ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Computers & Graphics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTERS & GRAPHICS

www.elsevier.com/locate/cag

Technical Section

Facial animation based on context-dependent visemes

José Mario De Martino[a,*], Léo Pini Magalhães[a], Fábio Violaro[b]

[a]*Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering, State University of Campinas, 13083-970 – Av. Albert Einstein, 400, Campinas, SP, Brazil*

[b]*Department of Communications, School of Electrical and Computer Engineering, State University of Campinas, 13083-970 – Av. Albert Einstein, 400, Campinas, SP, Brazil*

Abstract

This paper presents a novel approach for the generation of realistic speech synchronized 3D facial animation that copes with anticipatory and perseveratory coarticulation. The methodology is based on the measurement of 3D trajectories of fiduciary points marked on the face of a real speaker during the speech production of CVCV non-sense words. The trajectories are measured from standard video sequences using stereo vision photogrammetric techniques. The first stationary point of each trajectory associated with a phonetic segment is selected as its articulatory target. By clustering according to geometric similarity all articulatory targets of a same segment in different phonetic contexts, a set of phonetic context-dependent visemes accounting for coarticulation is identified. These visemes are then used to drive a set of geometric transformation/deformation models that reproduce the rotation and translation of the temporomandibular joint on the 3D virtual face, as well as the behavior of the lips, such as protrusion, and opening width and height of the natural articulation. This approach is being used to generate 3D speech synchronized animation from both natural and synthetic speech generated by a text-to-speech synthesizer.

Keywords: Facial animation; Speech synchronized facial animation; Visual speech synthesis; Visemes; Coarticulation

1. Introduction

Computer facial animation refers to techniques for specifying and controlling the positioning and movement of a synthetic face into and between facial expressions. Apart from cartoon-like animation, which admits exaggerations and fairly simple lip motion, one of the goals of computer facial animation is realism. Especially in this realm, it is possible to distinguish two complementary aspects. The first is the simulation of the visible movements displayed on the face during speech production. The second encompasses the portrayal of all other possible facial behaviors, such as emotions and speech related communication signs [1], that are, however, not strictly bounded to the physical restrictions of speech production. The first aspect defines a subset of issues that is properly classified as speech synchronized facial animation, which is the main concern of this article.

Speech is produced by the movements of specific organs or articulators of the vocal tract. Each distinct speech sound is related to a characteristic positioning of the articulators, and some of their movements are wholly or partially visible on the speaker's face, especially in the region around the mouth, which comprises the upper end of the vocal tract. The simulation of such visible movements during speech production is the key to the generation of realistic speech synchronized 3D facial animation.

The term viseme [2], a shorthand for visual phoneme, constitutes a central concept of speech synchronized facial animation. Visemes can be understood as the recognizable visual motor patterns usually common to two or more phonemes [3,4]. These phonemes are the smallest speech sound units in a language that are capable of conveying a distinction in meaning, such as the 'p' in 'pie' in contrast with the 'b' in 'bye'. These sounds are not produced in isolation however, but involve coarticulation. Coarticulation refers to the altering of the set of articulatory movements made in the production of one phoneme

*Corresponding author.
E-mail address: martino@dca.fee.unicamp.br (J.M. De Martino).

0097-8493/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cag.2006.08.017

segment by those made in the production of an adjacent or nearby phoneme [5]. Hence a viseme depends not only on the actual phoneme being uttered, but also on its phonetic context. Efforts to distinguish visemes can be roughly classified into two general approaches, one focused on perceptual issues, and the other concerned with measurable geometric aspects visible on the face. In the first category are the early observations derived from clinical experience with hearing impaired individuals and the various studies based on visual recognition tests [5–12]. The second approach relies on measurements of movements displayed on the face during speech production to derive geometric models for visemes. The latter approach is based on the understanding that visemes involve the grouping of phonemes according to similarities in visible geometry. For speech synchronized facial animation, this seems to be an appropriate approach.

2. Previous work

A number of techniques have been used to capture and model key mouth shapes for speech animation. Parke [13] employed rotoscoping of recorded frames of a real speaker to define speech poses for demonstration utterances. Hill et al. [14] derived viseme parameters from static photographs illustrating a book about lipreading. Waters and Levergood [15] defined a chart of mouth shapes constructed from the observation of real lips. These early solutions, however, ignored the effects of coarticulation.

Pelachaud [1] specified a set of visemes based on the informal observations of Jeffers and Barley [3] and proposed a three step algorithm to handle coarticulation that ranks visemes according to deformability. In this approach, the ideal lip shape of a deformable phoneme is influenced by the shape of a less deformable viseme. Cohen and Massaro [16] implemented a coarticulation model based on Löfqvist's gestural theory of speech production [17]. In this model, each speech segment is associated with a target and a dominance function. A weighted sum of dominance functions specifies the trajectory of the articulators. Le Goff and Benoît [18,19] extended Cohen and Massaros's coarticulation model to get an n-continuous function. Revéret et al. [20] adopted the Öhman's coarticulation model [21]. Albrecht et al. [22] also extended the original Cohen and Massaro model to speed up the computation of coarticulation effects. Pelachaud [23] used radial basis functions to model the trajectory of articulatory parameters. Beskow [24] implemented and compared four coarticulation models: those of Cohen and Massaro, and Öhman's, as well as, two artificial neural network-based models. All the models performed equally well in the perceptual evaluation realized. All of these approaches use blending functions to model coarticulation.

Our approach, however, does not define ideal targets and associated blending functions, but rather seeks to establish a set of context-dependent visemes for the modelling of coarticulation. The appropriate concatenation of context-dependent visemes can reproduce the visible articulatory movement during speech production.

3. Articulatory target determination

The steps of our methodology for the identification of a set of visemes for facial animation are presented in Fig. 1. In this figure, the arcs represent flow of information, while the circles show processing steps. The final result, represented by the box, is a parametric model that describes the movement of fiduciary points on the face.

The linguistic analysis here was carried out for Brazilian Portuguese; it was based on a linguistic corpus of nonsense words of two types:

- $'CV_1\ CV_2$ where C = [p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ], V_1 = [i, a, u] and V_2 = [ɪ, ɐ, ʊ] expressed with the symbols of the International Phonetic Alphabet [25]. As the alveolar tap [ɾ] never occurs in Portuguese at the beginning of a word, we chose to analyze only the non-sense words of the type $'[ɣ]V_1[ɾ]V_2$. The vowels V_1 and V_2 are the extreme vowels of Brazilian Portuguese in tonic and post-tonic positions;
- Diphthongs $'V_1V_2$, where V_1 = [i, e, ɛ, a, ɔ, o, u] and V_2 = [ɪ, ɐ, ʊ].

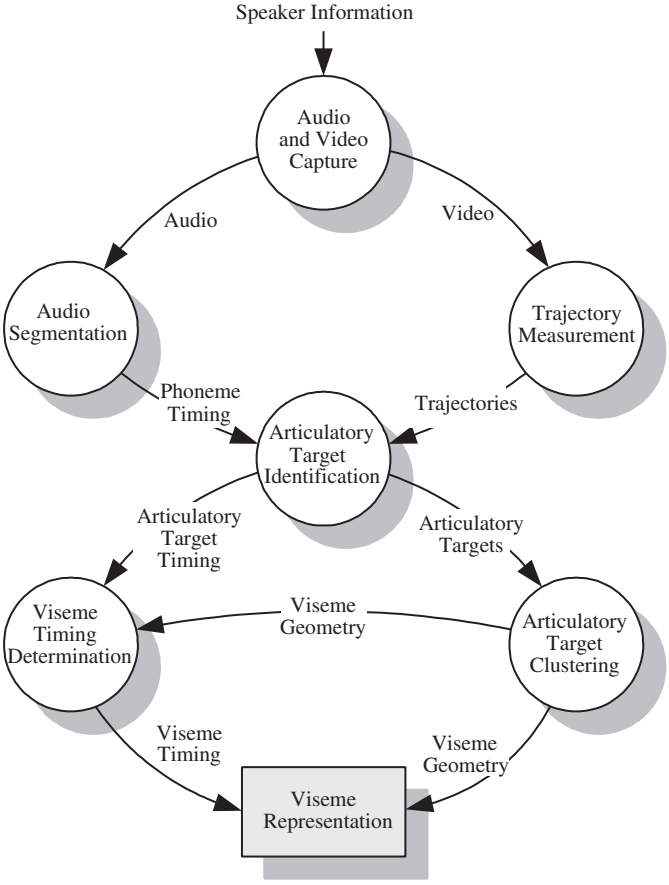

Fig. 1. Methodology for viseme determination.

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

The following realizations and phonetic contexts were analyzed:

- Consonants (18 realizations encompassing 12 different phonetic contexts): C[i] (3 realizations), C[a] (3 realizations), C[u] (3 realizations), [i]C[ɪ], [i]C[ɐ], [i]C[ʊ], [a]C[ɪ], [a]C[ɐ], [a]C[ʊ], [u]C[ɪ], [u]C[ɐ], and [u]C[ʊ].
- Tonic vowels [i, a, u] (30 realizations encompassing 12 different contexts): [p]V[p] (3 realizations), [t]V[t] (3 realizations), [k]V[k] (3 realizations), [f]V[f] (3 realizations), [s]V[s] (3 realizations), [ʃ]V[ʃ] (3 realizations), [l]V[l] (3 realizations), [ʎ]V[ʎ] (3 realizations), [ɣ]V[ɾ] (3 realizations), V[ɪ], V[ɐ], and V[ʊ].
- Post-tonic vowels [ɪ, ɐ, ʊ] (34 realizations encompassing 16 different phonetic contexts): [p]V (3 realizations), [t]V (3 realizations), [k]V (3 realization), [f]V (3 realizations), [s]V (3 realizations), [ʃ]V (3 realizations), [l]V (3 realizations), [ʎ]V (3 realizations), [ɣ]V (3 realizations), [i]V, [e]V, [ɛ]V, [a]V, [ɔ]V, [o]V, and [u]V;
- Non-extreme vowels [e, ɛ, ɔ, o] (3 realizations encompassing 3 different contexts): V[ɪ], V[ɐ], and V[ʊ].

To reduce the number of cases to be analyzed, the corpus was defined by exploiting the 'place of articulation' as a first order approximation of viseme grouping. Phonemes that share the same place of articulation constitute context-independent visemes traditionally accepted by speechreading researchers [26]. The initial grouping of the Brazilian Portuguese consonants into visemes is: [p, b,m], [f, v], [t, d, n], [s, z], [l], [ʃ, ʒ], [ʎ, ɲ], [k, g], [ɾ], and [ɣ]. The set of consonants in the present corpus is formed by a representative of each of these viseme groups. We tacitly assume that the context-dependent visemes found (Section 3.5) for a representative phoneme in the corpus are also valid for each phoneme in the same group.

Vowels also involve an initial grouping of the vocalic segments that differs only in relation to oral/nasal characteristics, within the same viseme category. Thus, for the vowels of Brazilian Portuguese, the following initial vocalic viseme grouping is considered: [i, ĩ], [e, ẽ], [a, ɐ̃], [o, õ], and [u, ũ]. Moreover, the vowels of the corpus were selected to emphasize the analysis of the extreme vowels [i, a, u] (tonic) and [ɪ, ɐ, ʊ] (post-tonic). Each of the other oral vowels of Brazilian Portuguese, [e, ɛ, ɔ, o], was analyzed in a few contexts just to identify the extreme vowel to which it is visually more similar.

A male speaker born and raised in the city of São Paulo, Brazil, was recorded producing the 102 stimuli composing the corpus. The utterance of the nonsense words of the corpus provided the input for the processing flow of Fig. 1. To capture the geometry information during speech production, fiduciary points were marked with white paint on the face of the speaker. Fig. 2 presents the location and labelling of the four fiduciary points, P_1, P_2 P_3 and P_4, for which 3D trajectories were measured and analyzed. The figure also shows the orientation of the Cartesian coordinate system consistently used throughout the text.

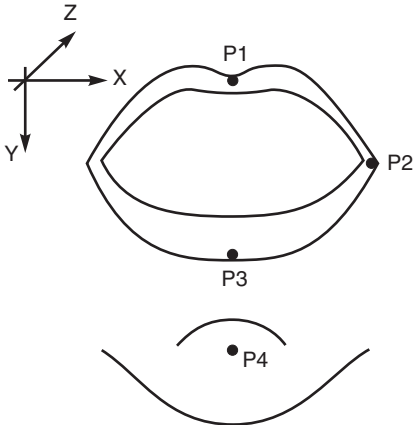

Fig. 2. Fiduciary points.

The nonsense words of the corpus were presented to and read by the speaker in a random order.

3.1. Audio and video capture

The speaker was recorded using two synchronized JVC KY27C video cameras positioned approximately 90° apart. The audio was captured by a Shure SM58 microphone placed directly in front of the speaker at a distance of approximately 1 m. The video and audio were recorded in S-VHS format using two VCRs, one a JVC BR-S622 DXU and the other a JVC BR-S822 DXU. The recording was realized in a sound-treated video production studio. The recorded analog material was converted to digital form using the video editing system iFinish V60, version 3.2.

3.2. Audio segmentation

After digitalization and word segmentation, the audio track was further segmented and labelled to identify the time span of the phonemes produced. This segmentation was carried out acoustically and visually by inspection of the signal spectrogram. As a result, the starting and end points of each phoneme were identified and marked on the audio and video tracks.

3.3. Trajectory measurement

The images of the video sequence were converted to a binary representation by level thresholding, after contrast enhancement and conversion from color to a gray scale representation. Moreover, operations of dilation, erosion and pruning were conducted on the binary image to extract artifacts other than the fiduciary points. The trajectories of these points were photogrammetrically measured using stereo vision techniques [27]. To provide the reference points needed for the measurements, a special helmet was constructed to provide a stable reference system despite involuntary head movement during speech production. Fig. 3 shows a sample frame of the captured video, with the

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

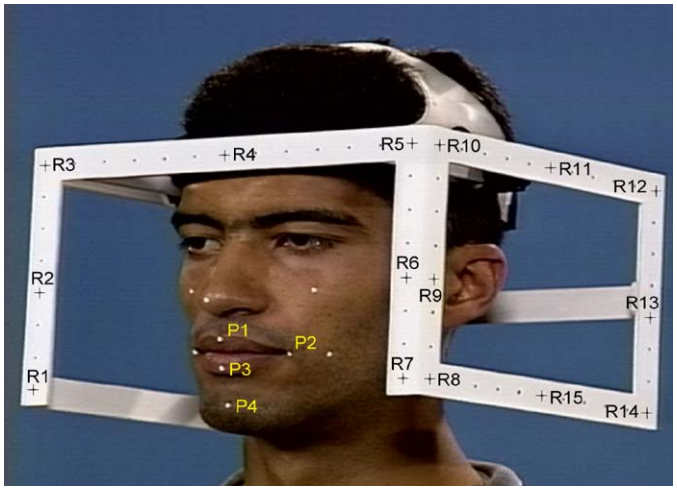

Fig. 3. Frame sample from the left side camera.

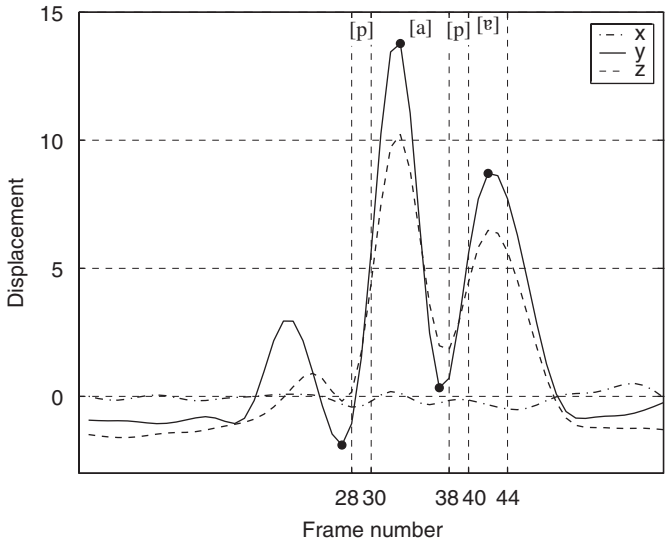

Fig. 4. Displacement of P_4 during production of ['papɐ].

15 reference points on each image R_i, $i = 1, \ldots, 15$, used for camera calibration identification.

3.4. Articulatory target identification

Once the trajectory of each fiduciary point for each nonsense word was calculated, the first stationary point (derivative equal zero) found during the production of each phoneme was identified and labelled as the phoneme articulatory target. This stationary point consisted of the component, x, y or z, revealing the greatest displacement. The time span for the acoustic realization of a phoneme was extended to include the two previous frames. This extension was necessary to account for plosive consonants, whose sound is produced during the release of the closure of the vocal tract that characterizes the phoneme. Fig. 4 presents the trajectory of point P_4 during the production of ['papɐ]. The articulatory targets of the phoneme are represented by solid black dots in the figure. The dashed vertical lines mark the borders of the acoustic production of the phoneme sequence. It can be seen that the articulatory target of the [p] always occurs prior to the acoustic realization, as expected since the bilabial [p] is a plosive consonant produced after the occlusion of the airflow.

3.5. Articulatory target clustering

The identification of visually distinguishable articulatory patterns from the trajectory of a phoneme in different phonetic contexts involved the clustering of the articulatory targets of a single phoneme using the Euclidian distance between them as the criterion of similarity. This clustering was performed using the K-Means algorithm of the SOM Toolbox for MatLab, developed by the Laboratory of Computer and Information Science of the Helsinki University of Technology. For each data set, we ran the algorithm 40 times, choosing as final result that clustering with the smallest Davies–Boulding index [28]. The clustering was realized in $\Re^{12}$ (three coordinates (x, y, z) of the four fiduciary points P_1, P_2, P_3 and P_4). The final results of the clustering procedure are presented in Table 1 (vocalic visemes) and Table 2 (consonantal visemes). The tables also present the phonetic contexts of each viseme. The centroid of the cluster was considered to be the articulatory target representing that cluster and was used to characterize the static geometric attributes of the viseme.

Table 1
Context-dependent vocalic visemes

Viseme	Contexts
⟨i_1⟩	All contexts but [tit] and [ʃiʃ]
⟨i_2⟩	[tit] [ʃiʃ]
⟨a⟩	All contexts
⟨u⟩	All contexts
⟨ɪ⟩	All contexts
⟨ɐ⟩	All contexts
⟨ʊ⟩	All contexts

The oral vowels [e, ɛ, ɔ, o] were classified in one of the seven vocalic visemes on the basis of the smallest distance from the centroid in its three productions from that of the articulatory targets of the vocalic visemes, resulting in the following associations: [e] → ⟨ɐ⟩, [ɛ] → ⟨a⟩, [ɔ] → ⟨ɐ⟩, and [o] → ⟨u⟩.

3.6. Viseme timing determination

Once the viseme clusters were established, an average value for the relative realization instants of all the articulatory targets composing the cluster was calculated. This value was adopted as the relative instant of realization

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

Table 2
Context-dependent consonantal visemes

Viseme	Contexts
⟨p1⟩	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]
⟨p2⟩	[pu] [upɪ] [upʊ]
⟨f1⟩	[fi] [fa] [ifi] [ifɐ] [ifʊ] [afi] [afɐ]
⟨f2⟩	[fu] [afʊ] [ufi] [ufɐ] [ufʊ]
⟨t1⟩	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]
⟨t2⟩	[ta] [atɐ]
⟨s1⟩	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]
⟨s2⟩	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]
⟨l1⟩	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]
⟨l2⟩	[la] [ilɐ] [alɪ] [alɐ]
⟨l3⟩	[lu]
⟨l4⟩	[ilʊ] [ulʊ]
⟨ʃ1⟩	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]
⟨ʃ2⟩	[ʃu] [uʃʊ]
⟨λ1⟩	[λi] [λa] [iλɪ] [iλɐ] [aλɪ] [aλɐ]
⟨λ2⟩	[λu] [uλɪ] [uλɐ]
⟨λ3⟩	[iλʊ] [aλʊ] [uλʊ]
⟨k1⟩	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]
⟨k2⟩	[ka] [akɐ]
⟨k3⟩	[ku] [ikʊ] [akʊ] [ukʊ]
⟨r1⟩	[ɣi] [ɣa] [irɪ] irɐ] [arɪ] [arɐ] [urɐ]
⟨r2⟩	[ɣu] [irʊ [arʊ] [urɪ] [urʊ]

of the representative articulatory target (centroid) of the cluster.

3.7. Viseme representation

Once the coordinates (x, y, z) and the timing of the articulatory target of the visemes were determined, the approximate trajectory between them was established using a smooth interpolation between the corresponding articulatory targets. The resulting interpolated curves must preserve geometric continuity and present derivatives equal to zero at the instant of realization of the articulatory targets. Although not strictly necessary, the interpolation curve adopted was the Hermite parametric cubic curve expressed by

$$\begin{bmatrix} x(t) \\ y(t) \\ x(t) \end{bmatrix} = \begin{bmatrix} I_x & F_x \\ I_y & F_y \\ I_z & F_z \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}, \quad (1)$$

where $x(t)$, $y(t)$ and $z(t)$ are the coordinates of the fiduciary point; I_x, I_y and I_z are the articulatory target coordinates of the first phoneme; F_x, F_y and F_z are the articulatory target coordinates of the second phoneme; and $0 \leqslant t \leqslant 1$ is the parametric variable.

To illustrate the interpolation process, Fig. 5 presents the measured and estimated displacements of the y coordinate of point P_4 during the production of ['aɐ]. It can be seen that the trajectory described by the model reproduces the profile of the measured trajectory fairly well. This was also true for all of the other trajectories.

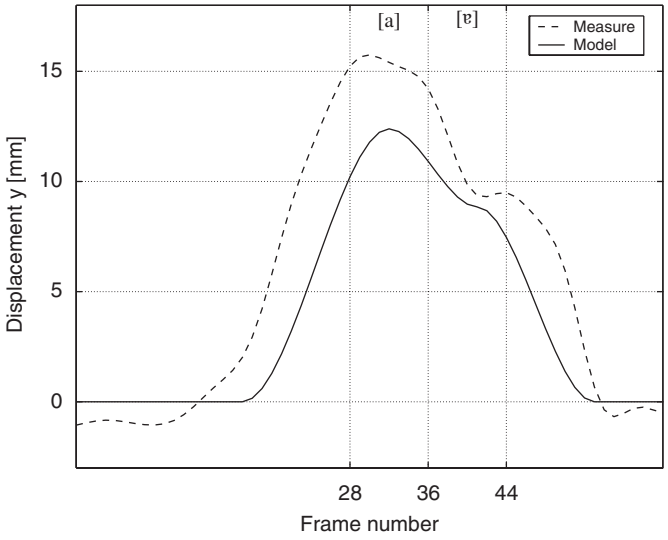

Fig. 5. Displacement of the y coordinate of point P_4 during the production of ['aɐ].

4. Major speech-related facial movements

For controlling the dynamic behavior of a 3D synthetic face, the speech related movements are broken down into three components: a rigid body component associated with the movement of the mandible, bounded by the behavior of the temporomandibular joint, and two non-rigid components associated with movements of the upper and lower lips. In the following paragraphs, we derive behavioral models for the temporomandibular joint and the lower lip based on the information provided by the trajectories of the fiduciary points. The upper lip behavior is directly given by the trajectory of P_1.

4.1. Temporomandibular joint behavior

The temporomandibular joint, or TMJ, is the joint connecting the mandible to the temporal bones at both sides of the head. The TMJ allows the occurrence of a diverse range of movements. The lower jaw can open and shut, move in and out and also side to side. During speech, the main movements of the TMJ are limited to the midsagittal plane, where the mandible rotates around an axis which suffers translation along that plane [29]. Fig. 6 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.

Considering the conventions in Fig. 6 and the measured coordinates $y_4(t)$ and $z_4(t)$ of the fiduciary point P_4 located on the chin, the TMJ angle of rotation can be

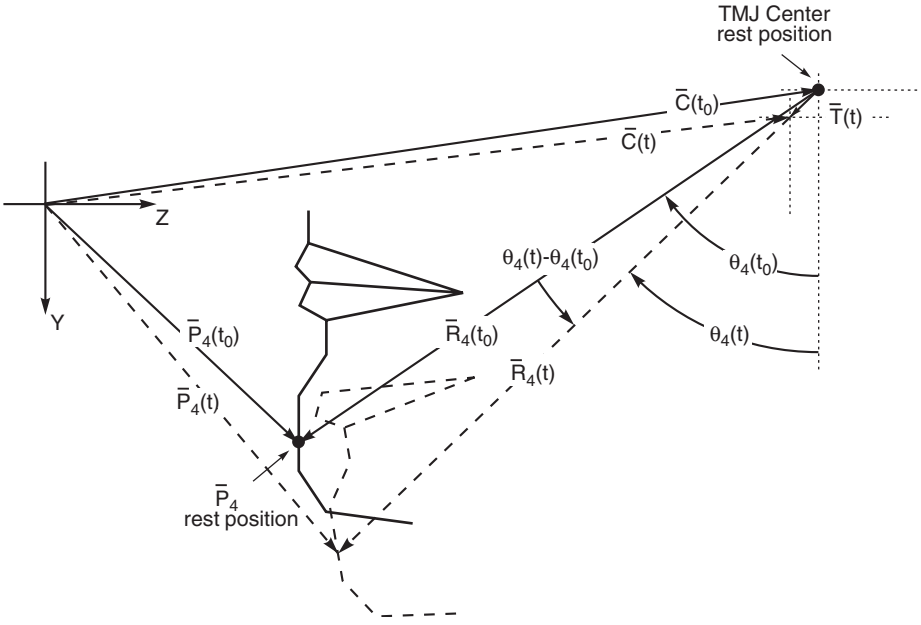

Fig. 6. Temporomandibular joint behavior.

expressed by

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

where $c_y(t_0)$ and $c_z(t_0)$ are the components on the midsagittal plane of $\overline{C}(t_0)$, with the TMJ center at rest position. The presumed location of the TMJ center of our speaker was also measured from the video.

The translations $t_y(t)$ and $t_z(t)$ of the TMJ center within the midsagittal plane can be expressed by

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

The radius r_4, the module of vector $\overline{R}_4(t)$, is a constant and is defined by the size of the mandible. It is possible to estimate r_4 in the position of rest by

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

Fig. 7 presents the modelled trajectory of point P_4 in the midsagittal plane during the production of ['aɐ]. The figure also shows the rotation associated with the movement of the TMJ. Fig. 8 shows the movement of the translation of the TMJ center during this phonetic production, with the head of the speaker in the same orientation as in Fig. 6.

4.2. Lower lip behavior

The lower lip behavior is related to the voluntary movement of the lower lip tissue (other than that caused by the mandible) necessary to produce specific speech gestures, such as lip protrusion. Lower lip behavior describes this kind of voluntary movement involved in the trajectory of the fiduciary point P_3. Non-rigid body deformation can be derived from this behavior to control

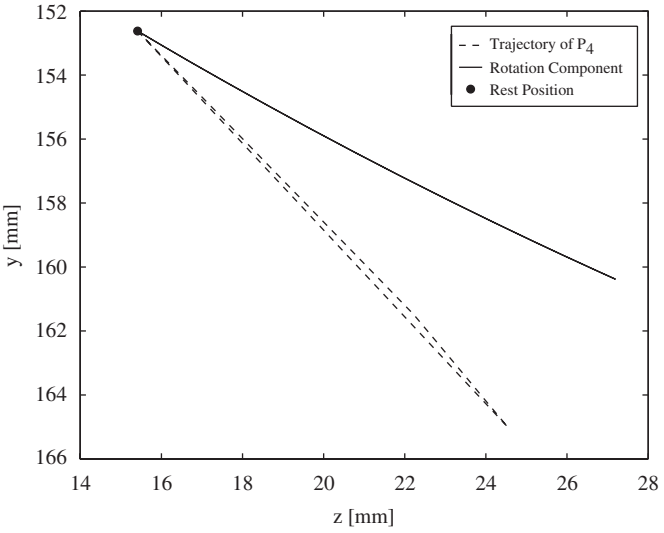

Fig. 7. Estimated trajectory of P_4 within the midsagittal plane during the production of ['aɐ] and associated rotation due to TMJ movement.

the animation of the lower region of the synthetic face around the mouth (see Section 4.4).

Assuming that $\overline{P}_3(t)$ is the trajectory of the fiduciary point P_3 and that $\overline{R}_3(t)$ is the movement of the lower lip due to that of the TMJ, the lower lip behavior can be expressed by $\overline{P}_3(t) - \overline{R}_3(t) - \overline{C}(t_0)$.

The components of the vector $\overline{R}_3(t)$ in the midsagittal plane are (directions Y and Z, respectively)

$$\begin{cases} r_{3y}(t) = r_3 \cos(\theta_3(t)) + t_y(t), \\ r_{3z}(t) = r_3 \sin(\theta_3(t)) + t_z(t). \end{cases} \tag{5}$$

Eq. (5) respects the same considerations of geometry underlying Fig. 6.

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

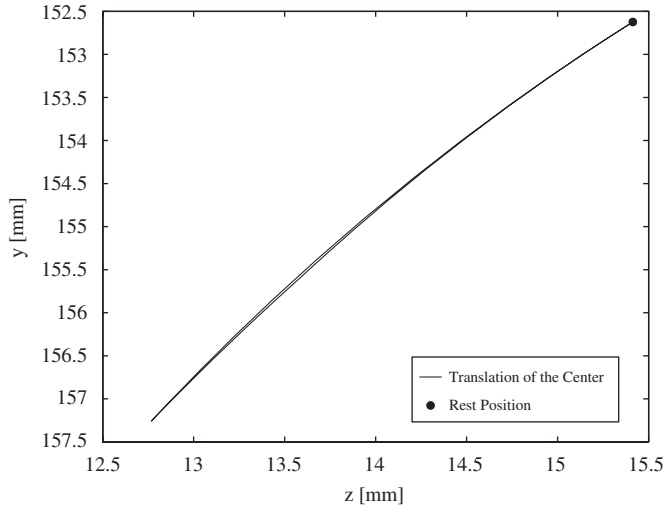

Fig. 8. Translation of the TMJ center within the midsagittal plane during the production of ['aɐ], as calculated from the estimated trajectory of P_4.

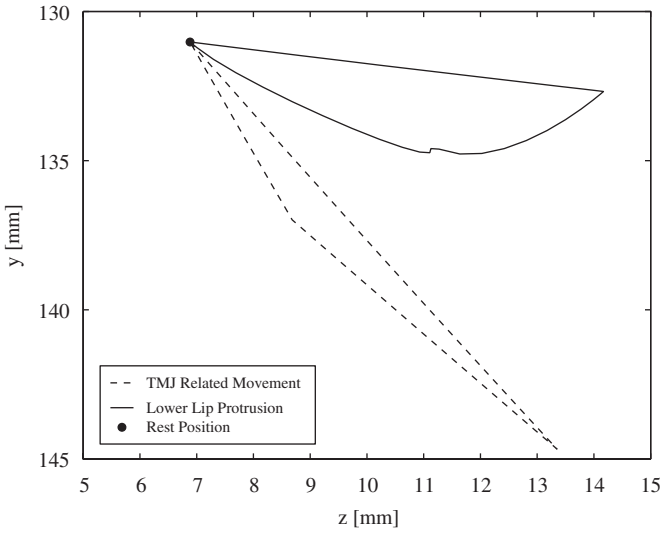

Fig. 9. Decoupled movements of the lower lip during the production of ['aɪ].

The distance r_3 between P_3 and the TMJ center can be calculated from the coordinates of P_3 and the location of the center of the TMJ at rest by

$$r_3 = \sqrt{[y_3(t_0) - c_y(t_0)]^2 + [z_3(t_0) - c_z(t_0)]^2}. \quad (6)$$

As the lower lip experiences the same variation of angle as P_4 due to the rotation of the TMJ, $\theta_3(t)$ can be expressed by

$$\theta_3(t) = \theta_4(t) - \theta_4(t_0) + \arctan\left(\frac{z_3(t_0) - c_z(t_0)}{y_3(t_0) - c_y(t_0)}\right). \quad (7)$$

The estimated trajectories of P_3 decoupled into movements related to that of the TMJ and that of lip protrusion during the production of ['aɪ], ['aɐ] and ['aʊ] are presented in Figs. 9, 10 and 11, respectively, with the speaker's head again maintaining the same orientation as in Fig. 6. These figures show that the analysis of the lower lip behavior correctly indicates the occurrence of lower lip protrusion during the production of [ʊ]. This protrusion occurs when the solid curve in the figures is to the left of the dashed one. When it is to the right, the lower lip is retracted, as in mouth spreading.

4.3. Upper lip behavior

As with lower lip behavior, upper lip behavior is associated with the voluntary movement of lip tissue. However, in contrast to what happens with the lower lip, the behavior of the upper lip associated with the fiduciary point P_1 is not entangled with TMJ movement. Thus, upper lip behavior can be determined directly from the trajectory of the fiduciary point P_1. Non-rigid body deformation is derived from this behavior to control the animation of region above the mouth of the synthetic face (see Section 4.4).

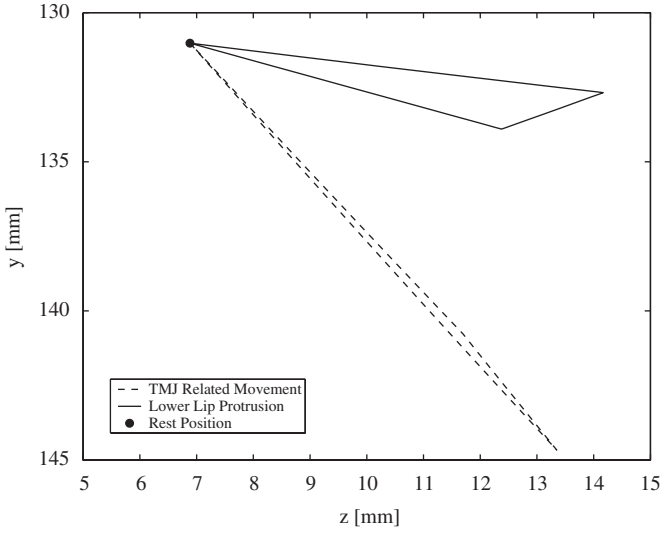

Fig. 10. Decoupled movements of the lower lip during the production of ['aɐ].

4.4. Virtual face manipulation

To reproduce the movements of the mandible in the synthetic face, the rigid body transformations of rotation and translation described by TMJ behavior are applied to the polygonal vertices of the geometric model. These transformed vertices are those within the region of the face alongside the mandibular bone, more precisely, 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 Fig. 12 are the vertices of the synthetic face submitted to these transformations. The axis of rotation defined by the center of the TMJ is presented in the figure as a white line in front of the ear.

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

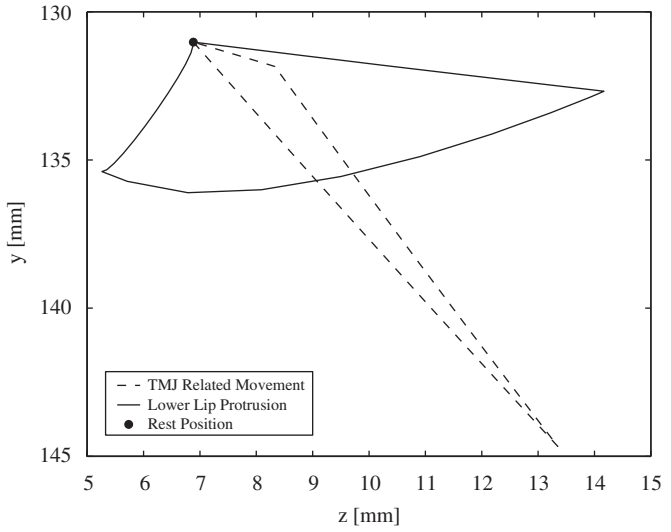

Fig. 11. Decoupled movements of the lower lip during the production of ['aʊ].

Fig. 12. Vertices that undergo rotation and translation affected by TMJ behavior during speech production.

The behavior of the upper and lower lips is 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 closely mimic the behaviors described by the measurements. 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. Third, other muscles which also influence the movement of the skin around the mouth are distributed asymmetrically with respect to the horizontal plane (with different groups of muscles inserted into the skin of the upper and lower regions of the mouth).

Based on these 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, which has an elliptical constitution [13], with a spheroid. To accommodate for the asymmetrical 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 behavior, while the internal one defines the points of maximum influence of that behavior. The influence of the behavior decays as one moves away from the surface of the internal spheroid, and ceases completely outside the external spheroid. The spheroids assume a Cartesian reference system centered in the mouth, with the same orientation as that of Fig. 2, and are generically expressed as

$$\frac{x^2}{a^2}+\frac{y^2}{b_i^2}+\frac{z^2}{b_i^2}=\begin{cases}1 & \text{internal spheroid}\\ F_i^2 & \text{external spheroid}\end{cases}\quad i=1,3. \tag{8}$$

The values of coefficients a and b_i are influenced by the geometry of the face. The following conditions define these coefficients:

(1) the corners of the mouth (fiduciary point P_2 and its symmetrically counterpart on the other side of the mouth) are the vertices in the major axis of the internal spheroid, i.e., the distance between P_2 and its counterpart equals $2a$; and
(2) the fiduciary point P_i is located on the surface of the internal spheroid, i.e., b_i equals the distance between the major axis of the spheroid and the fiduciary point P_i ($i=1$ for upper region, $i=3$ for lower region).

The scale factor F_1 (upper region) is defined by the distance from the upper lip to the bottom center edge of nose (to limit the region of influence at the columella-labial junction). For the lower region, the factor F_3 (lower region) limits the region of influence to a point halfway between the midpoint of the cleft and the tip of the chin. In other words, we let the skin of the chin be influenced while excluding the tip of the chin. Figs. 13 and 14 present the spheroids and polygon vertices of the geometric model of the lower and upper regions of influence, respectively.

Considering $\Delta\overline{P}_j$, $j=1,2,3$, the 3D displacement of the fiduciary point P_j from the rest position, the 3D displacement $\Delta\overline{V}$ of a vertex inside the region of influence is calculated by

$$\Delta\overline{V}=R_i[D_i\Delta\overline{P}_2+(1-D_i)\Delta\overline{P}_i],\quad i=1,3. \tag{9}$$

The interpolation factor $0\leqslant D_i\leqslant 1$ ($i=1$ for upper region, $i=3$ for lower region) is a function of the distance from the vertex to the fiduciary point P_i at rest position, and is defined as

$$D_i=\left[\cos\left(\frac{d_2}{d_2+d_i}\pi\right)+1\right]\Big/2,\quad i=1,3, \tag{10}$$

where d_j, with $j=1,2,3$, is the distance between the vertex and the fiduciary point P_j at rest.

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

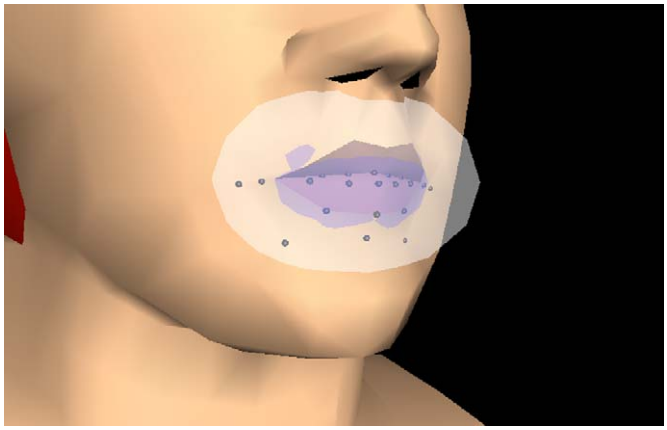

Fig. 13. Lower region of influence and associated vertices.

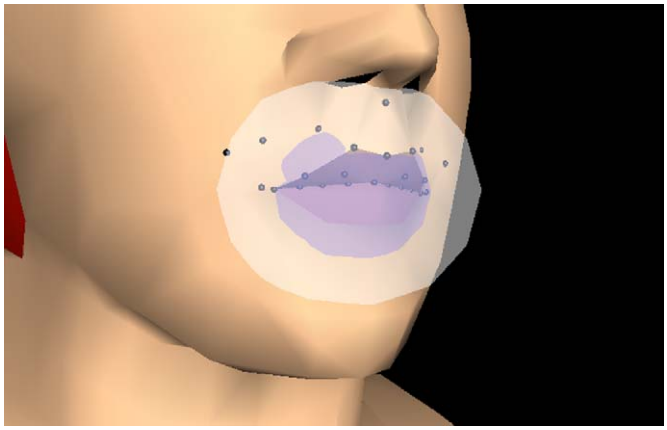

Fig. 14. Upper region of influence and associated vertices.

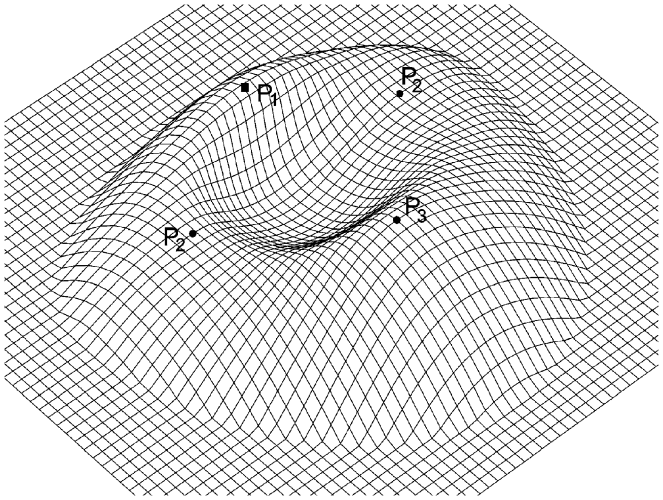

Fig. 15. Displacement of the surface around the mouth.

Depending on whether the vertex is inside or outside the internal spheroid, the fall-off factor $0 \leqslant R_i \leqslant 1$ is calculated by

$$\begin{cases} R_i = \cos((1 - S_i)(\pi/2)) & \text{inside}, \\ R_i = \cos[((S_i - 1)/(F_i^2 - 1))(\pi/2)] & \text{outside}. \end{cases} \quad (11)$$

The sphincteral factor S_i attenuates R_i as the location of the vertex moves away from the surface of the internal spheroid; it is obtained from the evaluation of the left side of Eq. (8) at the vertex location.

To illustrate the effects of upper and lower lip behavior, Fig. 15 presents a simulation of the displacement formulation applied to a regular planar grid, with the location of the fiduciary points indicated.

5. Conclusions

The visual representation of the phonemes, or visemes, is a crucial aspect of speech synchronized realistic facial animation. The recognizable characteristics of a viseme, however, can significantly change due to the effects of coarticulation. Traditionally, these effects have been modelled by fusion functions that blend phonetic context independent visemes. Context-independent visemes, however, are "pure" visemes that represent an idealization of expected speech postures. This idealization assumes that the speech segment, and therefore its viseme, is produced in an isolated form, without suffering interference from the articulatory movements of a nearby segment.

The characterization of a "pure" viseme and of its blending involves an intricate process and requires the gathering and analysis of extensive articulatory data. Nevertheless, despite practical cumbersomeness, the model based on blending functions proposed by Löfqvist [17] and adapted by Cohen and Massaro [16] to drive facial animation provides a powerful conceptual framework for the establishment of a general articulatory model of speech production. However, the implementation of this model using poorly defined parameters can easily produce deceptive results.

The approach we have suggested here is based on context-dependent visemes (see also [30]), namely, on the characterization of key speech postures which already accommodate the effects of coarticulation. The central idea of this approach is the characterization of a viseme not as a visual representation of an isolated speech segment, but as a composite including the influence of the segments that come before and after that segment. Considering the linguistic corpus analyzed so far, the viseme set presented in this article is able to cope with adjacent anticipatory and perseveratory coarticulation. The analysis was centered on triphone contexts of the type CVC and VCV. Although our methodology for the characterization of visemes was applied to Brazilian Portuguese, it is general enough to be used successfully with other languages. The main value of the approach described in this paper lies in its relatively simple and straightforward employment in the implementation of visemes and coarticulation effects. From a conceptual point of view, the proposed solution seeks to extract and model visual cues displayed on the speaker's face in order to drive facial animation without trying to lay down a general articulatory model of speech production. In this respect, we argue that we are pursuing "realism" in the

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017

sense that we are attempting to mimic the workings of reality.

An important issue that will be considered as a continuation of our work is the influence of the rate of speech on articulation/coarticulation. One promising possibility would be to extend the present articulatory analysis to establish what could be called speech rate-dependent and context-dependent visemes, with an initial attempt contemplating slow, normal and fast rates of speech.

Acknowledgment

The synthetic face used in this research is a modified version of the Miraface polygonal 3D face model developed at MIRALab, University of Geneva, and published by ISO as MPEG-4 reference software.

References

[1] Pelachaud C, Badler NI, Steedman M. Generating facial expressions for speech. Cognitive Science 1996;20(1):1–46.
[2] Fisher CG. Confusions among visually perceived consonants. Journal of Speech and Hearing Research 1968;11:796–804.
[3] Jeffers J, Barley M. Speechreading (Lipreading). Charles C. Thomas Publisher; 1971.
[4] Jackson PL. The theoretical minimal unit for visual speech perception: Visemes and coarticulation. The Volta Review 1988;11(5):99–115.
[5] Benguerel A-P, Pichora-Fuller MK. Coarticulation effects in lipreading. Journal of Speech and Hearing Research 1982;25:600–7.
[6] Erber NP. Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. Journal of Speech and Hearing Disorders 1972;15:413–22.
[7] Binnie CA, Montgomery AA, Jackson PL. Auditory and visual contributions to the perception of consonants. Journal of Speech and Hearing Research 1974;17:619–30.
[8] Walden BE, Prosek RA, Montgomery AA, Scheer CK, Jones CJ. Effects of training on the visual recognition of consonants. Journal of Speech and Hearing Research 1977;20(1):130–45.
[9] Walden BE, Erdman SA, Montgomery AA, Schwartz DM, Prosek RA. Some effects of training on speech recognition by hearing-impaired adults. Journal of Speech and Hearing Research 1981;24(1):201–16.
[10] Kricos PB, Lesner SA. Differences in visual intelligibility across talkers. The Volta Review 1982;84:219–25.
[11] Kricos PB, Lesner SA. Effect of talker differences on the speechreading of hearing-impaired teenagers. The Volta Review 1985;85:5–16.
[12] Owens E, Blazek B. Visemes observed by hearing-impaired and normal adult viewers. Journal of Speech and Hearing Research 1985;28:381–93.
[13] Parke FI, Waters K. Computer Facial Animation. A K Peters, Ltd; 1996.
[14] Hill DR, Pearce A, Wyvill B. Animating speech: an automated approach using speech synthesised by rules. The Visual Computer 1988;3(4):277–89.
[15] Waters K, Levergood TM. DECface: an automatic lip-synchronization algorithm for synthetic faces. Technical Report Series, DEC Cambridge Research Laboratory; September 1993.
[16] Cohen MM, Massaro DW. Modelling coarticulation in synthetic visual speech. In: Magnenat-Thalmann N, Thalmann D, editors. Models and techniques in computer animation. Berlin: Springer; 1993. p. 139–56.
[17] Löfqvist A. Speech as audible gesture. In: Hardcastle W, Marchal A, editors. Speech production and speech modeling. Dordrecht, the Netherlands: Kluwer Academic Publishers; 1990. p. 289–322.
[18] Le Goff B, Benoît C. A text-to-audiovisual-speech synthesizer for French. In: Proceedings of the fourth international conference on spoken language processing, vol. 4. Philadelphia, USA, 1996. p. 2163–66.
[19] Benoît C, Le Goff B. Audio-visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP. Journal of Speech and Hearing Research 1998;26:117–29.
[20] Revéret L, Bailly G, Badin P. Mother: a new generation of talking heads providing a flexible control for video-realistic speech animation. In: Proceedings of the sixth international conference on spoken language processing—ICSLP'00, ISCA, Beijing, China, 2000. p. 755–8.
[21] Öhman SEG. Numerical model of coarticulation. The Journal of the Acoustical Society of America 1967;41(2):310–20.
[22] Albrecht I, Haber J, Seidel H-P. Speech synchronization for physics-based facial animation. In: Proceedings of the tenth international conference in central Europe on computer graphics, visualization and computer vision—WSCG'2002, 2002. p. 9–16.
[23] Pelachaud C. Visual text-to-speech. In: Pandzic IS, Forchheimer R, editors. MPEG-4 facial animation. New York: Wiley; 2002. p. 125–40.
[24] Beskow J. Trainable articulatory control models for visual speech synthesis. International Journal of Speech Technology 2004;7(4): 335–49.
[25] TIP Association. Handbook of the international phonetic association—a guide to the use of the international phonetic alphabet. Cambridge: Cambridge University Press; 1999.
[26] Nitchie E. New lessons in lipreading. JB Lippincott, 1950.
[27] Ayache N. Artificial vision for mobile robots: stereo vision and multisensory perception. Edinburgh University Press; 1991.
[28] Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall; 1988.
[29] Vatikiotis-Bateson E, Ostry DJ. An analysis of the dimensionality of jaw motion in speech. Journal of Phonetics 1995; 101–17.
[30] De Martino JM. Speech synchronized facial animation: phonetic context dependent visemes for Brazilian Portuguese. PhD thesis, State University of Campinas, Brazil; July 2005 [in Portuguese].

Please cite this article as: José Mario De Martino et al., Facial animation based on context-dependent visemes, Computers & Graphics (2006), doi:10.1016/j.cag.2006.08.017