Benchmarking Speech Synchronized Facial Animation Based on Context-Dependent Visemes

José Mario De Martino
Dept. of Computer Eng. and Industrial Automation
School of Electrical and Computer Engineering
State University of Campinas, CP 6101
13083-970, Campinas, SP, Brazil
martino@dca.fee.unicamp.br

Fábio Violaro
Department of Communications
School of Electrical and Computer Engineering
State University of Campinas, CP 6101
13083-970, Campinas, SP, Brazil
fabio@decom.fee.unicamp.br

ABSTRACT

In this paper we evaluate the effectiveness in conveying speech information of a speech synchronized facial animation system based on context-dependent visemes. The evaluation procedure is based on an oral speech intelligibility test conducted with, and without, supplementary visual information provided by a real and a virtual speaker. Three situations (audio-only, audio+video and audio+animation) are compared and analysed under five different conditions of noise contamination of the audio signal. The results show that the virtual face driven by context-dependent visemes effectively contributes to speech intelligibility at high noise degradation levels (Signal to Noise Ratio (SNR) ≤ -18dB).

Keywords

Facial animation, Facial animation evaluation, Context-dependent visemes, Speech intelligibility test.

1. INTRODUCTION

Facial animation has been a research topic for more than three decades. Since Park's pioneer work [Par72], different approaches have been proposed to improve the various aspects involved in facial animation. Among them, the proper reproduction of the visible speech articulatory movements in synchronization with the audio is a crucial issue for many applications. Examples of such applications include virtual characters for films, embodied conversational agents and virtual helpers for language learning.

Viseme, the shorthand of visual phoneme, is the term used to denote the recognizable visual motor patterns common to two or more speech segments [Jac88]. A viseme is a visually contrastive unit usually associated with more than on speech segment. Speech segments represented by the same viseme are produced with similar visible speech articulatory movements and therefore are not visually distinguishable. Such segments are called homophenes. However a viseme representation can be affected by coarticulation. Coarticulation refers to the altering of the set of articulatory movements made in the production of one speech segment by those made in the production of an adjacent or nearby one [BP82]. Coarticulation can be classified as antecipatory or perseverative. Antecipatory coarticulation takes place when the articulatory movement of a given segment is influenced by the articulation of a following one. Perseverative coarticulation refers to the change in the articulation pattern of a given segment influenced by the production of a preceding one. Therefore visemes are dependent on the phonetic context of the acoustic production of the associated speech segment.

One strategy for speech synchronized facial animation aimed at reproducing the visible movements and associated coarticulation effects is based on context-dependent visemes [DMV06]. The central idea of the context-dependent viseme approach is the characterization of visemes not as visual representations of isolated speech segments, but as a composite including the influence of coarticulation. The main advantage of context-dependent visemes lies in its relatively simple way to handle antecipatory and perseverative coarticulation effects.

In this paper we assess the effective contribution to speechreading of context-dependent visemes

 ISBN 978-80-86943-98-5

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 [SP54]. The same test, with minor variations, has been widely applied to evaluate speech synchronized facial animation [BL98] [PO99] [Bes04] [MOC+05]. Essentially the intelligibility test involves the presentation of audio signals with, and without, supplementary visual information, conducted under different conditions of acoustic degradation.

2. RELATED WORK

Other strategies have been proposed to model coarticulation and visible articulatory movements in the context of facial animation. Pelachaud and colleagues [PBS96] specified a set of visemes based on the informal observations of Jeffers and Barley [JB71] 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 phoneme. Cohen and Massaro [CM93] implemented a coarticulation model based on Löfqvist's gestural theory of speech production [Lof90]. 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 [LB96] [BL98] extended Cohen and Massaros's coarticulation model to get an n-continuous function. Revéret and colleagues [RBB00] adopted the Öhman's coarticulation model [Ohm67]. Albrecht et al. [AJS02] also extended the original Cohen and Massaro model to speed up the computation of coarticulation effects. Pelachaud [Pel02] used radial basis functions to model the trajectory of articulatory parameters. Beskow [Bes04] 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.

In contrast, our approach does not define ideal targets and associated blending functions, but rather seeks to establish a set of context-dependent visemes that already incorporates the effects of coarticulation. The appropriate concatenation of context-dependent visemes reproduces the visible articulatory movements during speech production.

3. OUR APPROACH

Our approach characterizes visemes as transitions between context-dependent articulatory targets. The articulatory target of a speech segment is the distinctive conformation or posture of the vocal tract necessary for the acoustic production of the segment. Instead of treating speech segments in isolation, we consider the phonetic context of its production as well.

Articulatory targets

In order to identify context-dependent articulatory targets we measured and analysed the visible 3D trajectories of fiduciary points marked on the face of a real speaker during the speech production. The trajectories were measured from standard video sequences using stereovision photogrammetric techniques. The linguistic analysis was carried out for Brazilian Portuguese and based on a corpus of non-sense paroxytone words of the type $'CV_1CV_2$ and diphthongs of the type $'V_3V_2$, where C_1 = [p, t, k, f, s, ʃ, l, ʎ, ɣ(ɾ)], V_1 = [i, a, u], V_2 = [ɪ, ɐ, ʊ] and V_3 = [i, e, ɛ, a, ɔ, o, u], expressed with symbols of the International Phonetic Alphabet [IPA99]. As the alveolar tap [ɾ] never occurs in Portuguese at the beginning of a word, we chose, regarding [ɣ, ɾ], to analyze only the non-sense words of the type $[ɣ]V_1[ɾ]V_2$. More details about the reasoning behind the selection of this set of segments can be found in [DMV06].

A male speaker born and raised in the city of São Paulo, Brazil, was recorded producing the 102 stimuli composing the corpus. Figure 1 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 analysed.

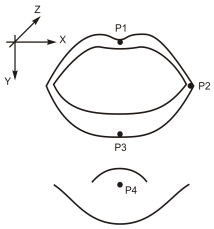

Figure 1: Fiduciary points.

The stationary points (derivative equal to zero) of the fiduciary point trajectories during segment production were considered as the segment articulatory target. The articulatory targets of a same segment in its various phonetic contexts in the corpus were then cluster analysed using the K-means clustering algorithm and the euclidian distance as a similarity criterion [Job92]. As result, a set of phonetic context-dependent articulatory targets was identified. Once the articulatory target clusters were established, an average value for the relative instants of realization of all the articulatory targets composing the cluster was adopted as the relative instant of

realization of the representative articulatory target (centroid) of the cluster. The coordinates (x, y, z) and the timing of the articulatory target are used as the driving parameters of our facial animation system. Given the timed sequence of speech segments that compose an utterance, the fiduciary point trajectories were approximated by the smooth interpolation between the corresponding articulatory targets using a Hermite parametric cubic curve. These trajectories, which define visual motor patterns, and therefore 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.

Speech-related facial movements

In order to control the dynamic behavior of a 3D synthetic face, the modeled 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.

3.1.1 Mandible movement

During speech, the mandible rotates and slides forward and backwards due to the temporomandibular joint. The temporomandibular joint, or TMJ, is the joint connecting the mandible to the temporal bones at both sides of the head. Figure 2 presents the typical behavior of the temporomandibular joint within the midsagittal plane during speech. The TMJ behavior can be modeled 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.

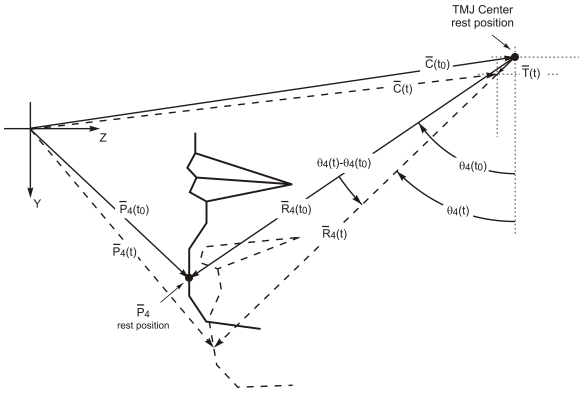

Figure 2: Temporomandibular joint behavior.

From the modeled trajectory of fiduciary point P_4, we calculate the rotation angle $q_4(t)$ and translation components $t_y(t)$ and $t_z(t)$ of the TMJ in the mid-sagittal plane xy using Equations 1 and 2, respectively.

$$q_4(t) = \frac{z_4(t) - c_z(t_0)}{y_4(t) - c_y(t_0)} \quad (1)$$

$$\begin{cases} t_y(t) = y_4(t) - c_y(t_0) - r_4 \cos(q4(t)) \\ t_z(t) = z_4(t) - c_z(t_0) - r_4 \sin(q4(t)) \end{cases} \quad (2)$$

In the above equations y_4 and z_4 are the coordinates y and z of the modeled 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 center of rotation, 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 size of the mandible. It is possible to estimate r_4 in the position of rest by Equation 3.

$$r_4 = \sqrt{[y_4(t_0) - c_y(t_0)]^2 + [z_4(t_0) - c_z(t_0)]^2} \quad (3)$$

Figure 3: Vertices associated with the mandible.

To reproduce the movements of the mandible 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 3 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.

3.1.2 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

 ISBN 978-80-86943-98-5

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.

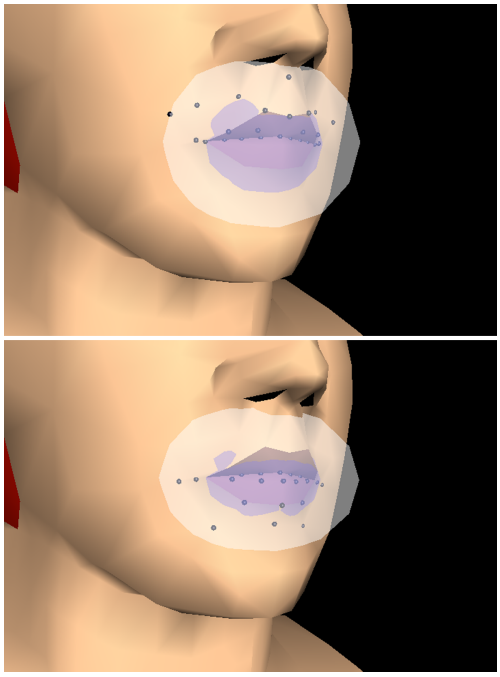

Figure 4: Vertices associated with the upper lip (above) and lower lip (below) movement.

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. 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 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 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 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 4a (internal) and Equation 4b (external). The spheroids assume a Cartesian reference system centered in the mouth, with the same orientation as that of Figure 1.

$$\frac{x^2}{a^2}+\frac{y^2}{b_i^{\,2}}+\frac{z^2}{b_i^{\,2}}=\begin{cases}1 & (4a)\\ F_i^{\,2} & (4b)\end{cases} \qquad i=1,3$$

The parameters a, b_i and F_i, with i = 1, 3 (i = 1 upper lip; i = 3 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 4 shows the region of influence defined by the spheroids and the vertices of the 3D face model included in the lower and upper region.

Figure 5: The posture of the synthetic face during the production of /a/ (left side) and /u/ (right side).

Equation 5 gives the displacement $\Delta\overline{V}$ of a vertex inside a region of influence.

$$\Delta\overline{V} = R_i\left[D_i\ \Delta\overline{P}_2 + (1-D_i)\,\Delta\overline{P}_i\right] \qquad (5)$$

where $\Delta\overline{P}_2$ is the displacement of the fiduciary point P_2; $\Delta\overline{P}_i$ is the displacement of fiduciary point P_i, i =1,

 ISBN 978-80-86943-98-5

3 (i = 1 for the upper region of influence; and i = 3 for the lower one); $0 \le D_i \le 1$ is an interpolation factor given by Equation 6; and $0 \le R_i \le 1$ is an attenuation factor given by Equation 7.

$$D_i = \left[\cos\left(\frac{p \; d_2}{d_i + d_2} \right) + 1 \right] \Big/ 2 \qquad (6)$$

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:

$$R_i = \begin{cases} \cos\left((1 - S_i) \dfrac{p}{2} \right) & \textit{inside} \\ \cos\left(\dfrac{(S_i - 1)}{(F_i^2 - 1)} \dfrac{p}{2} \right) & \textit{outside} \end{cases} \qquad (7)$$

The 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 5 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.

4. EVALUATION

Method

The intelligibility test carried out to evaluate our approach consisted on the presentation of a set of 27 non-sense words or logatomes conveyed in a vehicle phrase uttered by a Brazilian male speaker. The adopted vehicle phrase had the following structure: "Eu falo <*logatome*>" ("I say <*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 five different Signal-to-Noise-Ratio (SNR) conditions: 0dB, -6dB, -12dB, -18dB and -24dB.

The speaker was recorded in digital Mini-DV format using a JVC DV-GY500 camcorder and a Shure SM58 microphone. The camera was positioned to capture a front view of the speaker. The front view was chosen because it is the normal face-to-face conversation view. The recorded raw material was transferred to a video editing system iFinish V60, version 3.2, for further manipulation. After phrase segmentation, the audio was detached from the video. The audio-only files were used as the base material for the generation of new versions with different SNRs, resulting a total of 135 (27 x 5) audio files. The degradation of the acoustic signal was realized through the addition of noise.

The original, not degraded, audio signal was manually segmented at the speech segment level (phones). The resulting timed phonetic transcription was used to drive our facial animation system, which was configured to generate and store each frame of the animation in TGA format with the standard NTSC resolution of 720 x 468. Copies of the degraded audio files were then properly (re)synchronized with the corresponding real speaker video and with the facial animation frames. The material was packed in QuickTime ".mov" files with no audio compression and with video compression using the Soreson codec bundled in the iFinish system. As final result, 3 sets of 135 files were prepared for presentation: an audio-only, an audio+video, and an audio+animation set.

The resulting 405 files were organized for presentation in 15 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 special software tool was implemented to present, collect and record the vote of the test subjects.

Corpus

The logatomes used in the intelligibility test were paroxytone logatomes of the type ˈCVCV, with C = /p, t, k, f, s, ʃ, l, ʎ, ɣ/, V = /i, a, u/. The logatomes formed by the concatenation of two CV contexts were chosen to stimulate the production of coarticulation effects during its utterance.

The consonant set was selected considering the following homophene grouping of the Brazilian consonants: /p, b, m/, /f, v/, /t, d, n/, /s, z/, /l/, /ʃ, ʒ/, /ʎ, ɲ/, /k, g/, /ɣ/, and /ɾ/. Consonants of a same homophene group are not visually distinguishable. With one exception, a representative consonant of each homophene group was considered in the intelligibility test. As the /ɾ/ does not occur at the beginning of a word in Portuguese, we decided to let it out of our evaluation set. The vowel set used is formed by the three extreme vowels of Brazilian Portuguese.

Noise Contamination

The audio material was originally recorded at 48 kHz sampling rate and 16 bits resolution. The audio files

 ISBN 978-80-86943-98-5

were then downsampled to 16 kHz sampling rate, and manually segmented using spectrogram analysis. The timing and phonetic identification of the segment was stored in a text file and used for controlling and speech synchronizing the facial animation.

The downsampled audio files were converted to float format, and normalized to the interval [-1.0, 1.0] after a division by 32768. Five versions of each vehicle phrase were prepared with SNRs of 0, -6, -12, -18 and -24 dB. A uniform distribution random noise was employed. The noise was added over the whole phrases, but the SNR level was calculated only considering the logatome. After the noise addition, the audio amplitude was renormalized to the [-1.0, 1.0] interval and again quantized to 16 bits resolution and packed into RIFF WAVE files.

Subjects

33 subjects were invited to participate in the intelligibility test. This group of individuals consisted of normal-hearing undergraduate and graduate students, and administrative personnel of our School. They had no prior knowledge of the test purpose or any involvement with the facial animation research. Their age ranged from 19 to 55 years.

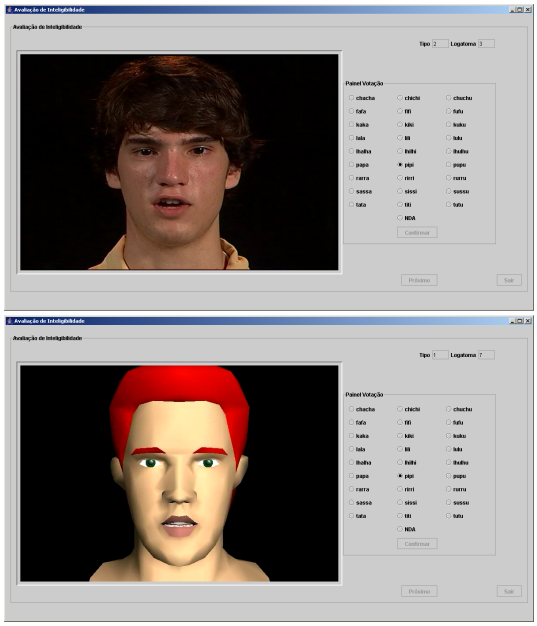

Figure 6: Screenshots of the voting tool showing above a frame of the real speaker video and below a frame of the facial animation.

Evaluation Environment

The intelligibility test was conducted on an isolated small dark room (3.5mx3.5m) with low ambient noise. A java application (j2sdk1.4.2_08) running on a Compaq Professional Workstation SP750 with an Intel Pentium III Xeon 733MHz processor was used to present the test material and collect the subject's responses. The workstation was equipped with a 21 inch Compaq X1100 color monitor. Figure 6 shows screenshots of the evaluation software. Each subject heard the audio material using a good quality headset plugged to the microcomputer audio card.

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. Figure 7 shows the voting panel in detail. After the vote confirmation realized through the Confirm button (*Confirmar* in Portuguese), the subject had to click the Next button (*Próximo* in Portuguese) in order to start the next presentation.

Figure 7: Detail of the voting panel.

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 subjects were informed that the logatomes were of the type CVCV and that different logatomes from those shown as options could be presented.

Results

The results of the intelligibility test are summarized in Figure 8. The figure shows the percentage of correct responses to the logatome stimulus. It can be seen that the visual information provided by both real and virtual speaker increases the speech intelligibility. This gain increases with the audio degradation, indicating that valuable and effective speech information is provided by the visual information. For SNR= -18 and -24 dB, the intelligibility gain provided by the inclusion of the facial animation is 25% and 10%, respectively, both

with a statistic significance $p < 0.001$. The corresponding gain of the real speaker's video amounts to 41% and 44%, respectively. These scores are well above the score by chance of 3.7%. For SNR = -12 dB, -6 dB and 0 db there is no statistical difference between the audio-only and audio+animation conditions.

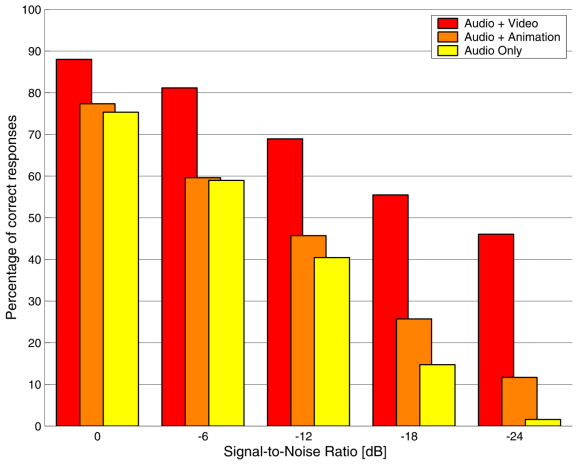

Figure 8: Percentage of correct responses.

A boxplot of the correct responses is shown in Figure 9. The 15 analysed cases are presented in the same order as in Figure 8. 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.

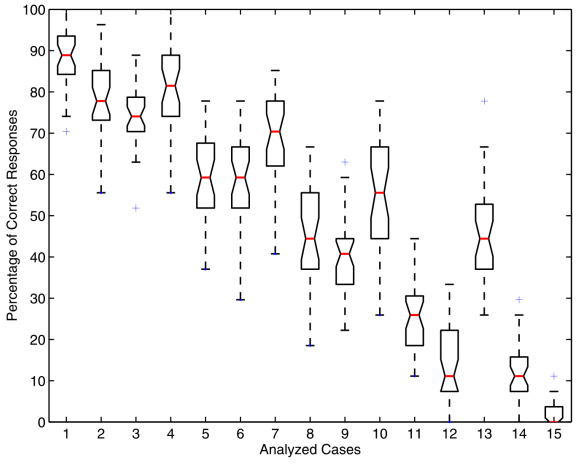

Figure 9: Boxplot of correct responses.

5. CONCLUSION & FUTURE WORK

The results so far 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 some room for improvement as it is not as effective as the real video. Some aspects of our approach still claim further elaboration. At least two aspects deserve special attention.

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 at normal video sampling rate (30 fps). 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. An improvement of our approach could be potentially achieved by the use of a muscle-based approach to drive the facial animation. Of course, 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 [LTW95]. We plan to test and compare in a near future the current version and the muscle-model version.

It is well known that viseme perception depends on talker, language, stimuli, subjects, response task, and environmental characteristics such as lighting, distance, and angle of observation. If all that is true for real visual speech, it gets worse in the context of facial animation. For facial animation, the coarticulation model and the strategy used to manipulate the virtual face also play an important role. It is not an easy task to compare and clearly identify the aspects that effectively contribute to the realism of speech synchronized facial animation. The systematic evaluation of the effective contribution to speechreading of a facial animation system seems to be an appropriate approach to the problem. In this paper we presented an evaluation of a facial animation system based on context-dependent visemes. The final result is a benchmark to be used for comparing developments using this and other approaches.

6. ACKNOWLEDGMENTS

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.

7. REFERENCES

[AJS02] I. Albrecht, J. Haber and H.-P. Seidel. Speech Synchronization for Physics-Based Facial Animation, In *Proceedings of the 10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision - WSCG '02*, Skala, V. (ed.), Union Agency, Plzen, Czech Republic, pp. 9-16, February 2002.

[Bes04] J. Beskow. Trainable Articulatory Control Models for Visual Speech Synthesis International. *Journal of Speech Technology*, 7(4), pp. 335-349, October 2004.

[BL98] C. Benoît and B. Le Goff. Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, 26(1-2), pp. 117-129, October 1998.

[BP82] A.-P. Benguerel and M.K. Pichora-Fuller. Coarticulation Effects in Lipreading. *Journal of Speech and Hearing Research*, Vol. 25, pp. 600-607, December 1982.

[CM93] M.M. Cohen and D.W. Massaro. Modeling Coarticulation in Synthetic Visual Speech. In *Models and Techniques in Computer Animation*, N. Magnenat-Thalmann and D. Thalmann (eds.), Springer-Verlag, pp. 139-156, 1993.

[DMV06] J.M. De Martino, L.P. Magalhães and F. Violaro. Facial Animation Based on Context-Dependent Visemes. *Computers and Graphics*, 30(6), December 2006 (accepted for publication).

[Jac88] P.L. Jackson. The Theoretical Minimal Unit for Visual Speech Perception: Visemes and Coarticulation. *The Volta Review*, 90(5), pp.99-115, 1998.

[JB71] J. Jeffers and M. Barley. *Speechreading (Lipreading)*, Charles C. Thomas Publisher, Springfield, Illinois, USA, 1971.

[Job92] J.D. Jobson. *Applied Multivariate Data Analysis – Volume II: Categorical and Multivariate Methods*, Springer-Verlag New York Inc., USA, 1992.

[IPA99] *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, International Phonetic Association, Cambridge University Press, 1999.

[LB96] B. Le Goff and C. Benoît. A Text-To Audiovisual-Speech Synthesizer for French. In *Proceedings of the 4th International Conference on Spoken Language Processing - ICSLP '96*, Vol.4, pp. 2163-2166, October 1996.

[LTW95] Y. Lee, D. Terzopoulos and K. Walters. Realistic face modeling and facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95*, pp. 55-62, 1995.

[Lof90] A. Löfqvist. Speech as Audible Gesture. In *Speech. Production and Speech Modeling*, W.J. Hardcastle and A. Marchal (eds.), Kluwer Academinc Publishers, pp. 289-322, 1990.

[MOC+05] D.W. Massaro, S. Ouni, M.M. Cohen and R. Clark. A Multilingual Embodied Conversational Agent. In *Proceedings of 38th Annual Hawaii International Conference on System Sciences - HICCS'05*, IEEE Computer Society, January 2-6 2005, CD-ROM 8 pages.

[Ohm67] S.E.G. Öhman. Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, 41(2), pp. 310-320, 1967.

[Par72] Parke, F. I. *Computer generated animation of faces*. Master's thesis, University of Utah, June 1972.

[Pel02] C. Pelachaud. Visual Text-to-Speech. In *MPEG-4 Facial Animation*, I.S. Pandzic and R. Forchheimer (eds), John Wiley and Sons, pp. 125-140, 2002. ISBN 0-470-84465-5.

[PBS96] C. Pelachaud and N.I. Badler and M. Steedman. Generating Facial Expressions for Speech. *Cognitive Science*, 20(1), pp. 1-46, January-March 1996.

[PO99] I.S. Pandzic, J. Ostermann and D. Millen. User evaluation: Synthetic talking faces for interactive services. *The Visual Computer*, 15(7-8), pp. 330-340, November 1999.

[RBB00] L. Revéret, G. Bailly and P. Badin. Mother: A new generation of talking heads providing a flexible control for video-realistic speech animation. In *Proceedings of the 6th International Conference on Spoken Language Processing - ICSLP'00*, Beijing, China, ISCA, pp. 755-758, 2000.

[SP54] W.H. Sumby and I. Pollack. Visual Contribution to Speech Intelligibility in Noise, *The Journal of the Acoustical Society of America*, 26(2), pp. 212-215, March 1954.