Animation of generic 3D Head models driven by speech

Lucas Terissi, Mauricio Cerda, Juan C. Gomez, Nancy Hitschfeld-Kahler, Bernard Girau, Renato Valenzuela

To cite this version:

Lucas Terissi, Mauricio Cerda, Juan C. Gomez, Nancy Hitschfeld-Kahler, Bernard Girau, et al.. Animation of generic 3D Head models driven by speech. IEEE International Conference on Multimedia and Expo - ICME 2011, Jul 2011, Barcelone, Spain. To appear. hal-00587016

HAL Id: hal-00587016

https://hal.archives-ouvertes.fr/hal-00587016

Submitted on 19 Apr 2011

ANIMATION OF GENERIC 3D HEAD MODELS DRIVEN BY SPEECH

Lucas Terissi[1] *Mauricio Cerda*[2] *Juan C. Gómez*[1]
Nancy Hitschfeld-Kahler[3] *Bernard Girau*[2] *Renato Valenzuela*[3]

[1]Lab. for System Dyn. & Signal Processing, Universidad Nacional de Rosario, CIFASIS, Argentina
[2]Loria - INRIA Nancy Grand Est, Cortex Team, Vandoeuvre-lès-Nancy, France
[3]Computer Science Department, FCFyM, Universidad de Chile, Santiago, Chile
{terissi, gomez}@cifasis-conicet.gov.ar, {Mauricio.Cerda, Bernard.Girau}@loria.fr, nancy@dcc.uchile.cl

ABSTRACT

In this paper, a system for speech-driven animation of generic 3D head models is presented. The system is based on the inversion of a joint Audio-Visual Hidden Markov Model to estimate the visual information from speech data. Estimated visual speech features are used to animate a simple face model. The animation of a more complex head model is then obtained by automatically mapping the deformation of the simple model to it. The proposed algorithm allows the animation of 3D head models of arbitrary complexity through a simple setup procedure. The resulting animation is evaluated in terms of intelligibility of visual speech through subjective tests, showing a promising performance.

***Index Terms*—** Facial Animation, Hidden Markov Models, Audio-Visual Speech Processing

1. INTRODUCTION

Animation of virtual characters is playing an increasingly important role due to the widespread use of multimedia applications such as computer games, online virtual characters, video telephony, and other interactive human-machine interfaces.

Several techniques have been proposed in the literature for facial animation. Among the main approaches, keyframe interpolation [1], direct parametrization, and muscles or physics based techniques [2], can be mentioned. On the other hand, the animation can be data-driven (e.g. by video, speech or text data) [3], manually controlled or a combination of both approaches. A thorough review of the different approaches for facial animation can be found in [4]. Most of the above mentioned animation techniques require a tedious and time-consuming preparation of the head model to be animated, in order to have control of the animation by a reduced set of parameters. For the particular case of speech-driven facial animation, the animation is controlled by a set of visual features estimated from speech, by means of a trained audio-visual model. Among the different approaches proposed in the literature to model audio-visual data, the ones based on Hidden Markov Models (HMM) have proved to yield realistic results when used in applications of speech driven facial animation. For speech-driven facial animation systems, Choi *et al* [5] have proposed a Hidden Markov Model Inversion (HMMI) method for audio-to-visual conversion. In HMMI, the visual output is generated directly from the given audio input and the trained HMM by means of an expectation-maximization (EM) iteration, leading to improvements in the estimation in comparison with other techniques based on Viterbi algorithm [6].

In this paper, a speech-driven animation system of generic 3D head models is proposed. A joint audio-visual Hidden Markov Model (AV-HMM) is trained using audio-visual data and then HMMI is used to estimate the visual features from speech data. These estimated visual features are used to animate a simple (small number of vertices) face model, which in turn is used to animate an arbitrary complex head model. This animation is performed by mapping and interpolating the movements of the vertices of the simple model to the ones of the complex model. An advantage of the proposed animation method is that it can be used to animate arbitrary head models, requiring a simple setup procedure.

2. SPEECH DRIVEN FACIAL ANIMATION SYSTEM

A block diagram of the proposed speech driven animation system is depicted in Fig. 1. An audiovisual database is used to estimate the parameters of a joint AV-HMM. This database consists of videos of a talking person. In the training stage, feature parameters of the audiovisual data are extracted. The audio part of the feature vector consists of mel-cepstral coefficients, while the visual part are the coefficients related to a set of facial movements. In the synthesis stage, the trained AV-HMM is used to estimate the visual features associated with a novel speech signal. These visual features, corresponding to different facial movements, allow the animation of a simple face model synchronized with the speech signal, which

Research partially supported by STIC-AmSud Project 09STIC06.

in turn is used to animate an arbitrary complex head model. This animation is performed by mapping and interpolating the movements of the vertices of the simple model to the ones of the complex model.

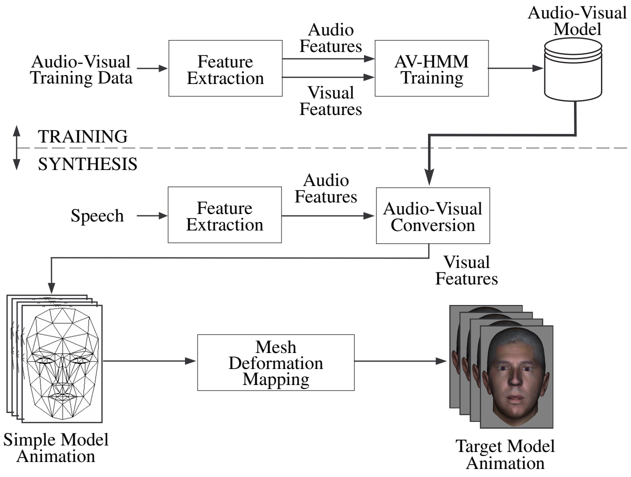

Fig. 1. Schematic representation of the speech driven animation system.

3. FEATURE EXTRACTION

The audio signal is partitioned in frames with the same rate as the video frame rate. A number of Mel-Cepstral Coefficients in each frame ($\mathbf{a}_t$) are used in the audio part of the feature vector. To take into account the audiovisual coarticulation, several frames are used to form the audio feature vector $\mathbf{o}_{at} = \left[\mathbf{a}_{t-t_c}^T, \ldots, \mathbf{a}_{t-1}^T, \mathbf{a}_t^T, \mathbf{a}_{t+1}^T, \ldots, \mathbf{a}_{t+t_c}^T\right]^T$ associated to the visual feature vector $\mathbf{o}_{vt}$.

Visual features are represented in terms of a simple 3D face model, namely *Candide-3* [7]. This 3D face model, depicted in Fig. 2(a), has been widely used in computer graphics, computer vision and model-based image-coding applications. The model defines two parameter vectors, denoted as $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$ to control its appearance and to perform facial movements, respectively. The values of $\boldsymbol{\sigma}$ are used to deform the model for the purposes of changing the position of the eyes, nose and mouth, making the mouth wider, etc. Similarly, the values of $\boldsymbol{\alpha}$ are used to control the movements of the mouth, eyes, eyebrows, etc. In this paper, the method proposed in [8] is used to extract visual features related with mouth movements during speech. This method allows the tracking of head pose and facial movements from videos. It also computes the values of vector $\boldsymbol{\alpha}$ associated to the facial movements of the person's face in the video. In Fig. 2(b), a frame of this tracking procedure is shown, where the animation of the *Candide-3* model is synchronized with the facial movements. The visual feature vector $\mathbf{o}_{vt}$ is composed by 4 components of vector $\boldsymbol{\alpha}$, related to the movements of the mouth.

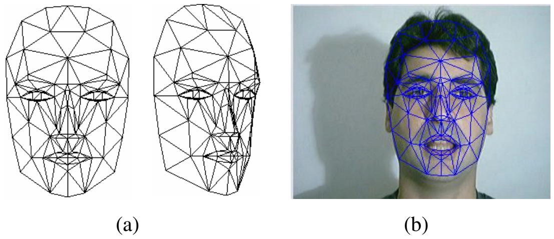

Fig. 2. *Candide-3* face model. (a) Triangular mesh. (b) Model synchronized with face movements.

4. AUDIO VISUAL MODEL

A joint AV-HMM is used to represent the correlation between the speech and facial movements. The AV-HMM, denoted as λ_{av}, is characterized by three probability measures, namely, the state transition probability distribution matrix ($\mathbf{A}$), the observation symbol probability distribution ($\mathbf{B}$) and the initial state distribution ($\boldsymbol{\pi}$), and a set of N states $S = (s_1, \ldots, s_j, \ldots, s_N)$, and audiovisual observation sequence $O_{av} = \{\boldsymbol{o}_{av1}, \ldots, \boldsymbol{o}_{avt}, \ldots, \boldsymbol{o}_{avT}\}$. The audiovisual observation $\boldsymbol{o}_{avt}$ is partitioned as $\boldsymbol{o}_{avt} \triangleq \left[\boldsymbol{o}_{at}^T, \boldsymbol{o}_{vt}^T\right]^T$, where $\boldsymbol{o}_{at}$ and $\boldsymbol{o}_{vt}$ are the audio and visual observation vectors, respectively. In addition, the observation symbol probability distribution at state j and time t, $b_j(\boldsymbol{o}_{avt})$, is considered a continuous distribution which is represented by a mixture of M Gaussian distributions

$$b_j(\mathbf{o}_{avt}) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(\mathbf{o}_{avt}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \qquad (1)$$

where c_{jk} is the mixture coefficient for the k-th mixture at state j and $\mathcal{N}(\mathbf{o}_{avt}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$ is a Gaussian density with mean $\boldsymbol{\mu}_{jk}$ and covariance $\boldsymbol{\Sigma}_{jk}$.

A single ergodic HMM is proposed to represent the audiovisual data. The model provides a compact representation of the audiovisual data, without the need of phoneme segmentation, which makes it adaptable to other languages.

4.1. AV-HMM Training

The training of the AV-HMM is carried out using an audiovisual database consisting of videos of a talking person. As described in section 3, audio-visual features are extracted from videos. Then, the audio-visual observation sequences O_{av} are used to estimate the parameters of an ergodic AV-HMM, as it is usual in HMM training, resorting to the standard Baum-Welch algorithm [9].

4.2. Audio-to-Visual Conversion

Choi and co-authors [5] used HMMI to estimate the visual features associated to audio features for the purposes of speech driven facial animation. Typically, it is assumed [5, 10] a diagonal structure for the covariance matrices of the Gaussian mixtures, invoking reasons of computational complexity. This assumption is relaxed in this paper allowing for full covariance matrices. This leads to more general expressions for the visual feature estimates.

The idea of HMMI for audio-to-visual conversion is to estimate the visual features based on the trained AV-HMM, in such a way that the probability that the whole audiovisual observation has been generated by the model is maximized, that is

$$\tilde{O}_v = \arg\max_{O_v} \{P(O_a, O_v|\lambda_{av})\} \qquad (2)$$

where O_a, O_v and $\tilde{O}_v$ denote the audio, visual and estimated visual sequences from $t = 1, \ldots, T$, respectively. It has been proved [9] that this optimization problem is equivalent to the maximization of Baum's auxiliary function $Q(\lambda_{av}, O_a, O_v; \lambda_{av}, O_a, \tilde{O}_v)$.The solution to this optimization problem can be computed by equating to zero the derivative of Q with respect to $\tilde{\mathbf{o}}_{vt}$, and it is given by

$$\tilde{\mathbf{o}}_{vt} = \left[\sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, j, k|\lambda_{av})\mathbf{\Phi}^v_{jk}\right]^{-1} \times$$
$$\times \sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, j, k|\lambda_{av}) \times$$
$$\times \left[\mathbf{\Phi}^v_{jk}\boldsymbol{\mu}^v_{jk} - \frac{1}{2}\mathbf{\Phi}^{va}_{jk}(\mathbf{o}_{at} - \boldsymbol{\mu}^a_{jk}) - \frac{1}{2}\left[(\mathbf{o}_{at} - \boldsymbol{\mu}^a_{jk})^T\mathbf{\Phi}^{av}_{jk}\right]\right] \qquad (3)$$

where

$$\mathbf{\Sigma}^{-1}_{jk} = \begin{bmatrix} \mathbf{\Phi}^a_{jk} & \mathbf{\Phi}^{av}_{jk} \\ \mathbf{\Phi}^{va}_{jk} & \mathbf{\Phi}^v_{jk} \end{bmatrix}$$

The derivation of this solution is not included here due to space limitations. For the case of diagonal matrices, $\mathbf{\Phi}^{va}_{jk}$ and $\mathbf{\Phi}^{av}_{jk}$ are zero matrices, and the expression (3) coincides with the one derived in [5].

Finally, given a sequence of acoustic observations O_a and the AV-HMM λ_{av}, the sequence of visual observations $\tilde{O}_v$ can be estimated by applying Eq. (3) to compute the visual parameters $\tilde{\mathbf{o}}_{vt}$ for each time t. These estimations are implemented in a recursive way, initializing the visual observation randomly.

5. ANIMATION

As described in section 3, visual feature vectors $\mathbf{o}_{vt}$ are composed by the values of a set of animation parameters of the *Candide-3* face model. Thus, the visual parameters estimated from speech can be used to animate this face model. The idea in this paper is to animate complex head models based on the speech-based animation of the *Candide-3* model. This is a difficult task due to the fact that, in contrast to the *Candide-3* model, realistic head models are very diverse and complex, usually defined by at least one complex textured triangular mesh, with possibly disconnected meshes for head, eyes, teeth, hair and tongue. In addition, the meshes corresponding to complex head models have a large number of vertices (10^4) in comparison to the *Candide-3* model (10^2). Moreover, the *Candide-3* model is not a head but a face model, so that some areas such as top and back of the head, neck and ears are not defined in the model. To overcame this difficulties, mesh deformation using a sub-set of points is one among the many available techniques [11] to perform animation. In particular, the use of the vertices of the *Candide-3* model as control points for the deformation of the target mesh is proposed in this paper.

The animation process for any head model has been divided into two sequential stages: (1) the projection or correspondence between each *Candide-3* vertex with one and only one vertex in the target head model and (2) the interpolation of the movement of any vertex in the target head model, using several projected points. In this section both stages are described.

5.1. Projection

The projection performs the mapping from each vertex of the *Candide-3* mesh to one (and only one) vertex in the head model. In order to perform this, both models are placed by the user in approximately the same position and orientation, see Figure 3, and then the anatomy of the *Candide-3* model (given by the parameter vector $\boldsymbol{\sigma}$) is adapted to the target head model. The adaptation process is rather simple but it is performed semi-automatically because there is a part of personal taste on how the anatomy is adapted, mainly when dealing with more abstract representations of the head (such as cartoons or inert objects). Despite this, the adaptation takes only a couple of minutes and it is sufficient to check head size/shape and lips/nose correspondences. It can be seen in Figure 3 how even a quick fit can deliver a good projection (red dots).

Once the *Candide-3* model is in place and it has been adapted using both position and deformation control sliders, see Figure 3, each vertex is projected into the target head mesh, selecting the closest one using the 3D Euclidean distance, and verifying that the same target head model vertex is not assigned to more than one *Candide-3* vertex. These special points in the head model are called "control points" and they are shown as red dots in Figure 3.

The projection stage links one vertex in the *Candide-3* model, with one vertex in the target head model. As the position and adaptation of both models is never perfectly performed by the user, the projection may have serious problems when the vertices forming a face (triangles in the surface mesh) of the *Candide-3* model are projected into disconnected

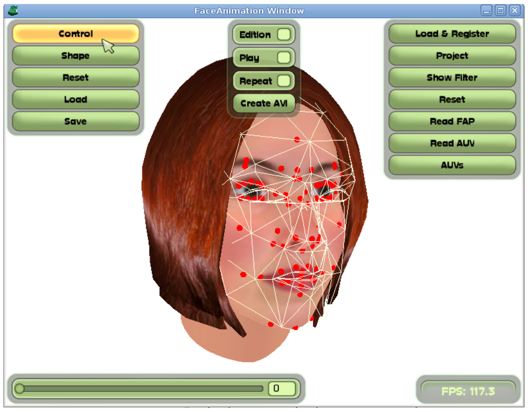

Fig. 3. Position and deformation of the *Candide-3* model for the Alice target head. *Candide-3* model is placed over the Alice head.

parts of the head models, *e.g.* one vertex of the superior lip get projected in the inferior lip because they are very close (in the Euclidean distance sense). To avoid this, the geodesic distance [12] between the vertices of each projected *Candide-3* triangle is checked, verifying that the distance is not too large.

5.2. Interpolation

Since control point positions in the target head model are available after the projection stage, the problem now is how to move the remaining points of the target head mesh (in the order of 10^4 points). The standard solution to this is to perform an interpolation taking into account the movements of the control points [13]. In this paper, the idea is to use the triangles that form the *Candide-3* model together with the control points in the target head model to form triangles in it. Then all vertices inside each triangle will depend (only) on the 3 vertices forming the triangle. To determine which is the triangle associated to each vertex, the projection onto the 2D texture map is employed, as can be seen in Fig. 4. Using this idea the interpolation can be expressed as,

$$\mathbf{v}_k = \gamma_{k1}\mathbf{z}_{i-1} + \gamma_{k2}\mathbf{z}_i + \gamma_{k3}\mathbf{z}_{i+1} \tag{4}$$

where $\mathbf{v}_k$ is the 3D position of the k-th vertex in the head model, expressed in terms of the coordinates of 3 control points ($\mathbf{z}_{i-1}$, $\mathbf{z}_i$ and $\mathbf{z}_{i+1}$). Considering Eq. (4) for the neutral position of the head model, the coefficient vector $\boldsymbol{\gamma}_k$ can be obtained as

$$\boldsymbol{\gamma}_k \triangleq [\gamma_{k1}, \gamma_{k2}, \gamma_{k3}]^T = \mathbf{Z}^{-1}\mathbf{v}_k \tag{5}$$

where $\mathbf{Z}$ is a matrix with the control points coordinates as columns. Once vector $\boldsymbol{\gamma}_k$ has been computed, the coordinates of vertex $\mathbf{v}_k$ can be obtained according to Eq. (4). It is not difficult to show that this interpolation method takes into account rigid rotations and translations of the model.

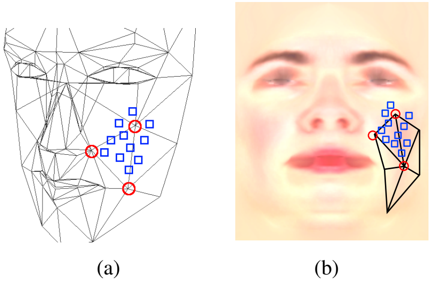
(a)
(b)

Fig. 4. (a) 3D *Candide-3*. (b) 2D texture map. Projected *Candide-3* vertices (red circles), some target head vertices (blue squares).

Special attention has to be paid to the case when some vertices of the complex model are not inside of any triangle of the simple model. This occurs in the areas of the neck, the back of the head or the interior part of the lips, which are not handled by the *Candide-3* model (see Fig. 2(a)). The details of the modification of the algorithms to cope this problem are not included here due to space limitations.

6. EXPERIMENTAL RESULTS

In this section, results concerning the proposed tracking and animation algorithms are included. In addition, the perceptual test to evaluate the quality of the animated avatars is described, and the results of this test are reported.

6.1. Audio-to-visual conversion

For AV-HMM training, videos of a person pronouncing a set of 120 sentences phonetically balanced were recorded at a rate of 30 frames per seconds, with a resolution of (320×240) pixels. The audio was recorded at 11025Hz synchronized with the video. Experiments were performed with AV-HMM with full covariance matrices, different number of states and mixtures in the ranges $[8, 40]$ and $[2, 7]$, respectively, and different values of the co-articulation parameter t_c in the range $[0, 7]$. In the experiments, the audio feature vector $\mathbf{a}_t$ is composed by the first eleven non-DC Mel-Cepstral coefficients, while the visual feature vector $\mathbf{o}_{vt}$ is composed with the 4 components of vector $\boldsymbol{\alpha}$ related to mouth movements. The performances of the different models were quantified by computing the Average Mean Square Error (AMSE)(ϵ), and the Average Correlation Coefficient (ACC)(ρ) between the true and estimated visual parameters, defined as

$$\epsilon = \frac{1}{4T}\sum_{r=1}^{4}\frac{1}{\sigma^2_{v_r}}\sum_{t=1}^{T}[\tilde{o}_{v_r t} - o_{v_r t}]^2 \tag{6}$$

$$\rho = \frac{1}{4T}\sum_{r=1}^{4}\sum_{t=1}^{T}\frac{(o_{v_r t} - \mu_{v_r})(\tilde{o}_{v_r t} - \tilde{\mu}_{v_r})}{\sigma_{v_r}\tilde{\sigma}_{v_r}} \tag{7}$$

where $o_{v_r t}$ is the value of the r-th visual parameter at time t and $\tilde{o}_{v_r t}$ its estimated value, μ_{v_r} and $\sigma^2_{v_r}$ denote the mean and variance of the r-th visual parameter in the time interval $[1, T]$, and $\tilde{\mu}_{v_r}$ and $\tilde{\sigma}^2_{v_r}$ denote the mean and variance of the r-th estimated visual parameter in $[1, T]$. For the quantification of the visual estimation accuracy, a separate audio-visual dataset, different from the training dataset, was employed.

The true and estimated visual parameters for the case of full covariance matrices with $N = 28$ states, $M = 5$ mixtures and co-articulation parameter $t_c = 3$ are represented in Fig. 5, where a good agreement can be observed. The AMSE and the ACC are for this case $\epsilon = 0.47$ and $\rho = 0.81$, respectively. This combination of number of states, number of mixtures and co-articulation parameter was among the ones that yield better results. Videos showing speech-driven animated avatars, using the proposed method, can be downloaded from http://www.fceia.unr.edu.ar/∼jcgomez/BAVI/videos/.

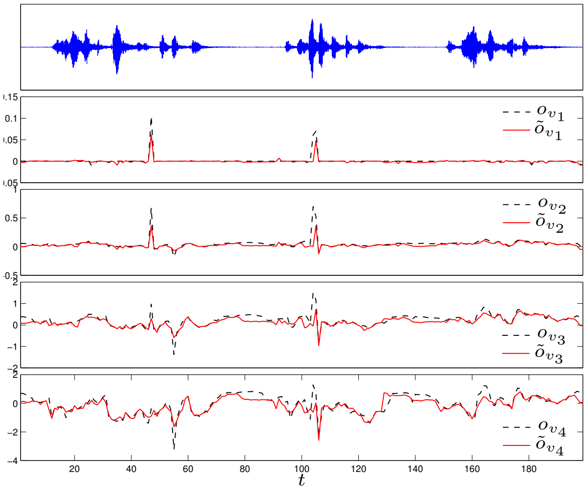

Fig. 5. Visual parameters estimated (lower four plots) for the audio signal consisting in 3 sentences (top plot).

6.2. Perceptual Evaluation

One way to evaluate the quality of animated avatars, is to analyze the visual contribution that the avatar provide to intelligibility of speech. This is usually measured through perceptual experiments. In these experiments, the recognition rate of a set of utterances (words, syllables or sentences) by a group of observers, is computed under at least two different conditions, namely, unimodal auditory and bimodal audio-visual conditions [14]. The same acoustic signals, corrupted by noise, are used in the unimodal and bimodal conditions. To actually measure the visual contribution of speech intelligibility of the avatar, the signal-to-noise ratio (SNR) has to be such that it makes it difficult to understand speech.

Table 1. Average intelligibility scores C_A, C_N and C_S, and the relative visual contribution C_V of the animated avatar driven by speech, for 3 different SNRs.

	C_A	C_N	C_S	C_V
-10dB	0.5704	0.8426	0.7071	0.7407
-15dB	0.4852	0.7944	0.6431	0.6796
-20dB	0.3704	0.7389	0.5919	0.5963

In this paper, perceptual tests were carried out to evaluate the avatar's visual contribution to speech intelligibility using a set of 27 consonant-vowel syllables, built by the combination of the sets $\{/b/, /d/, /f/, /k/, /m/, /p/, /s/, /t/, /\int/\}$ and $\{/a/, /i/, /u/\}$. During the test, the observer was presented with the utterances of the 27 syllables in random order, under only audio, natural audio-visual (original video), synthetic audio-visual animated from original audio, and synthetic audio-visual animated from true visual parameters conditions. Through a graphical user interface, the person had to indicate the perceived syllable within the set of 27 syllables. This was repeated for 3 different noise levels (SNR: -10dB, -15dB and -20dB), resulting in a total of $27 \times 4 \times 3 = 324$ syllables to be recognized in each test.

To evaluate the relative visual contribution of the animated avatar with respect to the real person visual contribution, the metric proposed in [14] was employed. This relative visual contribution (C_V) is defined as

$$C_V \triangleq 1 - \frac{C_N - C_S}{1 - C_A}, \tag{8}$$

where C_N, C_S and C_A are the bimodal natural face, bimodal synthetic face and unimodal auditory intelligibility scores, respectively. These scores are defined as the average, over all the participants, of the ratio between the number of correctly recognized syllables and the total number of syllables. As it is described in [14], this metric is designed to evaluate the performance of a synthetic talker compared to a natural talker when the acoustic channel is degraded by noise. The quality of the animated speech, measured by C_V, approaches the real visible speech as this measure increases from 0 to 1.

A group of 20 participants were enrolled in the perceptual experiments, with ages between 22 and 35. The group was conformed by 7 females and 13 males, reporting normal hearing and seeing abilities. The perceptual test results are summarized in Table 1, where the obtained values for C_A, C_N, and C_S and C_V for speech-driven animation are shown. As can be seen from Table 1, for each SNR condition (each row of the table), the recognition rates for the cases of bimodal natural (C_N) and synthetic (C_S) face stimulus are better than for the case of unimodal auditory (C_A) stimulus. This indicates that the visualization of the original video or the corresponding avatar animation during speech, leads to improvements in noisy speech intelligibility. As expected, the best recognition rate is obtained for the case of natural audio-visual stimulus. Figure 6 shows the average values of C_A, C_N and C_S, together with the corresponding plus-minus standard deviation intervals, for the three SNR values. The values of

C_S corresponding to the case of animating the avatar with the true visual parameters are almost indistinguishable from the ones corresponding to speech-driven animation. This would indicate that visual parameters estimation errors are not affecting significantly the final animation.

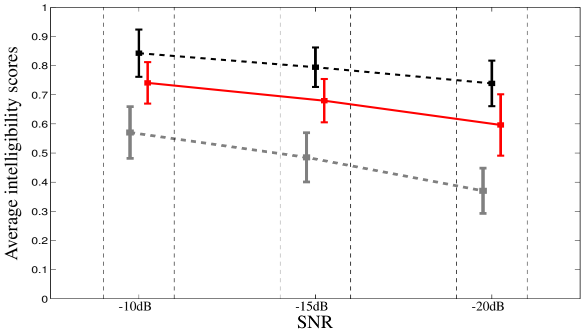

Fig. 6. Average intelligibility scores C_A (grey dashed), C_N (black dashed) and C_S (red solid) for the three SNRs.

Concerning the relative visual contribution of the animated avatar (C_V), the results in Table 1 indicate that the visual performance of the animated avatar reaches the 59%-73% of the visual performance of the natural face. A similar perceptual experiment is reported in [14], where the animation of the synthetic head *Baldi* [15], was performed using the Rapid Application Design (RAD) tools from the CSLU speech toolkit (http://cslu.cse.ogi.edu/toolkit/). The *Baldi* model was specifically designed and trained to be animated synchronized with speech. The animation was performed by Viterbi aligning and manually adjusting the model's facial movements to match the real speaker phonemes pronunciation. In [14], the relative visual contribution of the *Baldi* model was in the range 80%-90%. Compared to the above mentioned results in [14], the results in the present paper (Table 1) are promising taking into account that the proposed speech-driven facial animation technique does not require phoneme segmentation (language independent), and the animation requires a simple calibration stage to animate an arbitrary generic head model.

7. CONCLUSIONS

In this paper, a system for speech-driven animation of generic 3D head models was presented. A joint AV-HMM was proposed to represent the audio-visual data and an algorithm for HMM inversion was derived for the estimation of the visual parameters, considering full covariance matrices for the observations. Estimated visual speech features were used to animate a simple face model, which in turn was employed to animate a complex head model by automatically mapping the deformation of the simple model to it. The proposed animation technique requires a simple setup procedure. The resulting animation was evaluated in terms of intelligibility of visual speech through subjective tests. The perceptual quality of the animation proved to be satisfactory, showing that the visual information provided by the animated avatar improves the recognition of speech in noisy environments.

8. REFERENCES

[1] F. Pighin, J. Hecker, D. Lischinski, R. Szerisky, and D. Salesin, "Synthezing realistic facial expressions from photographs," in *Proc. of ACM SIGGRAPH98*, San Antonio, 1998, pp. 75–84.

[2] B. Choe, H. Lee, and H.S. Ko, "Performance-driven muscle-based facial animation," *Journal of Visualization and Computer Animation*, vol. 12, no. 2, pp. 67–79, May 2001.

[3] A. Savrana, L. M. Arslana, and L. Akarunb, "Speaker-independent 3D face synthesis driven by speech and text," *Signal Processing*, vol. 86, no. 10, pp. 2932–2951, January 2006.

[4] N. Ersotelos and F. Dong, "Building highly realistic facial modeling and animation: a survey," *Visual Computer*, vol. 28, pp. 13–30, 2008.

[5] K. Choi, Y. Luo, and J.N. Hwang, "Hidden Markov Model inversion for audio-to-visual conversion in an MPEG-4 facial animation system," *Journal of VLSI Signal Processing*, vol. 29, no. 1-2, pp. 51–61, 2001.

[6] S. Fu, R. Gutierrez-Osuna, A. Esposito, P.K. Kakumanu, and O.N. Garcia, "Audio/visual mapping with cross-modal Hidden Markov Models," *IEEE Trans. on Multimedia*, vol. 7, no. 2, pp. 243–252, April 2005.

[7] Jörgen Ahlberg, "An updated parameterized face," Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden, 2001.

[8] L. D. Terissi and J. C. Gómez, "3D head pose and facial expression tracking using a single camera," *Journal of Universal Computer Science*, vol. 16, no. 6, pp. 903–920, 2010.

[9] L. E. Baum and G. R. Sell, "Growth functions for transformations on manifolds," *Pacific Journal of Mathematics*, vol. 27, no. 2, pp. 211–227, 1968.

[10] L. Xie and Z.Q. Liu, "A coupled HMM approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, pp. 2325–2340, 2007.

[11] N. Ersotelos and F. Dong, "A survey of computer vision-based human motion capture," *The Visual Computer*, vol. 24, no. 1, pp. 13–30, 2008.

[12] S. Fortune, "Voronoi diagrams and Delunay triangulations," *Handbook of discrete and computational geometry*, pp. 377–388, 1997.

[13] Koray Balci, "XfacEd: authoring tool for embodied conversational agents," in *ICMI*, G. Lazzari, F. Pianesi, J. L. Crowley, K. Mase, and S. Oviatt, Eds. 2005, pp. 208–213, ACM.

[14] S. Ouni, M. Cohen, H. Ishak, and D. Massaro, "Visual contribution to speech perception: Measuring the intelligibility of animated talking heads," *EURASIP Journal on Audio, Speech and Music Processing*, no. Article ID 47891, 12 pages, 2007, DOI:10.1155/2007/47891.

[15] M. Cohen, D. Massaro, and R. Clark, "Training a talking head," in *Proc. of the IEEE Fourth International Conference on Multimodal Interfaces*, Pittsburgh, PA, 2002, pp. 499–510.