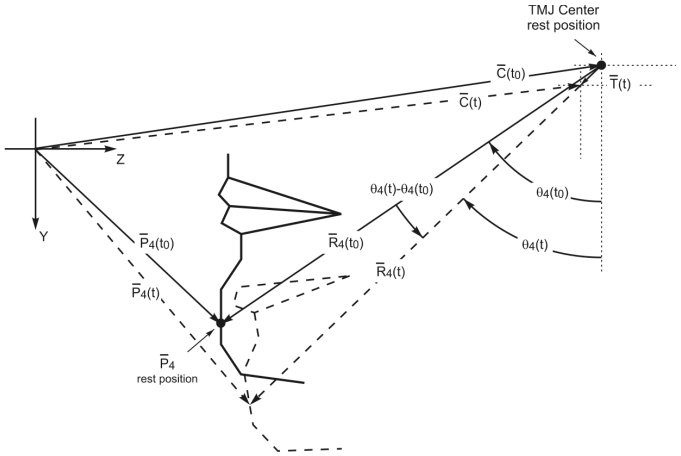

Fig. 5. Temporomandibular joint behavior.

From the modelled trajectory of fiduciary point P_4, we calculate the rotation angle $\theta_4(t)$ and translation components $t_y(t)$ and $t_z(t)$ of the TMJ in the mid-sagittal plane xy using Equations 2 and 3, respectively.

$$\theta_4(t) = \arctan\left(\frac{z_4(t) - c_z(t_0)}{y_4(t) - c_y(t_0)}\right) \quad (2)$$

$$\begin{cases} t_y(t) = y_4(t) - c_y(t_0) - r_4 \cos(\theta_4(t)) \\ t_z(t) = z_4(t) - c_z(t_0) - r_4 \sin(\theta_4(t)) \end{cases} \quad (3)$$

In the above equations y_4 and z_4 are the y and z coordinates of the modelled trajectory of point P_4, and $c_y(t(0))$ and $c_z(t(0))$ are the y and z components of vector $\overline{C}(t)$, the TMJ rotation center at rest position ($t = t_0$). The radius r_4, the module of vector $\overline{R_4}(t)$, is a constant and is defined by the mandible size. It is possible to estimate r_4 in the rest position by using Equation 4.

$$r_4 = \sqrt{[y_4(t_0) - c_y(t_0)]^2 + [z_4(t_0) - c_z(t_0)]^2} \quad (4)$$

In Figure 6 it is shown the y and z coordinates of P_4 measured during the production of the logatome ['aɐ], frame

by frame, and the estimated rotation component based on Equations 2 and 4 and the measured data. In Figure 7 it is shown the y and z coordinates of P_4 and the rotation component predicted by the proposed model and just using the target value of visemes <a, ɐ>, and the interpolation between the articulatory targets described in the previous section [8].

Figures 8, 9 show the measured (from the frame data and Equation 3) and estimated (articulatory targets and interpolation) translation of the TMJ. It can be noted that, adding to the extreme point of the curves in Figures 6, 7, in the opposite side of the rest position, the translation shown in Figures, 8, 9 respectively, it is possible to reach the extreme point of the P_4 trajectory (in the opposite side of the rest position).

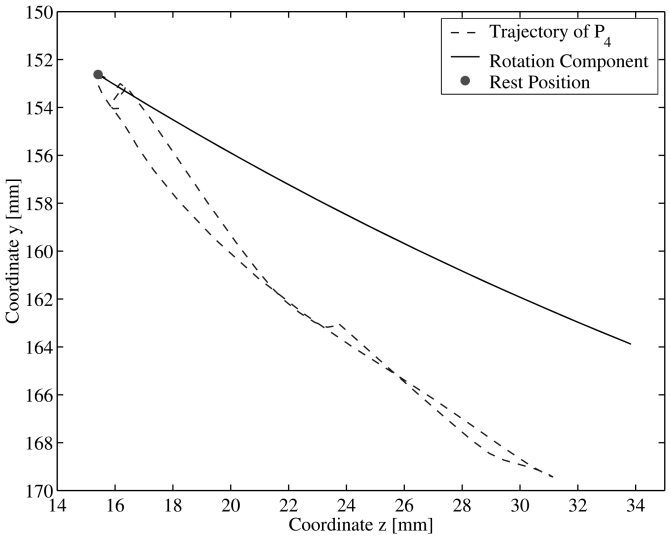

Fig. 6. TMJ rotation during the production of [’aɐ] estimated from measured data.

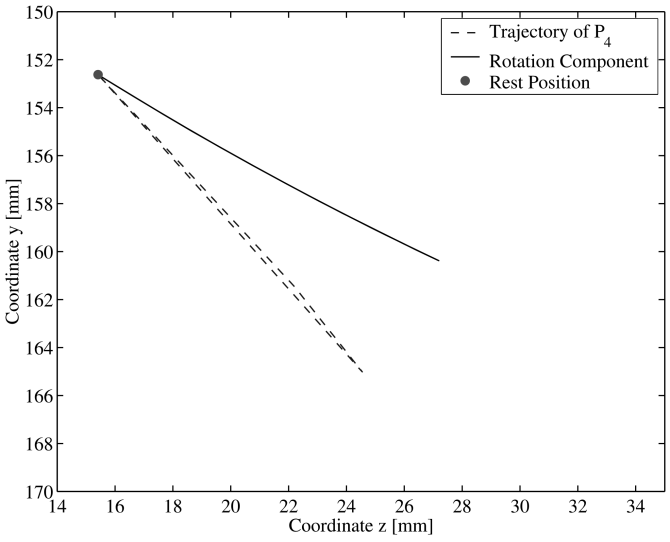

Fig. 7. TMJ Rotation during the production of [’aɐ] given by the model.

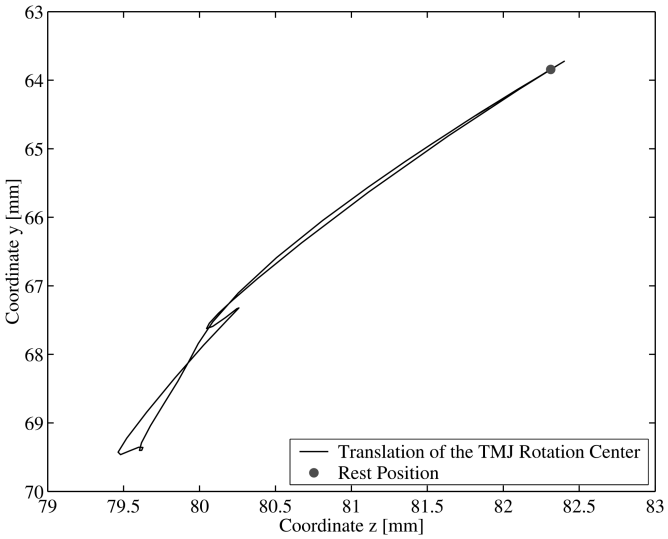

Fig. 8. TMJ translation during the production of [’aɐ] estimated from measured data.

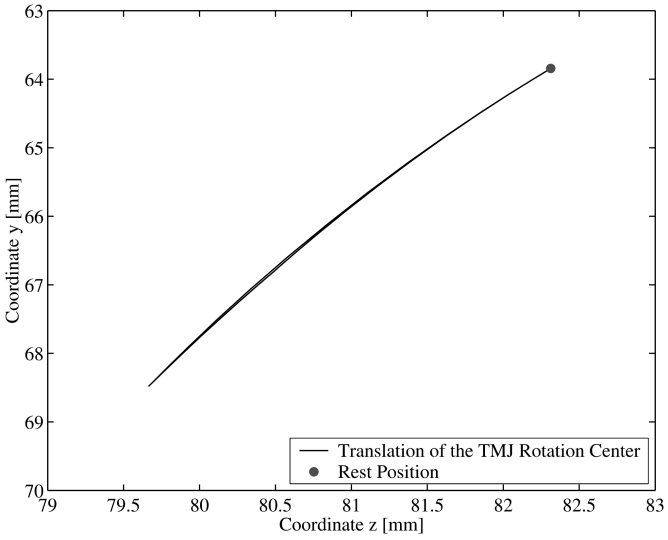

Fig. 9. TMJ translation during the production of [’aɐ] given by the model.

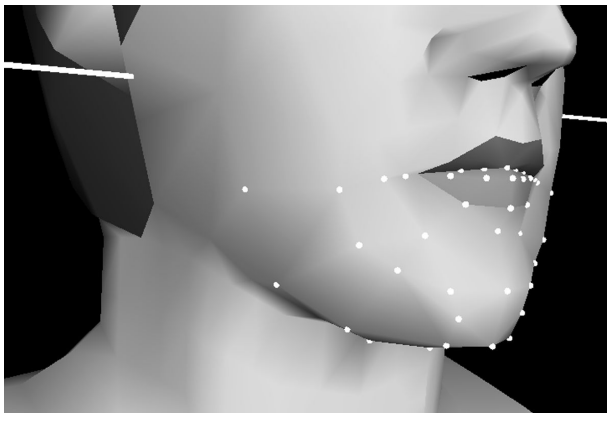

Fig. 10. Vertices associated with the mandible.

To reproduce the mandible movements in the synthetic face, the rigid body transformations of rotation and translation described by TMJ behavior were applied to the polygonal vertices of the geometric model. The transformed vertices were those within the region of the face alongside the mandibular

bone. More precisely, the vertices in the region below and including the lower lip and the lateral side of the face below an imaginary plane defined by TMJ rotation axis and the corners of the mouth. The lower bound of the mandibular region is defined by the neck. The white dots shown in Figure 10 show the vertices of the synthetic face submitted to these transformations. The rotation axis defined by the TMJ is represented in the figure by the white cylinder piercing the surface of the virtual face in front of the ears.

B. Lip movements

The movement of the fiduciary point P_3 located on the lower lip can be decomposed into two components. The first one is due to the mandible rotation and translation. The second one is the voluntary movement of the lower lip tissue necessary to produce specific speech gestures, such as lip protrusion. This first component is directly derived from the TMJ movement previously discussed. The second component is given by subtracting the movement of the TMJ from the trajectory of P_3. Differently, the displacement of P_1, located on the upper lip, is only driven by the voluntary movement of the upper lip tissue.

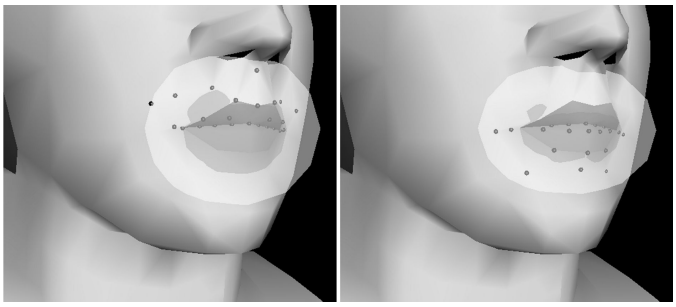

Fig. 11. Vertices associated with the upper lip (left) and lower lip (right) movement.

The behavior of the upper and lower lips was mapped onto the synthetic face on the basis of three main considerations. First, the points on the geometric model corresponding to the fiduciary points must exactly follow the fiduciary point trajectories expressed by the viseme model. Second, during speech production, the skin tissue around the mouth, including the lips, suffers deformations primarily attributed to the sphincteral behavior of the *Orbicularis Oris* muscle, which has an elliptical constitution [10]. Third, other muscles, which also influence the movement of the skin around the mouth, are distributed asymmetrically with respect to the horizontal plane.

Based on the above considerations, a geometric model was derived to express the visible characteristics of the skin around the mouth during speech production. We approximated the region of influence of the *Orbicularis Oris* muscle with a spheroid. To accommodate for the asymmetric characteristics of muscle insertions, the area around the mouth is divided into two regions, with the upper and lower regions influenced by the behavior of the upper and lower lips, respectively. Actually, each region of influence is defined by two spheroids: an internal one and an external one.

The external spheroid, which is merely a scaled instance of the internal one, defines the limits of influence of the lip behavior, while the internal one defines the points of maximum influence of that behavior. The influence of the lip behavior decays as one moves away from the surface of the internal spheroid, and ceases completely outside the external spheroid. The spheroids are expressed by Equation 5. The spheroids assume a Cartesian reference system centered in the mouth, with the same orientation as that of Figure 2.

$$\frac{x^2}{a^2}+\frac{y^2}{b_i^2}+\frac{z^2}{b_i^2}=\begin{cases}1 & \text{internal spheroid}\\ F_i^2 & \text{external spheroid}\end{cases} \quad i=1,3 \qquad (5)$$

The parameters a, b_i and F_i, with $i = 1, 3$ ($i = 1$ for the upper lip; $i = 3$ for the lower lip), are defined by the face geometry. The parameter a is equal to half of the distance between the corners of the mouth, that is, half the distance between P_2 and its counterpart on the other side of the mouth; the parameter b_i is equal to the distance between the major axis of the spheroid and the fiduciary point P_i; the scale factor F_1 is defined by the distance from the upper lip to the bottom center edge of the nose (to limit the upper region of influence at the columella-labial junction); F_3 is defined to limit the lower region of influence to the point halfway between the midpoint of the cleft and the tip of the chin. Figure 11 shows the region of influence defined by the spheroids and the vertices of the 3D face model included in the lower and upper region.

Fig. 12. The posture of the synthetic face during the production of /a/ (left side) and /u/ (right side).

The displacement $\Delta\overline{V}$ of a vertex inside a region of influence is given by Equation 5.

$$\Delta\overline{V} = R_i \ \left[\ D_i\ \Delta\overline{P}_2 + (1-D_i)\ \Delta\overline{P}_i\ \right] \quad i=1,3 \qquad (6)$$

where $\Delta\overline{P}_2$ is the displacement of fiduciary point P_2; $\Delta\overline{P}_i$ is the displacement of fiduciary point P_i, $i = 1, 3$ ($i = 1$ for the upper region of influence; and $i = 3$ for the lower one); $0 \leq D_i \leq 1$ is an interpolation factor given by Equation 6; and $0 \leq R_1 \leq 1$ is an attenuation factor given by Equation 7.

$$D_i = \left[cos\left(\frac{d_2}{d_2 + d_i}\pi\right) + 1 \right] /2 \quad i = 1, 3 \tag{7}$$

where d_2 is the distance between the vertex, whose displacement is being calculated, and the fiduciary point P_2 at rest position; and d_i is the distance between the vertex and fiduciary point P_i, $i = 1, 3$, at rest position. Depending on whether the vertex is inside or outside the internal spheroid, the fall-off factor R_i is calculated by:

$$\begin{cases} R_i = cos\left((1 - S_i)\ (\pi/2)\right) & \text{inside} \\ R_i = cos\left[\left((S_i - 1)\ /\ \left(F_i^2 - 1\right)\right)\ (\pi/2)\right] & \text{outside} \end{cases} \tag{8}$$

Factor S_i attenuates R_i as the location of the vertex moves away from the surface of the internal spheroid. S_i is obtained from the evaluation of the left side of Equation 4 at the vertex location.

To illustrate, Figure 12 presents snapshots showing the virtual face postures during the production of phonemes /a/ and /u/. Note the lip protrusion and rounding during /u/ and opening during /a/, in perfect agreement with real articulation.

VI. Speech Intelligibility Evaluation Test

In order to evaluate the naturalness of the talking head articulatory movements, a speech intelligibility test was conducted under low signal-to-noise ratio conditions in which the listeners always unconsciously make some kind of speechreading [11]. The audio and video of a real male speaker were recorded. The audio material consisted of a vehicle phrase "*Eu falo <logatome>*" ("*I say <logatome>*"), with 27 different logatomes of the type CV_1CV_2, C=[p,t,k,f,s,ʃ,l,λ,ɣ], V_1=[i,a,u] and V_2=[ɪ,ɐ,ʊ]. For example: [pipɪ], [papɐ], [pupʊ], [titɪ], [tatɐ], [tutʊ], etc. As the [ɾ] does not occur at the beginning of a word in Portuguese, it was left out of our evaluation set. The vehicle phrase was devised as a mean to get the subject's attention prior to the logatome utterance.

The audio was manually segmented in phones to synchronize the talking head movements. The movements incorporate an intrinsic synchronization error of 1/2 video frame, or 16.6 ms.

A white random noise with uniform distribution was added to the test phrases, assuring signal-to-noise ratios of 0 dB, -6 dB, -12 dB, -18 dB and -24 dB measured in the logatome segment. 33 subjects without any hearing impairment were invited to participate in the evaluation tests. In these tests the audio (through a headset) plus the video signal (through the microcomputer monitor) were presented to the subjects in a random order and under three different conditions: a) the audio together with the real speaker video; b) the audio together with the virtual talking head; c) the audio alone. For each presented vehicle phrase, the subjects were invited to choose in a list the more likely spoken logatome or to select the "none of the above - NDA" option.

Each intelligibility evaluation section lasted around one hour. 27 logatomes, under 5 different signal-to-noise ratios and in the three aforementioned conditions (audio only, audio plus real speaker video and audio plus virtual talking head), totalizing 405 files, composed the material presented to each subject in each evaluation section. Prior the test, the subjects were oriented to guess even if they were in doubt between different options, and only to choose the "none of the above" alternatives when they had absolutely no clue about the logatome presented or they had "heard" something else not provided as one of the logatome options.

Figure 13 presents a screenshot of the voting tool, showing a frame of the facial animation on the left side and the voting panel on the right side. Figure 14 shows the percentage of correct responses under the different signal-to-noise ratios.

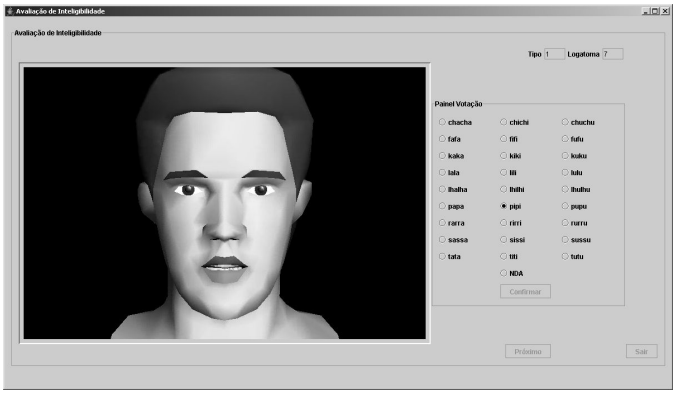

Fig. 13. Screenshot of the voting tool showing a frame of the facial animation.

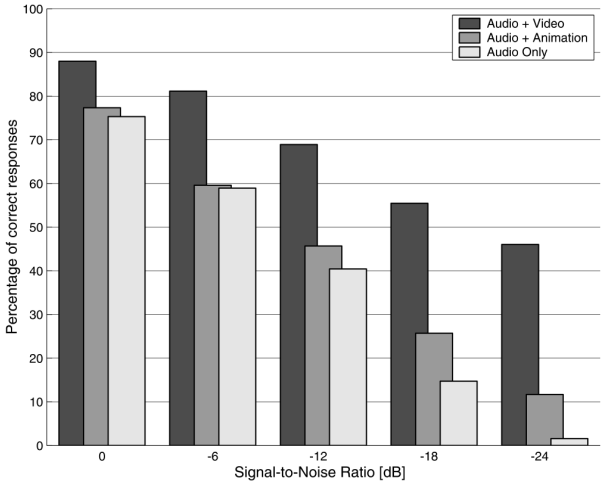

Fig. 14. Percentage of correct responses.

Although in high signal-to-noise ratios the real person video or the talking head video have no influence on the logatome intelligibility, under -18 dB and -24 dB the intelligibility gain reaches +41% and +44% respectively for the real person video presentation and +11% and +10% for the talking head presentation, both results with a statistic significance $p<10^{-3}$. In analyzing these data, it should be reminded that in the real face visualization not only the articulatory movements can help the increase in intelligibility, but also other movements produced during the articulation, as brow and eyes movement.

VII. Conclusion

The results presented in this paper show an improvement of the speech intelligibility due to the facial animation based on context-dependent visemes. The comparison with the real speaker video defines a reference to orient our future work. Although the context-dependent viseme approach helps speechreading, there is room for improvement as it is not as effective as the real video. Some aspects of our approach still claim further elaboration.

The first aspect is the refinement of the viseme model. The quality of the viseme representation in our approach is at some extent dependent on the number and distribution of the fiduciary points. Currently, the results are based on the analysis of only 4 points (P_1, P_2, P_3 and P_4) at normal video sampling rate (30 frames per second). It is expected that the analysis of more points in space and time will result in a better viseme representation.

The second aspect concerns the strategy used to manipulate the synthetic face. We used a pure geometric strategy to map the viseme representation given by the fiduciary point trajectories onto the virtual face, abstracting the existence of the skin and muscles and their biomechanical properties.

VIII. Future Work

An improvement of our approach could be potentially achieved by the use of a muscle-based approach to drive the facial animation. But it is an open question if and in what extent a more complex muscle-based approach could improve speech intelligibility, specially because there are many biomechanical parameters whose values are not easy to determine but yet have to be properly tuned.

To assess and compare results, a version based on biomechanics simulation is currently being implemented. This version is based on the muscle-model approach proposed by Lee, Terzopoulos and Waters [12]. We plan to test and compare in a near future the current version and the muscle-model version.

Referências

[1] Plínio A. Barbosa, Fábio Violaro, Eleonora C. Albano, F. Simões, P. Aquino, S. Madureira, and E. Françoso. Aiuruetê: A high-quality concatenative texto-to-speech system for Brazilian Portuguese with demisyllabic analysis-based units and a hierarchical model of rhythm production. In *Proceedings of the 6th European Conference on Speech Communication and Technology - Eurospeech'99*, pages 2059–2062, Budapest, Hungary, September 1999. ISCA.

[2] N. Ayache. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception*. Edinburgh University Press, 1991.

[3] José Mario De Martino. *Speech synchronized facial animation: phonetic context-dependent visemes for Brazilian portuguese*. PhD thesis, State University of Campinas - Brazil, July 2005. (in portuguese).

[4] Pamela L. Jackson. The Theoretical Minimal Unit for Visual Speech Perception: Visemes and Coarticulation. *The Volta Review*, 11(5):99–115, 1988.

[5] Janet Jeffers and Margaret Barley. *Speechreading Lipreading*. Charles C. Thomas Publisher, 1971.

[6] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey 07632, 1988.

[7] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes. *Computer Graphics Principles and Practice*. Addison-Wesley Publishing Company, 2th edition, 1990.

[8] José Mario De Martino, Léo Pini Magalhães, and Fábio Violaro. Facial animation based on context-dependent visemes. *Computers and Graphics*, 30(6):971–980, December 2006.

[9] Eric Vatikiotis-Bateson and David J. Ostry. An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics*, pages 101–117, 1995.

[10] Johannes Sobotta. *Atlas de anatomia humana*. Guanabara Koogan, Rio de Janeiro, RJ, 1990. Tradução de Helcio Werneck do original Atlas der Anatomie des Menschen.

[11] W. H. Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2):212–215, March 1954.

[12] Yuencheng Lee, Demetri Terzopoulos, and Keith Walters. Realistic modeling for facial animation. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, New York, NY, USA, 1995. ACM Press.

José Mario De Martino was born in Campinas, São Paulo, Brazil, on April 27, 1958. He holds a bachelor's, master's and doctor's degrees in Electrical Engineering from the School of Electrical and Computer Engineering of the State University of Campinas - Unicamp.

He is assistant professor at the School of Electrical and Computer Engineering of the State University of Campinas. His area of concentration is Computer Graphics. The research interests include: computer graphics, computer facial animation, embodied conversational agents, computer games.

He is member of the European Association for Computer Graphics - Eurographics and the Brazilian Computer Society - SBC.

Fábio Violaro was born in Campinas, São Paulo, Brazil, on December 8, 1950. He received the bachelor (1973), master (1975) and doctor (1980) degrees in Electrical Engineering from the School of Electrical and Computer Engineering of the State University of Campinas - Unicamp

He is full professor at the School of Electrical and Computer Engineering of the State University of Campinas. His area of concentration is Speech Processing. His research interest topics include: speech analysis, speech recognition and text-to-speech synthesis.

He is member of the Brazilian Telecommunications Society (SBrT) and the Luso-Brazilian Association of Speech Sciences (LBASS).