EURASIP Journal on Audio, Speech, and Music Processing

# A comprehensive system for facial animation of generic 3D head models driven by speech

*EURASIP Journal on Audio, Speech, and Music Processing* 2013, **2013**:5 doi:10.1186/1687-4722-2013-5

Lucas D Terissi (terissi@cifasis-conicet.gov.ar)
Mauricio Cerda (mauriciocerda@med.uchile.cl)
Juan C Gómez (gomez@cifasis-conicet.gov.ar)
Nancy Hitschfeld-Kahler (nancy@dcc.uchile.cl)
Bernard Girau (Bernard.Girau@loria.fr)

| | |
|---|---|
| **ISSN** | 1687-4722 |
| **Article type** | Research |
| **Submission date** | 31 July 2012 |
| **Acceptance date** | 8 January 2013 |
| **Publication date** | 1 February 2013 |
| **Article URL** | http://asmp.eurasipjournals.com/content/2013/1/5 |

# A comprehensive system for facial animation of generic 3D head models driven by speech

Lucas D Terissi*[1]
*Corresponding author
Email: terissi@cifasis-conicet.gov.ar

Mauricio Cerda[2]
Email: mauriciocerda@med.uchile.cl

Juan C Gómez[1]
Email: gomez@cifasis-conicet.gov.ar

Nancy Hitschfeld-Kahler[3]
Email: nancy@dcc.uchile.cl

Bernard Girau[4]
Email: Bernard.Girau@loria.fr

[1]Laboratory for System Dynamics & Signal Processing, Universidad Nacional de Rosario and CIFASIS, Rosario, Argentina

[2]SCIAN-Lab, Faculty of Medicine, Universidad de Chile, Santiago, Chile

[3]Computer Science Department, FCFyM, Universidad de Chile, Santiago, Chile

[4]Loria - INRIA Nancy Grand Est, Cortex Team, Vandoeuvre-lès-Nancy, France

## Abstract

A comprehensive system for facial animation of generic 3D head models driven by speech is presented in this article. In the training stage, audio-visual information is extracted from audio-visual training data, and then used to compute the parameters of a single joint audio-visual hidden Markov model (AV-HMM). In contrast to most of the methods in the literature, the proposed approach does not require segmentation/classification processing stages of the audio-visual data, avoiding the error propagation related to these procedures. The trained AV-HMM provides a compact representation of the audio-visual data, without the need of phoneme (word) segmentation, which makes it adaptable to different languages. Visual features are estimated from the speech signal based on the inversion of the AV-HMM. The 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, using a small number of control points for the interpolation. 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 perceptual tests, showing a promising performance. The computational complexity of the proposed system is analyzed, showing the feasibility of its real-time implementation.

## Keywords

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. In these approaches, 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 [5], by means of a trained audio-visual model [6]. Several methods have been proposed in the literature for computing the facial movements from a given speech input, such as the ones based on rules [7], vector quantization (VQ) [8], neural networks (NN) [9], nearest neighbor (KNN) [5], Gaussian mixture models (GMM) [10], and hidden Markov models (HMM) [11, 12]. Regardless of the method employed to model the audio-visual data and to compute facial movements from speech, the different approaches usually divide the training audio-visual data into *classes*. This clustering is performed taking into account the acoustic data, the visual data, or a combination of both. For each cluster, an audio-visual model is trained using the corresponding audio-visual data. The first step in the synthesis stage is the computation of the class that the novel audio data belongs to. The trained audio-visual model associated with the selected class is then used to estimate the corresponding visual parameters. In these approaches, the performance of the whole system is directly affected by errors derived by these classification procedures.

For instance, in [8] a static mapping between phonemes and visemes was proposed, where a phoneme recognition engine is used to segment the input speech, and then the recognized phonemes are directly mapped to lip shapes using a viseme codebook. In the study by Beskow et al. [13], in the context of the SynFace project [14], different rule-based and data-driven articulatory control models were proposed to animate the movements of the talking head. This system takes time-stamped recognized phonetic symbols as input and produces articulatory control parameter trajectories to drive the face model. In [15], word-based and phoneme-based methods were proposed, where for each cluster an HMM model is trained using audio-visual data, and then, the HMM model associated with the recognized phoneme (word) is used to compute facial animation parameters. Kshirsagar and Thalmann [16] proposed a syllable-based approach, where captured facial motions are categorized into syllable motions and the new speech animation is achieved by concatenating syllable motions optimally chosen from the syllable motion database. Gutierrez et al. [5, 17] used a training audio-visual database to generate a lookup table of audio-visual dynamic pairs. Given a new speech signal, they find the examples in the audio-visual lookup table of the $k$ closest audio parameters using the Euclidean distance, then, the visual parameter is computed as the average of the visual parameters associated with the $k$ closest neighbors. In [18] an approach based on Fused HMM is proposed. The visual parameters of the training audio-visual data are classified into several visual clusters by $k$-means method, and for each cluster two individual HMMs are trained independently using the associated audio and visual parameters, and also a Fused HMM is trained. In the synthesis stage, for each audio input, the audio HMM of each cluster and the Viterbi algorithm are used to get the best HMM state alignment. Then, the aligned audio HMM states and the visual cluster centers are used to find the best visual cluster for the input audio, and to compose new facial animation parameters. Similar strategies, in the sense that the audio-visual data is classified and a particular model is computed for each cluster, have been also proposed in the literature for applications such as emotion recognition [19] and for head motion synthesis [20, 21] from speech.

Some approaches not requiring an *a priori* classification of the audio-visual data for the training stage have been also proposed in the literature. These techniques have the advantage of avoiding the propagation of possible errors during the classification stage. For instance, in [22] a shared Gaussian process latent variable model (SGPLVM) is introduced to perform a mapping between facial motion and speech signal. A shared latent space is computed by maximizing the joint likelihood function of the audio and visual data sets, using Gaussian kernels. During the synthesis stage, intermediate latent points are obtained from the audio data, and then used to predict the corresponding visual data by means of the Gaussian process mapping. Visual data is represented in terms of active appearance models (AAMs) [23], trained with shape and texture data provided by a set of annotated prototype face images. Given a set of AAM parameters estimated from audio data, novel frames of the animation are generated by first reconstructing the shape and texture separately and then warping the texture to the shape. The limitation of this approach is that it does not actually animate 3D head models, but rather the resulting animation consists of a sequence of synthetic face images generated from the AAM. Another approach not requiring a priori classification, was introduced in the late 90's by Massaro and colleagues in [24], based on their text-to-speech driven talking head *Baldi* [25]. An artificial neural network (ANN) is trained with an only audio database and the associated visual parameters, where the visual parameters are not extracted from real videos but are computed from the corresponding audio transcriptions. Given a novel audio signal, the ANN produces as an output the set of estimated visual parameters to control *Baldi*'s animation. This approach is constrained to using *Baldi*'s movements as visual parameters, computed from the corresponding text transcriptions, not allowing the use of visual data from other speakers. The disadvantage of the technique in [25] is that it is only capable of animating *Baldi* 3D head model.

Among the different approaches to model audio-visual data, the ones based on HMM have proved to yield realistic results when used in applications of speech driven facial animation [18]. HMM-based methods have the advantage that context information can be easily represented by state-transition probabilities. Earlier approaches, such as the studies in [8, 10, 26, 27], resort to different HMM structures and require the use of Viterbi optimization algorithm [28] to determine a particular HMM state transition sequence for the training or synthesis stage. Due to the high noise sensitivity of Viterbi algorithm, the Viterbi search may be easily misguided by noise in the audio input. As a result, this dependency would lead to visual parameter predictions of limited quality [29]. To address this limitation, Choi et al. [15] have proposed a hidden Markov model inversion (HMMI) method for audio-visual conversion, which was originally introduced in [30] in the context of robust speech recognition. 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, thus avoiding the use of the Viterbi sequence and improving the performance of the estimation [11]. However, the HMMI-based approach proposed by Choi et al. [15] classifies the audio-visual data. In particular, word-based and phoneme-based methods were proposed, and an HMM model is trained for each cluster. In the synthesis stage, the input phoneme (word) is recognized and its associated HMM is used to compute the visual parameters by applying HMMI. Hence, this approach is susceptible to errors related to the classification procedures in the training and synthesis stages.

To assess the performance of the different speech-driven approaches proposed in the literature, quantitative and perceptual evaluations are usually employed. In the case of quantitative evaluations, the estimated visual parameters are compared with their associated ground truth parameters extracted from the database [6, 11, 15, 18, 29]. On the other hand, in the case of perceptual evaluations, the quality of the generated animations is judged by a group of persons. For instance, in [29, 31] the participants were asked about how natural and realistic are the animations in a five-point scale, and in [17] the individuals were asked to choose the more realistic animation between two different ones, generated by different methods. Other kind of perceptual evaluations are reported in [32, 33], where the contribution of the animated avatar to intelligibility of speech in noisy conditions is analyzed. This evaluation approach has the advantage of objectively quantify the perceived quality of the animation.

In this article, a complete pipeline for speech-driven animation of generic 3D head models is proposed. In the training stage, audio-visual information is extracted from audio-visual training data, and then used to compute the parameters of a single joint audio-visual hidden Markov model (AV-HMM). The proposed training method does not require prior segmentation or classification of the audio-visual data, avoiding in this way the errors derived from these classification procedures. The trained AV-HMM provides a compact representation of the audio-visual data, without the need of phoneme (word) segmentation, which makes it adaptable to different languages. In the synthesis stage, HMMI is used to estimate the visual features from speech data. An extension of the HMMI method in [15] is presented in this article, which allows the use of full covariance matrices for the Gaussian mixtures in the AV-HMM, leading to more general expressions to compute the associated visual features. A similar speech driven animation approach, presenting only preliminary results, was reported by the present authors in [34]. The experimental results presented in this article indicate that, in comparison with the method proposed in [15], the proposed extension of the HMMI method significantly reduces the computational load in the synthesis stage, making it more adequate for real-time applications. The estimated visual features are used to animate an arbitrary complex head model by means of the animation of a simple (small number of vertices) face model, requiring only a quick setup procedure. This is an advantage compared to most of the existing animation techniques that require a tedious and time-consuming preparation of the head model to be animated. The animations generated by the proposed system are perceptually evaluated in terms of intelligibility of visual speech. The results from these tests indicate that the visual information provided by the animated avatar improves the recognition of speech in noisy environments. The computational complexity of each stage of the proposed speech-driven animation system is also analyzed, showing the feasibility of its real-time implementation.

The rest of the article is organized as follows. An overview of the speech driven facial animation system is presented in Section 2. In Section 3, the proposed algorithm for feature extraction from videos is described. The AV-HMM is introduced in Section 4, where an HMMI algorithm for the general case of considering full covariance matrices for the audio-visual observations is also derived. The proposed technique for animating generic 3D head models driven by speech is presented in Section 5. In Section 6, experimental results regarding the accuracy of the visual features estimation from speech, the quality of the animations in terms of intelligibility of visual speech, and the computational complexity of the proposed algorithms, are presented. Finally, some concluding remarks and perspectives for future study are included in Section 7.

## 2 Speech driven facial animation system

A block diagram of the proposed speech driven animation system is depicted in Figure 1. An audio-visual 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 audio-visual data are extracted. The audio part of the feature vector consists of Mel-Cepstral Coefficients [35], while the visual part consists of parameters 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 (small number of vertices) face model synchronized with the speech signal, 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.

**Figure 1 Schematic representation of the speech driven animation system.**

In this article, the performance of the proposed system is evaluated on two different audio-visual databases, recorded by only one speaker, used for both the training and the testing stages. The quality of the animations is evaluated through perceptual tests on intelligibility of visual speech by a group of 20 individuals.

## 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 the current frame $t$, denoted as $\mathbf{a}_t$, are used in the audio part of the feature vector. To take into account the audio-visual co-articulation, $t_c$ preceding and $t_c$ subsequent frames are used to form the audio feature vector $\mathbf{o}_{at} = \left[\mathbf{a}^T_{t-t_c}, \ldots, \mathbf{a}^T_{t-1}, \mathbf{a}^T_t, \mathbf{a}^T_{t+1}, \ldots, \mathbf{a}^T_{t+t_c}\right]^T$ associated with the visual feature vector $\mathbf{o}_{vt}$ at the current frame $t$. The feature extraction strategy is schematically depicted in Figure 2. As already mentioned, no segmentation/classification of the training audio-visual data is required in the proposed approach.

**Figure 2 Schematic representation of the feature extraction strategy.**

Visual features are represented in terms of a simple 3D face model, namely *Candide-3* [36]. This 3D face model, depicted in Figure 3a, has been widely used in computer graphics, computer vision and model-based image-coding applications. The advantage of using the *Candide-3* model is that it is a simple generic 3D face model, adaptable to different real faces, that allows to represent facial movements with a small number of parameters, and, as it is proposed in this article, it also allows to animate more complex face models in a simple way. *Candide-3* 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 article, the method proposed by the present authors in [37] is used to extract visual features related to 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 with the facial movements of the person's face in the video. In Figure 3b, 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 formed by 4 of the components of vector $\boldsymbol{\alpha}$, related to the movements of the mouth ($\mathbf{o}_{v_1}$: upper lip vertical movement, $\mathbf{o}_{v_2}$: lip stretching, $\mathbf{o}_{v_3}$: mouth opening control, and $\mathbf{o}_{v_4}$ : lip corner movement). These parameters allow the representation of mouth movements during speech.

**Figure 3 *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 speech and facial movements. The AV-HMM, denoted as $\lambda_{av}$, is characterized by the state transition probability distribution matrix ($\mathbf{A}$), the observation symbol probability distribution ($\mathbf{B}$), the initial state distribution ($\boldsymbol{\pi}$), a set of $N$ states $S = (s_1, \ldots, s_j, \ldots, s_N)$, and the audio-visual observation sequence $O_{av} = \{\boldsymbol{o}_{av1}, \ldots, \boldsymbol{o}_{avt}, \ldots, \boldsymbol{o}_{avT}\}$. The audio-visual 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 as a continuous distribution which is represented by a mixture of $M$ Gaussian distributions

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

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 audio-visual data. The model provides a compact representation of the audio-visual data, without the need of phoneme segmentation, which makes it adaptable to different languages.

### 4.1 AV-HMM training

The training of the AV-HMM is carried out using an audio-visual 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 [38].

### 4.2 Audio-to-visual conversion

As mentioned before, HMMI is used to estimate the visual features associated with the input audio features. In particular, an extension of the HMMI method proposed in [15] is presented in this article. Typically, it is assumed a diagonal structure for the covariance matrices of the Gaussian mixtures [15, 29], which reduces the training complexity. This assumption is relaxed in this article allowing for full covariance matrices. This leads to more general expressions for the visual feature estimates, where the covariance matrices are not constrained to a particular structure. Experimental results, reported in Section 6, indicate that the proposed extension of the HMMI method using full covariance matrices, even though increases the training complexity, significantly reduces the computational load in the synthesis stage, making it more suitable for real-time applications.

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 audio-visual observation has been generated by the model is maximized, that is

$$\tilde{O}_v = \text{arg max}\{O_v P(O_a, O_v|\lambda_{av})\} \ , \tag{2}$$

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 [38] that this optimization problem is equivalent to the maximization of the following auxiliary function

$$
\begin{aligned}
&Q(\lambda_{av}, O_a, O_v; \lambda_{av}, O_a, \tilde{O}_v) \triangleq \\
&\triangleq \sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, j, k \mid \lambda_{av}) \ \log \ P(O_a, \tilde{O}_v, j, k \mid \lambda_{av}) \ , \\
&= \sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, j, k \mid \lambda_{av}) \\
&\left[ \log \ \pi_{j_0} + \sum_{t=1}^{T} \log \ a_{j_{t-1}j_t} + \sum_{t=1}^{T} \log \ \mathcal{N}(\mathbf{o}_{at}, \tilde{\mathbf{o}}_{vt}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) + \sum_{t=1}^{T} \log \ c_{jk} \right] \ , \qquad (3)
\end{aligned}
$$

where $\pi_{j_0}$ denotes the initial probability for state $j$ and $a_{j_{t-1}j_t}$ denotes the state transition probability from state $j_{t-1}$ to state $j_t$. The solution to this optimization problem can be computed by equating to zero the derivative of $Q$ with respect to $\tilde{\mathbf{o}}_{vt}$. Considering that the only term that depends on $\tilde{\mathbf{o}}_{vt}$ is the one in the last line, involving the Gaussians, this derivative can be written as

$$
\begin{aligned}
&\frac{\partial Q(\lambda_{av}, O_a, O_v; \lambda_{av}, O_a, \tilde{O}_v)}{\partial \tilde{\mathbf{o}}_{vt}} = \\
&= \sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, q_t = S_j, k_t = k|\lambda_{av}) \times \frac{\partial}{\partial \tilde{\mathbf{o}}_{vt}} \left[ \log \ \mathcal{N}(\mathbf{o}_{at}, \tilde{\mathbf{o}}_{vt}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \right] \ . \qquad (4)
\end{aligned}
$$

Equating to zero this derivative and considering that

$$
\log \ \mathcal{N}(\mathbf{o}_{avt}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) = \log \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_{jk}|^{1/2}} - \frac{1}{2} \begin{bmatrix} \mathbf{o}_{at} - \boldsymbol{\mu}^a_{jk} \\ \tilde{\mathbf{o}}_{vt} - \boldsymbol{\mu}^v_{jk} \end{bmatrix}^T \underbrace{\begin{bmatrix} \boldsymbol{\Phi}^a_{jk} & \boldsymbol{\Phi}^{av}_{jk} \\ \boldsymbol{\Phi}^{va}_{jk} & \boldsymbol{\Phi}^v_{jk} \end{bmatrix}}_{\boldsymbol{\Sigma}^{-1}_{jk}} \begin{bmatrix} \mathbf{o}_{at} - \boldsymbol{\mu}^a_{jk} \\ \tilde{\mathbf{o}}_{vt} - \boldsymbol{\mu}^v_{jk} \end{bmatrix} \ , \qquad (5)
$$

where $d$ is the dimension of $\mathbf{o}_{avt} \triangleq \left[\mathbf{o}^T_{at}, \tilde{\mathbf{o}}^T_{vt}\right]^T$, $|\Sigma|$ stands for the determinant of matrix $\Sigma$, and where the inverse of the covariance matrix $\boldsymbol{\Sigma}_{jk}$ has been partitioned, taking into account the dimensions of the visual and audio vectors, as

$$
\boldsymbol{\Sigma}^{-1}_{jk} = \begin{bmatrix} \boldsymbol{\Phi}^a_{jk} & \boldsymbol{\Phi}^{av}_{jk} \\ \boldsymbol{\Phi}^{va}_{jk} & \boldsymbol{\Phi}^v_{jk} \end{bmatrix} \ , \qquad (6)
$$

it is not difficult to show that at each time $t$, the estimated visual observation is given by

$$
\begin{aligned}
\tilde{\mathbf{o}}_{vt} \ = \ & \left[ \sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, j, k|\lambda_{av}) \boldsymbol{\Phi}^v_{jk} \right]^{-1} \\
& \times \sum_{j=1}^{N}\sum_{k=1}^{M} P(O_a, O_v, j, k|\lambda_{av}) \times \left[ \boldsymbol{\Phi}^v_{jk}\boldsymbol{\mu}^v_{jk} - \frac{1}{2}\boldsymbol{\Phi}^{va}_{jk}(\mathbf{o}_{at} - \boldsymbol{\mu}^a_{jk}) - \frac{1}{2}\left[(\mathbf{o}_{at} - \boldsymbol{\mu}^a_{jk})^T \boldsymbol{\Phi}^{av}_{jk}\right] \right] . \\
& \qquad (7)
\end{aligned}
$$

For the case of diagonal matrices, $\boldsymbol{\Phi}^{va}_{jk}$ and $\boldsymbol{\Phi}^{av}_{jk}$ are zero matrices, and the expression (7) coincides with the one derived in [15].

Finally, given a sequence of acoustic observations $O_a$ and the trained AV-HMM $\lambda_{av}$, the sequence of visual observations $\tilde{O}_v$ can be estimated by applying (7) 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. The estimation strategy is schematically depicted in Figure 4.

**Figure 4 Schematic representation of the proposed strategy for the estimation of the sequence of visual observations $\tilde{O}_v$ from a given sequence of acoustic observations $O_a$ and the trained AV-HMM $\lambda_{av}$.**

## 5 Animation

As described in Section 3, the visual feature vectors $\mathbf{o}_{vt}$ are composed of 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 article 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, 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 (on the order of $10^4$) in comparison to the *Candide-3* model (on the order of $10^2$) and commonly each part of the complex head model has an associated 2D texture. 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.

The animation process for the head model has been divided into two sequential stages: (*i*) the projection or correspondence between each *Candide-3* vertex with one vertex in the target head model and (*ii*) 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 vertex in the head model, a task commonly called registration.

Manual and semi-automatic methods have been proposed to perform the registration stage in face animation. Manual methods [39, 40] have the disadvantage of being slow and subject to error but with a desirable (tight) control over the control points correspondence. Semi-automatic methods, like [41], consist in the minimization of an energy function starting from an initial manual guess. In the study of Sanchez et al. [41] that function consists of three terms that relate the two face models: the distance between both models, bending and strain, these last two to ensure a smooth result. The result of the semi-automatic method seems promising, yet the control does not seem well suited for cartoon-like face models, where the animation artist could require a more precise manual control of exaggerated features (large mouth, disproportionate face) that do not agree with distance or smoothness criteria.

In order to have a fast yet fine-grained control of the registration process, a semi-automatic registration algorithm is presented. The first stage for the user is to place both models in approximately the same position and orientation, see Figure 5, and then the anatomy of the *Candide-3* model (given by the parameter vector $\boldsymbol{\sigma}$) is manually adapted to the target head model using 10 sliders. 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 5 how even a quick fit can deliver a good projection (red dots).

**Figure 5 Position and deformation of the *Candide-3* model for the Alice target head. *Candide-3* model is placed over the Alice head.**

Once the *Candide-3* model is in place and it has been adapted using both position and deformation control sliders (shown in Figure 5), each vertex of the *Candide-3* model is projected into the target head mesh. The projection is performed by selecting the closest vertex 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 5.

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 non-contiguous parts of the head model, e.g., one vertex of the superior lip gets projected in the inferior lip because they are very close (in the Euclidean distance sense).

To avoid this, a verification must be carried out for each *Candide-3* projected triangle vertex: along the geodesic (computed using a greedy algorithm) between any of these three vertices, the Euclidean distance must be a decreasing function. If this is not the case, one of the vertices must be replaced, selecting the next closest (in the 3D Euclidean distance sense) vertex. This procedure is performed until the condition for the three vertices is verified.

### 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 (on the order of $10^4$ points). The general solution to this is to perform an interpolation taking into account the movements of the control points.

In the literature, face animation has been addressed before in the context of interpolation, where a few control points determine the position of many other points. Among these approaches some authors have proposed to directly use a subset of close control points [40], radial basis functions [42] and variations of free-form deformation (FFD) adapted to surface meshes, like surface oriented FFD (SoFFD) [43], or more elaborated algorithms like planar-bones [41]. SoFFD is specially interesting because only three control points determine the position of each vertex in the complex model (simple case). The SoFFD applied to the *Candide-3* case corresponds to three points forming a triangle and its normal vector determining the position of close vertices in the complex mesh. In this study a variation of the SoFFD algorithm is presented, where the same three control points are used but also the texture coordinates to perform the assigning of control points. The use of the texture avoids the use of a discontinuity map as it was proposed in the planar-bones algorithm (a face adaptation of SoFFD). 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 three vertices forming the triangle. To determine which is the triangle associated with each vertex, the projection onto the 2D texture map is employed, as can be seen in Figure 6. Using this idea, the displacement of the position of the *k*th vertex in the target head model from its neutral position, can be expressed as,

$$\begin{aligned}\Delta \mathbf{v}_k &\triangleq \mathbf{v}_k - \mathbf{v}_{k0} \\ &= \gamma_{k1}\Delta\mathbf{z}_{k1} + \gamma_{k2}\Delta\mathbf{z}_{k2} + \gamma_{k3}\Delta\mathbf{z}_{k3} \ ,\end{aligned} \qquad (8)$$

where $\mathbf{v}_k$ is the 3D position of the *k*th vertex in the head model and $\mathbf{v}_{k0}$ is its neutral position, and $\Delta\mathbf{z}_{k1}$, $\Delta\mathbf{z}_{k2}$, and $\Delta\mathbf{z}_{k3}$ are the displacements of the three control points associated with $\mathbf{v}_k$, see Figure 6c. The coefficient vector $\boldsymbol{\gamma}_k$ can be obtained as

$$\boldsymbol{\gamma}_k \triangleq [\,\gamma_{k1}, \gamma_{k2}, \gamma_{k3}]^T = \mathbf{Z}^{-1}\mathbf{v}_{k0} \ , \qquad (9)$$

where $\mathbf{Z}$ is a matrix with the neutral position control point coordinates as columns. It is not difficult to show that this interpolation method takes into account rigid rotations of the model. The computation of vector $\boldsymbol{\gamma}_k$ is performed only once.

**Figure 6 Vertices interpolation.** **(a)** 3D *Candide-3*. **(b)** 2D texture map. Projected *Candide-3* vertices (red circles), some target head vertices (blue squares). **(c)** *k*th vertex and its associated control points.

The proposed interpolation requires to associate for each head mesh vertex a certain *Candide-3* triangle. As all vertices are 3D, the notion of being inside a certain *Candide-3* triangle has been defined as in 2D using the available texture information. This solution has the inconvenience of generating points that could be outside all *Candide-3* triangles, like for instance in the neck, the back of the head or the interior part of the lips. To handle these cases the following interpolation is proposed,

$$\Delta \mathbf{v}_k = \beta \sum_{i=1}^{3} \Delta\mathbf{z}_{ki} e^{-\frac{|d_i - d_m|^2}{d_m^2}} \ , \qquad (10)$$

where the gain $\beta$ is manually selected and fixed for all head mesh vertices, $\Delta\mathbf{z}_{k1}$, $\Delta\mathbf{z}_{k2}$, and $\Delta\mathbf{z}_{k3}$ are the displacements of the three closest directly connected vertices projected from the *Candide-3* mesh, and $d_m$ is the smallest value of $d_i$ with $i = 1, 2, 3$, using the Euclidean distance. High values of $\beta$ ($\beta > 0.6$) make the mesh to strictly follow the control points when animated, generating discontinuities. The opposite effect is obtained for low values ($\beta < 0.4$) where smoothness is assured but movements may not have the appropriate amplitude, thus an intermediate value ($\beta = 0.5$) was chosen in all animations.

## 6 Experimental results

In this section, results concerning the proposed algorithms for animation of 3D head models driven by speech are presented. In particular, results regarding the accuracy of visual feature estimation from speech are included. In addition, a perceptual test used to evaluate the quality of the animated avatars in terms of intelligibility of speech is described, and the corresponding results are analyzed. The computational complexity of the whole system is also studied at the end of this section.

### 6.1 Audio-to-visual conversion evaluation

In order to evaluate the proposed strategy for estimating the visual parameters from the speech signal, two audio-visual databases were compiled. One consists of videos of a person pronouncing the digits, and the other one consists of videos of the same person pronouncing a set of 120 phonetically balanced sentences. In both cases, the videos were recorded without compression, at a rate of 30 frames per second, with a resolution of 320 × 240 pixels, and the audio was recorded at 11025 Hz synchronized with the video. Audio-visual features were extracted from these databases as described in Section 3, and then used to train the AV-HMM. In particular, in the experiments performed in this article, the audio feature vector $\mathbf{a}_t$ is composed of the first eleven non-DC Mel-Cepstral coefficients, and the visual feature vector $\mathbf{o}_{vt}$ is formed by four components of vector $\boldsymbol{\alpha}$ related to mouth movements. The first database (digits utterances) was used to evaluate the influence of the structure of the covariance matrices (full or diagonal) on the visual estimation. To evaluate the performance of the proposed conversion algorithm in a more complex and realistic scenario, AV-HMMs were trained using the second database (120 phonetically balanced sentences).

The performances of the different models were quantified by computing the average mean square error (AMSE)($\epsilon$), and the average correlation coefficient (ACC)($\rho$) between the true and estimated visual parameters, defined as

$$\epsilon = \frac{1}{4T}\sum_{r=1}^{4}\frac{1}{\sigma_{v_r}^2}\sum_{t=1}^{T}\left[\tilde{o}_{v_r t} - o_{v_r t}\right]^2 , \tag{11}$$

$$\rho = \frac{1}{4T}\sum_{r=1}^{4}\sum_{t=1}^{T}\frac{(o_{v_r t} - \mu_{v_r})(\tilde{o}_{v_r t} - \tilde{\mu}_{v_r})}{\sigma_{v_r}\tilde{\sigma}_{v_r}} , \tag{12}$$

respectively, 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, $\mu_{v_r}$ and $\sigma_{v_r}^2$ denote the mean and variance of the $r$th visual parameter in the time interval $[\,1, T]$, respectively, and $\tilde{\mu}_{v_r}$ and $\tilde{\sigma}_{v_r}^2$ denote the mean and variance of the $r$th estimated visual parameter in $[\,1, T]$, respectively. For the quantification of the visual estimation accuracy, the models were trained using an 80 % of the training database, while the remaining 20 % was used for testing.

### *Influence of the covariance matrix structure*

To evaluate the influence of the structure of the covariance matrices on the audio-to-visual conversion algorithm, experiments were performed using AV-HMMs with diagonal and full covariance matrices, different number of states and mixtures in the ranges from 2 to 19, and from 3 to 20, respectively, and different values of the co-articulation parameter $t_c$ in the range from 0 to 7. This evaluation was performed using the digits database, where each digit was pronounced 40 times in different order. To train the AV-HMMs, digit utterances were randomly extracted from the database and then used to generate sequences of different duration, e.g., “7 5 4” and “2 0 3 9”. To maintain the natural coarticulation between digit utterances, the digits were segmented preserving silence frames at the beginning and at the end of the utterances. Through this procedure, a set of sequences is generated and used to train the models. In this way, the trained models are less sensitive to the order in which the digits appear in the sequences. Figure 7 shows the AMSE ($\epsilon$) and the ACC ($\rho$) as a function of the number of states and the number of mixtures for AV-HMMs with full and diagonal covariance matrices. In all the cases, the co-articulation parameter was set to $t_c = 3$, which proved to be the optimal value in the given range.

**Figure 7** ***AMSE* ($\epsilon$) and *ACC* ($\rho$) as a function of the number of states *N* and the number of mixtures *M*.** Where **(a)** and **(b)** correspond to the case of full covariance matrices and, **(c)** and **(d)** correspond to the case of diagonal covariance matrices. The big circles indicate the best accuracy for each case.

As can be observed in Figure 7a,b, for the case of full covariance matrices, as the number of states and the number of mixtures increase, the AMSE increases and the ACC decreases, indicating that the accuracy of the estimation deteriorates. This is probably due to the bias-variance tradeoff inherent to any estimation problem. The optimal values for the number of states and mixtures would be for this case $N = 4$ and $M = 3$, respectively, corresponding to $\epsilon = 0.47$ and $\rho = 0.75$. A similar accuracy can be obtained with

$N = 7$ and $M = 3$, corresponding to $\epsilon = 0.48$ and $\rho = 0.76$. On the other hand, Figure 7c,d shows the AMSE ($\epsilon$) and ACC ($\rho$) obtained for the case of AV-HMMs with diagonal covariance matrices. It can be observed that a similar accuracy, to the case of full covariance matrices, can be obtained using a larger number of states or mixtures. The optimal values for the number of states and mixtures would be for this case $N = 10$ and $M = 7$, respectively, corresponding to $\epsilon = 0.49$ and $\rho = 0.75$, or $N = 19$ and $M = 3$, corresponding to $\epsilon = 0.48$ and $\rho = 0.76$. Thus, there is no significant influence of the structure of the AV-HMM's covariance matrices in the visual parameters estimation accuracy. However, it must be noted that the number of states and mixtures of the AV-HMM affect the computational complexity during the training and synthesis stages. Since the training stage is performed only once and carried out off-line, this does not represent a problem. During the synthesis stage, the computation of the visual estimation for a sequence of length $T$-frames takes on the order of $MN^2T$ operations. Table 1 summarizes the computational complexity ($MN^2T$) associated with the configurations corresponding to the best accuracies for the cases of full and diagonal covariance matrices. Considering these results, the computational load for the case of AV-HMMs with full covariance matrices ($N = 4$ and $M = 3$, or $N = 7$ and $M = 3$) is significantly less expensive (up to 20 times faster) than for the case of considering AV-HMMs with diagonal covariance matrices ($N = 10$ and $M = 7$, or $N = 19$ and $M = 3$). The above arguments would indicate that the structure of the covariance matrices of the AV-HMM does not affects significantly the accuracy of the visual estimation but it affects the computational complexity during the synthesis stage. From this point of view, the use of AV-HMMs with full covariance matrices is preferable for faster implementations.

**Table 1 Computational complexity ($MN^2T$) associated with the best accuracies for the cases of full and diagonal covariance matrices**

| Cov. matrix structure | $N$ | $M$ | $\epsilon$ | $\rho$ | $MN^2T$ (per frame, $T = 1$) |
|---|---|---|---|---|---|
| **Full (proposed)** | 4 | 3 | 0.47 | 0.75 | **48** |
| **Full (proposed)** | 7 | 3 | 0.48 | 0.76 | **147** |
| Diagonal | 10 | 7 | 0.49 | 0.75 | 700 |
| Diagonal | 19 | 3 | 0.48 | 0.76 | 1083 |

Results associated with the proposed configurations (full covariance matrices) are in bold.

### *Overall evaluation*

To evaluate the proposed audio-to-visual conversion algorithm in a more complex scenario, the second database, consisting of videos of a person pronouncing a set of 120 phonetically balanced sentences (uttered 3 times), was employed. Experiments were performed with AV-HMM with full covariance matrices, different number of states and mixtures in the ranges from 8 to 40, and from 2 to 7, respectively, and different values of the co-articulation parameter $t_c$ in the range from 0 to 7. These experiments do not reveal a particular trend about the values of the optimal parameters. The best values of $\epsilon$ (*AMSE*) and $\rho$ (*ACC*) were obtained for $20 \leq N \leq 35$ states and $3 \leq M \leq 7$ mixtures. In these experiments, the optimal value of the co-articulation parameter was $t_c = 3$, which corresponds to a co-articulation time[a] of 100 ms. Compared to the previous experiment, where a reduced vocabulary (digits utterances) was employed, the number of states needed in this case increases considerably due to the fact that a larger vocabulary is being represented by the AV-HMM. 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 Figure 8, where a good agreement between them can be observed. This combination of number of states, number of mixtures and co-articulation parameter was among the ones that yielded the best results. The AMSE and the ACC are for this case $\epsilon = 0.47$ and $\rho = 0.81$, respectively, over the testing audio-visual database.

**Figure 8 (Top plot) Audio signal corresponding to the following three Spanish utterances: "*Primero hay que aceptar y después optar*", "*Pepita pasó esta tarde*", "*Este atlas no es étnico*". (Lower four plots) True (dashed line) and estimated (solid line) visual parameters associated with the audio signal, $\mathbf{o}_{v_1}$: upper lip vertical movement, $\mathbf{o}_{v_2}$: lip stretching, $\mathbf{o}_{v_3}$: mouth opening control, and $\mathbf{o}_{v_4}$ : lip corner movement.**

A fair comparison, in terms of AMSE($\epsilon$) and ACC ($\rho$), between the proposed audio-to-visual conversion algorithm and previous research is not an easy task since there is neither a common audio-visual corpus for evaluation, nor a common quality metric [15, 17, 29, 31, 44, 45]. However, some conclusions can be drawn in relation to the strategy used for the estimation. As already mentioned, in this article the audio-visual training data is represented by a single joint AV-HMM. This is in contrast to other methods in the literature that classify the audio-visual data, and train an individual model for each class. In these cases, the visual predictions are strongly influenced by the results of the classification procedures in the training and synthesis stages. Furthermore, the use of a classification step in the synthesis stage, usually requires a smoothing procedure to concatenate the visual parameters estimated from each class [18]. In the case of the present study, the classification stages are not required since the predicted parameters are computed from a single AV-HMM, which implicitly incorporate the smoothing of the estimated visual parameters. As a consequence, the proposed strategy leads to faster and more stable predictions.

Videos showing speech-driven animated avatars generated with the proposed method are included with the additional files associated with this article, [see Additional file 1 and Additional file 2]. Video *movie1.mpg*, in [Additional file 1], shows the animation of three different 3D head models, see Figure 9, driven by three different speech signals. These animations were generated with the same AV-HMM trained with a single speaker. On the other hand, video *movie2.mpg*, in [Additional file 2], shows the simultaneous animation of two of the avatars driven by the same speech signal.

**Figure 9 Virtual head models. (a)** *Juan*, **(b)** *Alice*, and **(c)** *Jhonny*.

### 6.2 Perceptual evaluation

One way to evaluate the quality of animated avatars, is to analyze the visual contribution that the avatar provides 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 [33]. 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.

In this article, perceptual tests were carried out to evaluate the avatar's visual contribution to speech intelligibility using a set of Spanish consonant-vowel syllables. In this experiment, three different presentation conditions were considered: (a) unimodal auditory, (b) bimodal natural talker, and (c) bimodal synthetic talker.

***Participants***

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. All the participants were right handed and spoke Spanish as their native language.

### *Test stimuli*

A set of 27 consonant-vowel syllables, built by the combination of the sets $\{/p/, /b/, /m/, /f/, /t/, /d/, /s/, /\int/, /k/\}$ and $\{/a/, /i/, /u/\}$, was used as stimuli in these tests. These syllables were presented under three conditions: unimodal auditory, bimodal natural talker and bimodal synthetic talker. This was repeated using the original acoustic signals corrupted with three different white noise levels, corresponding to three different signal-to-noise ratios (SNR: $-10$, $-15$, and $-20$ dB). Thus, each participant was presented to a total of $27 \times 3 \times 3 = 243$ syllables to be recognized.

### *Presentation*

For each condition, the observer was presented with the utterances of the 27 syllables in random order. Through a graphical user interface, the person had to indicate the perceived syllable within the set of 27 syllables. The natural and synthetic talkers are shown in Figure 10. The visual stimulus, for both the natural and synthetic talkers, were presented in the center of the graphical user interface, by means of uncompressed videos with a resolution of 640 × 480 pixels. The auditory speech was taken from the natural talker videos. The synthetic face animations were generated using the original speech signal from the natural speaker as input of the facial animation system proposed in this article. The system was trained using the phonetically balanced sentences database.

**Figure 10 Natural and synthetic speaker used in the perceptual evaluation.**

### *Results*

To evaluate the relative visual contribution of the animated avatar with respect to the real person visual contribution, the metric proposed in [33] was employed. This relative visual contribution ($C_V$) is defined as

$$C_V \triangleq 1 - \frac{C_N - C_S}{1 - C_A} , \tag{13}$$

where $C_N$, $C_S$, and $C_A$ are the bimodal natural face, bimodal synthetic face (generated from estimated visual parameters) and unimodal auditory intelligibility scores, respectively. These scores are defined as the ratio between the number of correctly recognized syllables and the total number of syllables. As it is described in [33], 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.

The perceptual test results are summarized in Table 2, where the obtained average values (over all participants) for $C_A$, $C_N$, and $C_S$ and $C_V$, denoted as $\overline{C}_A$, $\overline{C}_N$, $\overline{C}_S$, and $\overline{C}_V$, respectively, are shown. As can be seen from Table 2, for each SNR condition (each row of the table), the recognition rates for the cases of bimodal natural ($\overline{C}_N$) and synthetic face ($\overline{C}_S$) stimuli are better than for the case of unimodal auditory ($\overline{C}_A$) 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.

**Table 2 Average intelligibility scores $\overline{C}_A$, $\overline{C}_N$, and $\overline{C}_S$, and the average relative visual contribution $\overline{C}_V$ of the animated avatar driven by speech, for three different SNRs**

| | $\overline{C}_A$ | $\overline{C}_N$ | $\overline{C}_S$ | $\overline{C}_V$ |
|---|---|---|---|---|
| −10 dB | 0.5704 | 0.8426 | 0.7407 | 0.7628 |
| −15 dB | 0.4852 | 0.7944 | 0.6796 | 0.7770 |
| −20 dB | 0.3704 | 0.7389 | 0.5963 | 0.7735 |

Boxplots for the intelligibility scores $C_A$, $C_N$, and $C_S$, for all the participants and the three different noise conditions are shown in Figure 11. As it is customary, the top and bottom of each box are the 75th and 25th percentiles of the samples, respectively. The line inside each box is the sample median, while the circle is the sample mean, over all the participants. The notches display the variability of the median between samples. The width of a notch was computed so that box plots whose notches do not overlap have different medians at the 5 % significance level. As can be observed, for each noise level, the notches corresponding to unimodal auditory ($C_A$) and bimodal synthetic face ($C_S$) intelligibility scores do not overlap, which is a visual indication that the difference between the corresponding medians are statistically significant at the 5 % significance level. The same holds for the notches corresponding to unimodal auditory ($C_A$) and bimodal natural face ($C_N$) intelligibility scores. The $p$-values [46] of the significance tests between $C_A$, $C_S$, and $C_N$, for the three different noise levels are in all the cases less than 0.0001 ($p < 0.0001$). These values support the claim that the difference between the medians are statistically significant at the 5 % significance level.

**Figure 11 Boxplots of intelligibility scores $C_A$ (grey), $C_S$ (solid black), and $C_N$ (dashed black) for the three SNRs.**

Figure 12 depicts the average syllable recognition rates, pooled across participants, vowels and the three noise levels, for the unimodal auditory, bimodal synthetic and bimodal natural conditions, ordered by the place of articulation from front to back. It can be noted that there are global improvements of the syllable recognition rates between unimodal auditory and bimodal synthetic conditions. In particular, there are significant improvements for the syllables starting with phonemes $/p/$, $/d/$, $/m/$, $/s/$, and $/\int/$, and partial improvements for the ones starting with phonemes $/t/$, $/k/$ and $/f/$. As expected, there is no important improvement on the recognition rate for both bimodal synthetic and bimodal natural conditions, for the case of phoneme $/k/$, since the articulatory production of this phoneme is not visible. It can also be observed that for the case of phoneme $/b/$, there is no improvement, but a deterioration, in the bimodal synthetic condition over the unimodal auditory condition, indicating that there are some problems in the animation of the syllables starting with this phoneme. This figure also shows that the animation of syllables starting with phonemes $/p/$, $/b/$, and $/f/$ needs some improvements, since there are important differences between the recognition rates for the cases of bimodal synthetic and bimodal natural conditions.

**Figure 12 Average syllable recognition rates grouped by initial consonant.** The bar graph depicts the average recognition rates pooled across all vowels, participants and noise levels, for the cases of unimodal auditory, bimodal synthetic, and bimodal natural conditions.

The syllable confusion matrices, pooled across participants, vowels and noise levels, for the cases of bimodal synthetic (left) and bimodal natural (right) conditions, are shown in Figure 13. It can be observed from Figure 13, that the deterioration in the recognition rate for syllables starting with phoneme $/b/$, mentioned in the previous paragraph, is due to the fact that phoneme $/b/$ is partially perceived as phoneme $/d/$. This could be caused by the fact that there are no tongue movements in the animation, which are important to visually disambiguate these phonemes. It can also be observed that syllables starting with phonemes $/s/$ and $/\int/$ are mutually confusable for both presentation conditions. In addition, the phoneme $/s/$ is partially confused with the phoneme $/t/$. This phenomenon is more noticeable for the bimodal synthetic condition. A possible explanation for this could be the masking of these phonemes by the Gaussian white noise present in the audio signal. For the case of syllables starting with $/t/$ and $/d/$, for both bimodal natural and synthetic conditions, the recognition rate for $/t/$ is worse than for $/d/$, in spite of the fact that there are few visual differences between both phonemes. Similarly, this difference in the recognition rates could be related to a masking effect, produced by the Gaussian noise present in the auditory stimuli, which is more noticeable for the perception of the voiceless consonant $/t/$ than for the voiced consonant $/d/$.

**Figure 13 Pooled syllable confusion matrices, averaged across vowels, participants, and noise levels.** Left: Bimodal synthetic condition. Right: Bimodal natural condition.

Figure 14 shows the confusion matrix for all the consonant-vowels syllables, for the bimodal synthetic condition, where a more detailed information about the confusability among the syllables is presented. For instance, it can be observed that the confusion between phoneme $/b/$ and $/d/$ is more accentuate for the case of vowels $/i/$ and $/u/$, and it is not that important for the case of vowel $/a/$.

**Figure 14 Syllables confusion matrix for bimodal synthetic condition.**

As pointed out above, the system presents some problems in the animation of syllables starting with $/b/$ and $/p/$. However, it is important to note that the overall improvement in the intelligibility is reasonably good, considering that the system is trained in an unsupervised way, and that it does not perform a separate training for each phoneme.

Concerning the relative visual contribution of the animated avatar ($C_V$), the results in Table 2 indicate that the visual performance of the animated avatar is around 77 % of the visual performance of the natural face for the three different noise levels. Note that as stated in [33], the relative visual contribution of the animated avatar remains approximately invariant over the different noise levels. A similar perceptual experiment is reported in [33], where the animation of the synthetic head *Baldi* [47], 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 [33], the relative visual contribution of the *Baldi* model was in the range 80–90 %. Compared to these results, the results in the present article (Table 2) 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.

### 6.3 Implementation issues

In this section, the computational complexity for each stage of the proposed speech-driven animation system is analyzed. Since the training of the AV-HMM is performed off-line, only the complexity of the synthesis stage is reported.

#### *Feature extraction*

As described in Section 3, the audio feature vector of the input speech signal consists of the concatenation of the Mel-Cepstral Coefficients of $(2t_c + 1)$ frames. However, only one new Mel-Ceptral computation is needed for each frame. The complexity of this computation is bounded above by the corresponding to the fast Fourier transform (FFT), *viz.*, $O(N_F \log_2(N_F))$, where $N_F$ is the number of frequencies in the FFT.

### *Audio-to-visual conversion*

As described in Section 4.2, the visual feature vector is estimated according to (7) in a recursive way. Since matrices $\mathbf{\Phi}^a_{jk}$, $\mathbf{\Phi}^{av}_{jk}$, $\mathbf{\Phi}^{va}_{jk}$, and $\mathbf{\Phi}^v_{jk}$, and vector $\boldsymbol{\mu}^a_{jk}$, have been pre-computed in the training stage, the computation in (7) for a sequence of length $T$-frames, takes on the order of $MN^2T$ operations [48], where, as already defined, $N$ is the number of states of the AV-HMM, and $M$ is the number of mixtures.

### *Animation*

As described in Section 5, the proposed algorithm is composed of two stages: projection (registration) and interpolation. One important aspect of the presented method is that a linear interpolation model allows for fast calculations. Assuming that the *Candide-3* mesh has $N_C$ vertices and the 3D head mesh has $N_T$ vertices, and that the interpolation vector $\boldsymbol{\gamma}$ is known, any face deformation requires $O(N_T)$ operations (points are in 3D). The computation of $\boldsymbol{\gamma}$ for each vertex is more expensive but it is performed only once. To compute $\boldsymbol{\gamma}$, first the closest vertex must be computed in $O(N_T N_C)$ and then the topology must be checked. In the worst case this may imply again $O(N_T N_C)$ operations. After the projection step, it must be verified whether each 3D head mesh vertex is inside some *Candide-3* triangle. This is also done in $O(N_T N_C)$ operations. For the case when a vertex is not inside any *Candide-3* triangle a new distance must be computed. This step could need in the worst case $O(N_T N_C)$ operations. Then, the complete algorithm would take $O(N_T(N_C + 1))$ operations. However, the operations can be reduced to $O(N_C + N_T)$ by using data structures where close vertices can be found in constant time.

The whole system was implemented and tested in a dual-core 2.5 GHz processor, 2 GB RAM, personal computer. Considering a 33 ms-length frame of the speech signal (30 frames per second), sampled at 11025 Hz, the feature extraction and audio-to-visual stages take about 18 ms per frame, where most of the computational time is employed for the visual estimation stage (feature extraction stage takes about 0.1 ms). The corresponding animation stage is capable of rendering up to 150 frames per seconds (about 6.6 ms). This result in approximately 24.6 ms per frame processing time, which is smaller than the speech signal frame rate. It must be noted that the system has a constant time delay to take into account the co-articulation in the audio-visual features. For the optimal value of the co-articulation parameter $t_c = 3$ set in this article, the system has a time delay of approximately 100 ms. These computational times are suitable for a real-time operation of the system.

## 7 Conclusions

In this article, a comprehensive system for facial animation of generic 3D head models driven by speech was presented. In contrast to most of the methods in the literature, the proposed approach does not require segmentation/classification processing stages of the audio-visual data, avoiding the error propagation related to these procedures, resulting in faster and more stable visual feature predictions. 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. The model provides a compact representation of the audio-visual data, without the need of phoneme segmentation, which makes it adaptable to different languages. 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. 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. The analysis of the computational complexity of the proposed algorithms shows that a real-time operation of the system is feasible.

The experimental results show that the performance of the proposed system is comparable to other studies of the state-of-the-art. The main advantage of this system is that it presents a complete automatic pipeline from audio recording to generic head animation. The results suggest that further improvements can be achieved by including a tongue model, in both training and synthesis stages, in order to disambiguate lip movements. Also, enhancing lips depth movements information is expected to improve the final animation. This could be accomplished by using multiple cameras for visual feature extraction. Work is in progress to incorporate new audio features such as prosody information and fundamental frequency. Regarding the visual features, additional ones related to facial expressions, the movements of other regions of the face, such as eyebrows and checks, could be incorporated to improve the performance of the animation.

## Endnote

[a] In the literature, coarticulation times have been reported in the range from 100 to 200 ms and some authors also argue that it is language-dependent [17].

## Competing interests

The authors declare that they have no competing interests.

## Acknowledgments

This study was partially supported by STIC-AmSud Project 09STIC06, Millennium Scientific Initiative (ICM P09-015-F), Fondecyt Project No. 1120495, Chile, and the National University of Rosario and CIFASIS-CONICET, Argentina.

## References

1. F Pighin, J Hecker, D Lischinski, R Szerisky, D Salesin, in *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*. Synthesizing realistic facial expressions from photographs (San Antonio, TX, 1998). pp. 75–84

2. B Choe, H Lee, HS Ko, Performance-driven muscle-based facial animation. J. Visual. Comput. Animat. **12**(2), 67–79 (2001)

3. A Savrana, LM Arslana, L Akarunb, Speaker-independent 3D face synthesis driven by speech and text. Signal Process. **86**(10), 2932–2951 (2006)

4. N Ersotelos, F Dong, Building highly realistic facial modeling and animation: a survey. Visual Comput. **28**, 13–30 (2008)

5. R Gutierrez-Osuna, PK Kakumanu, A Esposito, O Garcia, A Bojorquez, JL Castillo, I Rudomin,

Speech-driven facial animation with realistic dynamics. IEEE Trans. Multimedia **7**, 33–42 (2005)

6. Z Deng, U Neumann, J Lewis, TY Kim, M Bulut, S Narayanan, Expressive facial animation synthesis by learning speech coarticulation and expression spaces. IEEE Trans. Visual. Comput. Graph. **12**(6), 1523–1534 (2006)

7. C Pelachaud, NI Badler, M Steedman, Generating facial expressions for speech. Cognitive Sci. **20**, 1–46 (1996)

8. E Yamamoto, S Nakamura, K Shikano, Lip movement synthesis from speech based on Hidden Markov Models. Speech Commun. **26**(1–2), 105–115 (1998)

9. P Hong, Z Wen, T Huang, Real-time speech-driven face animation with expressions using neural networks. IEEE Trans. Neural Netws. **13**(4), 916–927 (2002)

10. R Rao, T Chen, R Mersereau, Audio-to-visual conversion for multimedia communication. IEEE Trans. Indus. Electron. **45**, 15–22 (1998)

11. S Fu, R Gutierrez-Osuna, A Esposito, P Kakumanu, O Garcia, Audio/visual mapping with cross-modal Hidden Markov Models. IEEE Trans. Multimedia **7**(2), 243–252 (2005)

12. T Hazen, Visual model structures and synchrony constraints for audio-visual speech recognition. IEEE Trans. Audio Speech Lang. Process. **14**(3), 1082–1089 (2006)

13. J Beskow, I Karlsson, J Kewley, G Salvi, SynFace: a talking head telephone for the hearing-impaired. Computers helping people with special needs, Lecture Notes in Computer Science **3118**, 1178–1186 (2004)

14. SynFace Project. http://www.speech.kth.se/synface [Last visited on November 2012]

15. K Choi, Y Luo, J Hwang, Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system. J. VLSI Signal Process. **29**(1–2), 51–61 (2001)

16. S Kshirsagar, N Magnenat-Thalmann, Visyllable based speech animation. Comput. Graph. Forum **22**(3), 631–639 (2003)

17. P Kakumanu, A Esposito, ON Garcia, R Gutierrez-Osuna, A comparison of acoustic coding models for speech-driven facial animation. Speech Commun. **48**(6), 598–615 (2006)

18. J Tao, L Xin, P Yin, Realistic visual speech synthesis based on hybrid concatenation method. IEEE Trans. Audio Speech Lang. Process. **17**(3), 469–477 (2009)

19. J Tao, S Pan, M Yang, Y Li, K Mu, J Che, Utterance independent bimodal emotion recognition in spontaneous communication. EURASIP J. Adv. Signal Process. **4**(4) (2011)

20. C Busso, Z Deng, U Neumann, SS Narayanan, Natural head motion synthesis driven by acoustic prosodic features. J. Comput. Animat. Virtual Worlds **16**(3–4), 283–290 (2005)

21. C Busso, Z Deng, M Grimm, U Neumann, S Narayanan, Rigid head motion in expressive speech animation: analysis and synthesis. IEEE Trans. Audio Speech Lang. Process. **15**(3), 1075–1086 (2007)

22. S Deena, A Galata, Speech-driven facial animation using a shared gaussian process latent variable model. Adv. Visual Comput. Lecture Notes in Computer Science **5875**, 89–100 (2009)

23. T Cootes, G Edwards, C Taylor, Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)

24. DW Massaro, J Beskow, MM Cohen, CL Fry, T Rodriguez, in *Proceedings of International Conference on Auditory-Visual Speech Processing*. Picture my voice: audio to visual speech synthesis using artificial neural networks (Santa Cruz, CA, 1999), pp. 133–138

25. DW Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle* (MIT Press, Cambridge, 1998)

26. T Chen, Audiovisual speech processing. IEEE Signal Process. Mag. **18**, 9–21 (2001)

27. M Brand, in *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, Voice puppetry (Los Angeles, CA, USA, 1999), pp. 21–28

28. AJ Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans. Inf. Theor. **13**, 260–269 (1967)

29. L Xie, ZQ Liu, A coupled HMM approach to video-realistic speech animation. Pattern Recogn. **40**, 2325–2340 (2007)

30. S Moon, J Hwang, in *Proceedings of IEEE International Conf. Acoust., Speech, Signal Processing*, Noisy speech recognition using robust inversion of hidden Markov models (Seattle, WA, 1995), vol 1, pp. 145–148

31. KH Choi, JN Hwang, Constrained optimization for audio-to-visual conversion. Trans. Signal Process. **52**(6), 1783–1790 (2004)

32. R Carlson, B Granström, Data-driven multimodal synthesis. Speech Commun. **47**(1–2), 182–193 (2005)

33. S Ouni, MM Cohen, H Ishak, DW Massaro, in *EURASIP Journal on Audio, Speech and Music Processing*, Visual contribution to speech perception: measuring the intelligibility of animated talking heads (2007), pp. 1–12

34. LD Terissi, M Cerda, JC Gomez, N Hitschfeld-Kahler, B Girau, R Valenzuela, in *Proceeding of the IEEE International Conference on Multimedia and Expo (ICME)*, Animation of generic 3D head models driven by speech (Barcelona, Spain, 2011), pp. 1–6

35. L Rabiner, BH Juang, *Fundamentals of Speech Recognition* Signal Processing Series (Prentice Hall, New Jersey, 1993)

36. J Ahlberg, An updated parameterized face. Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden (2001)

37. LD Terissi, JC Gómez, 3D Head Pose and Facial Expression Tracking using a Single Camera. J. Univ. Comput. Sci. **16**(6), 903–920 (2010)

38. LE Baum, T Petrie, G Soules, N Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals Math. Stat. **41**, 164–171 (1970)

39. K Balci, in *Proceedings of the Seventh International Conference on Multimodal Interfaces*. XfacEd: authoring tool for embodied conversational agents (ICMI, 2005), pp. 208–213

40. S Kshirsagar, S Garchery, N Magnenat-Thalmann, in *Proceedings of the Workshop on Deformable Avatars*, Feature point based mesh deformation applied to MPEG-4 facial animation (Deventer, The Netherlands, 2001), pp. 24–34

41. M Sanchez Lorenzo, JD Edge, SA King, S Maddock, in *Proceedings of Vision, Video, and Graphics*, ed. by P Hall, P Willis. Use and re-use of facial motion capture data (Bath, UK, 2003), pp. 135–142

42. N Kojekine, V Savchenko, M Senin, I Hagiwara, in *In Short papers proceedings of Eurographics*, Real-time 3D deformations by means of compactly supported radial basis functions (Saarbruecken, Germany, 2002), pp. 35–43

43. K Singh, E Kokkevis, in *Graphics Interface*, Skinning characters using surface-oriented free-form deformations (2000), pp. 35–42

44. P Aleksic, A Katsaggelos, Speech-to-video synthesis using MPEG-4 compliant visual features. IEEE Trans. Circ. Systs. Video Technol. **14**(5), 682–692 (2004)

45. D Cosker, A Marshall, P Rosin, Y Hicks, Speech-driven facial animation using a hierarchical model. IEE Proc. Vision Image Signal Process. **151**(4), 314–321 (2004)

46. JD Gibbons, S Chakraborti, *Nonparametric Statistical Inference*, 4th edn. (Marcel Dekker, Inc., New York, 2003)

47. M Cohen, D Massaro, R Clark, in *Proceeding of the IEEE Fourth International Conference on Multimodal Interfaces*, Training a talking head (Pittsburgh, PA, 2002), pp. 499–510

48. L Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)

## Additional files

**Additional_file_1 as MPEG**
**Additional file 1: Animation example generated with the proposed system.**

- Filename: *movie1.mpg*
- Description: This video shows the animation of three different 3D head models driven by three different speech signals.

**Additional_file_2 as MPEG**
**Additional file 2: Animation example generated with the proposed system.**

- Filename: *movie2.mpg*
- Description: This video shows the animation of two avatars driven by the same speech signal.

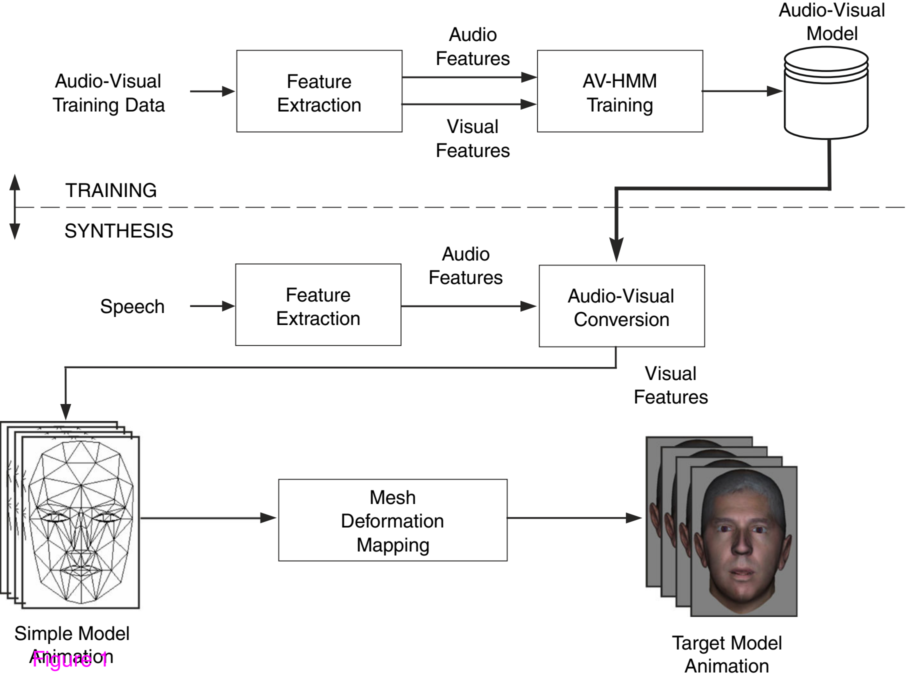

Audio-Visual
Training Data
Feature
Extraction
Audio
Features
Visual
Features
AV-HMM
Training
Audio-Visual
Model
TRAINING
SYNTHESIS
Speech
Feature
Extraction
Audio
Features
Audio-Visual
Conversion
Visual
Features
Mesh
Deformation
Mapping
Simple Model
Animation
Target Model
Animation

Figure 1

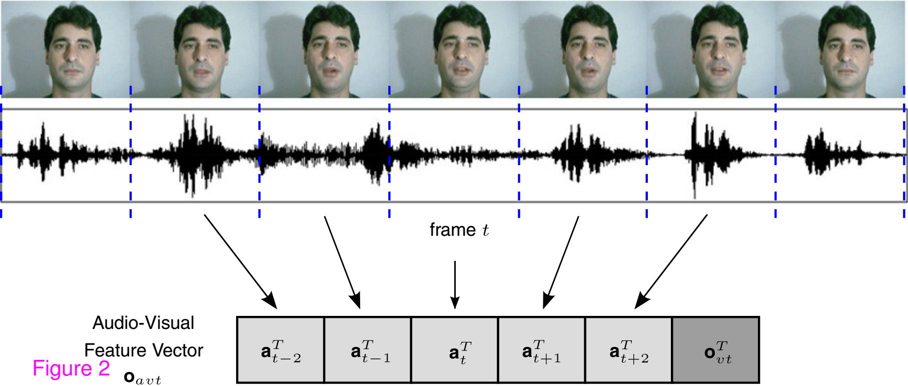

frame $t$
Audio-Visual
Feature Vector
$\mathbf{o}_{avt}$
$\mathbf{a}_{t-2}^{T}$
$\mathbf{a}_{t-1}^{T}$
$\mathbf{a}_{t}^{T}$
$\mathbf{a}_{t+1}^{T}$
$\mathbf{a}_{t+2}^{T}$
$\mathbf{o}_{vt}^{T}$

Figure 2

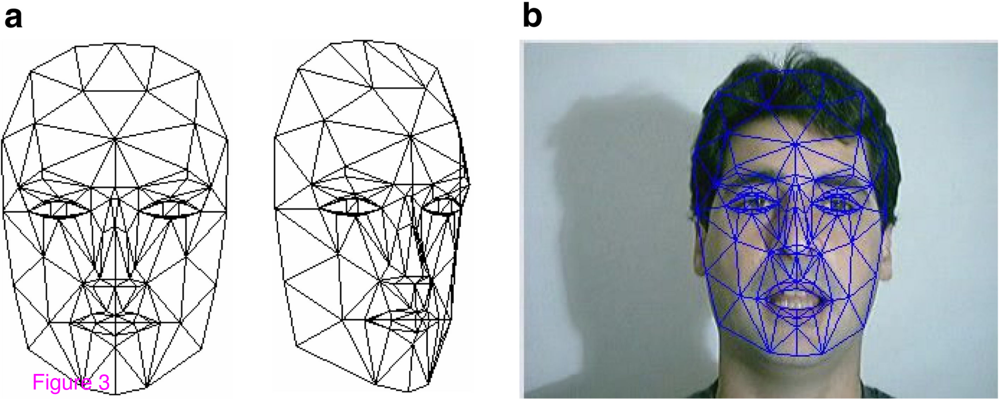


Figure 3

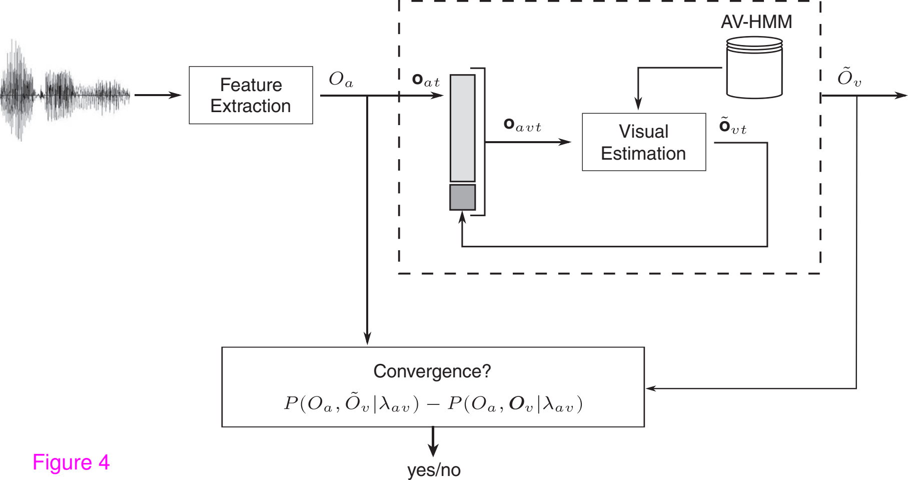

Feature Extraction
$O_a$
$\mathbf{o}_{at}$
AV-HMM
$\tilde{O}_v$
$\mathbf{o}_{avt}$
Visual Estimation
$\tilde{\mathbf{o}}_{vt}$
Convergence?
$P(O_a, \tilde{O}_v | \lambda_{av}) - P(O_a, \boldsymbol{O}_v | \lambda_{av})$
yes/no

Figure 4

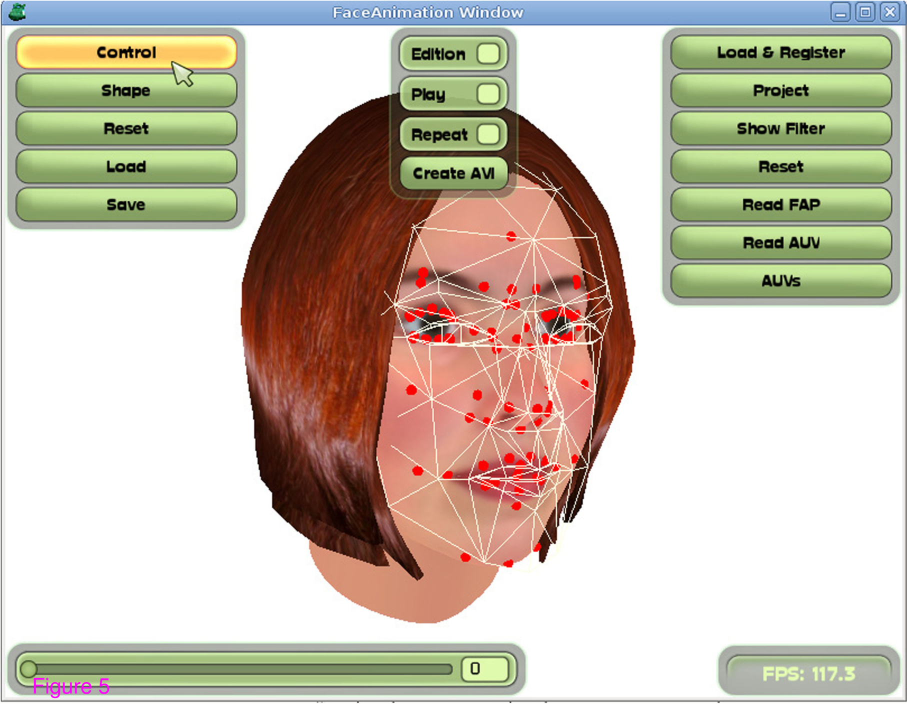
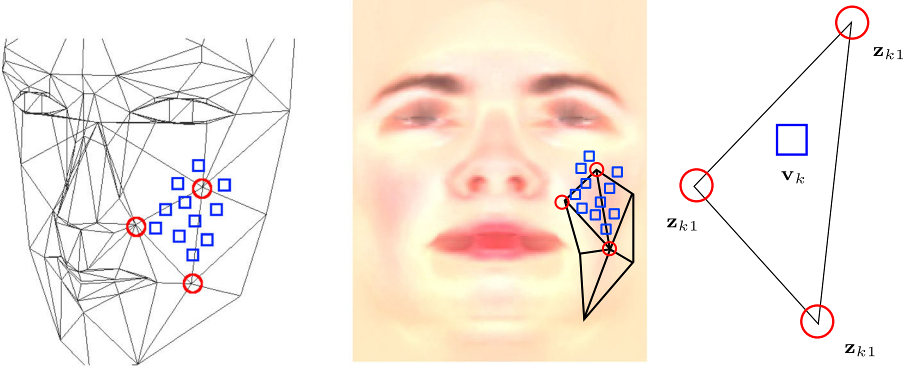

FaceAnimation Window
Control
Shape
Reset
Load
Save
Edition
Play
Repeat
Create AVI
Load & Register
Project
Show Filter
Reset
Read FAP
Read AUV
AUVs
0
FPS: 117.3

Figure 5

Figure 6



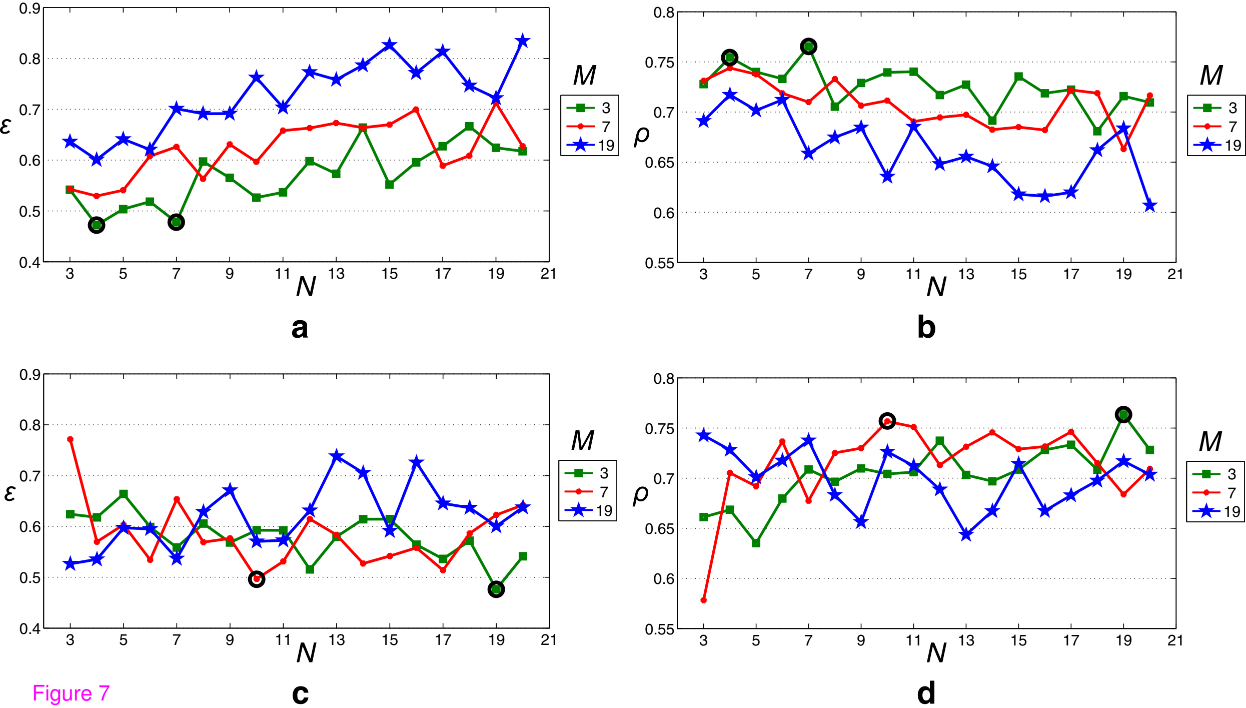

ε
0.4
0.5
0.6
0.7
0.8
0.9
3
5
7
9
11
13
15
17
19
21
N
M
3
7
19
a
ρ
0.55
0.6
0.65
0.7
0.75
0.8
3
5
7
9
11
13
15
17
19
21
N
M
3
7
19
b
ε
0.4
0.5
0.6
0.7
0.8
0.9
3
5
7
9
11
13
15
17
19
21
N
M
3
7
19
c
ρ
0.55
0.6
0.65
0.7
0.75
0.8
3
5
7
9
11
13
15
17
19
21
N
M
3
7
19
d

Figure 7

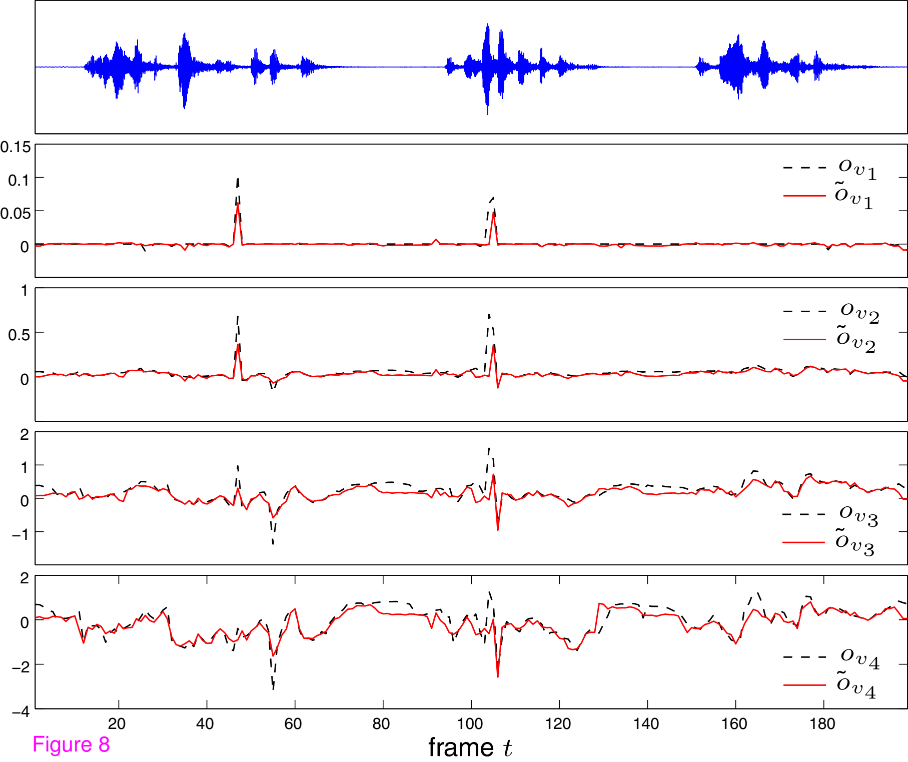

0.15
0.1
0.05
0
$o_{v_1}$
$\tilde{o}_{v_1}$
1
0.5
0
$o_{v_2}$
$\tilde{o}_{v_2}$
2
1
0
−1
$o_{v_3}$
$\tilde{o}_{v_3}$
2
0
−2
−4
$o_{v_4}$
$\tilde{o}_{v_4}$
20
40
60
80
100
120
140
160
180
frame $t$

Figure 8

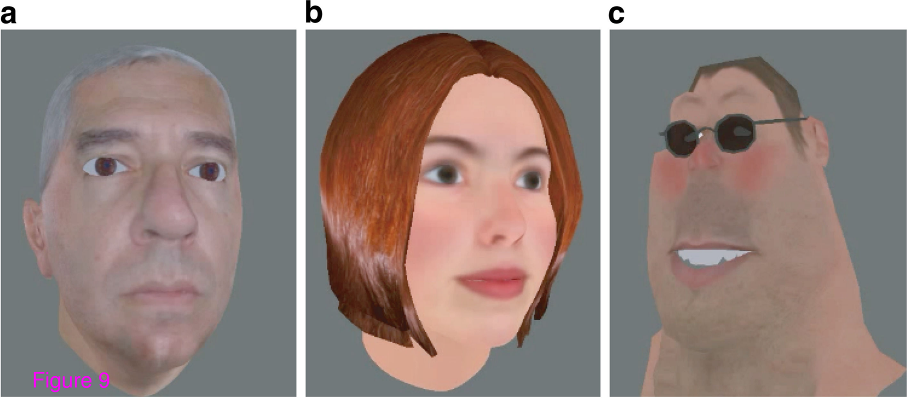


Figure 9

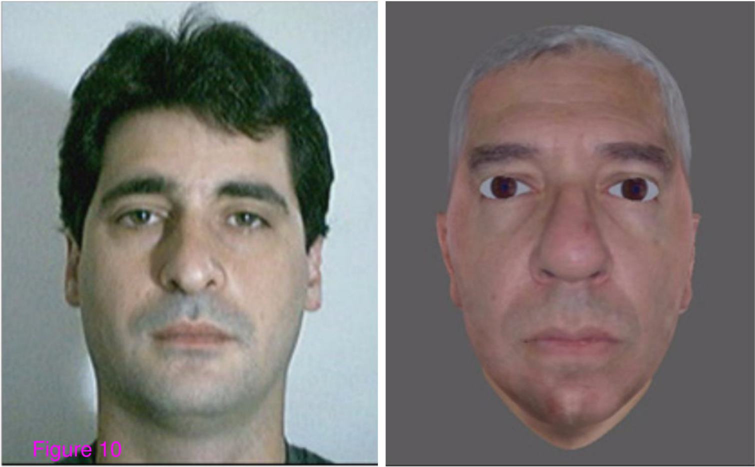


Figure 10

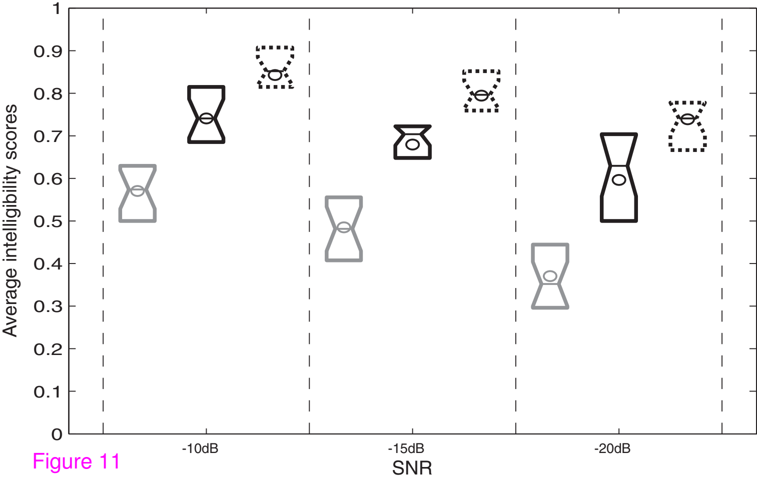

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Average intelligibility scores
-10dB
-15dB
-20dB
SNR

Figure 11

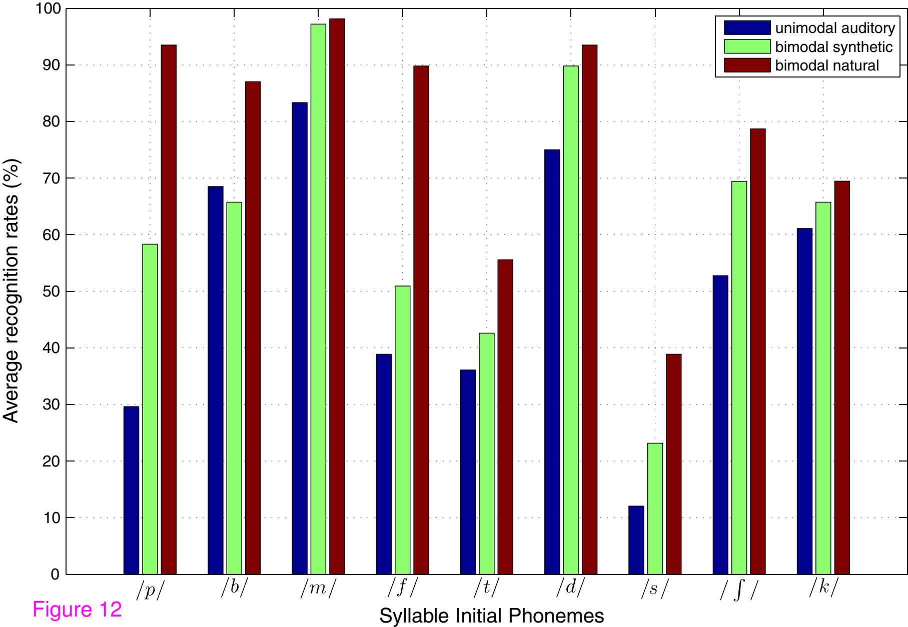

100
90
80
70
60
50
40
30
20
10
0
Average recognition rates (%)
unimodal auditory
bimodal synthetic
bimodal natural
$/p/$
$/b/$
$/m/$
$/f/$
$/t/$
$/d/$
$/s/$
$/\int/$
$/k/$
Syllable Initial Phonemes

Figure 12

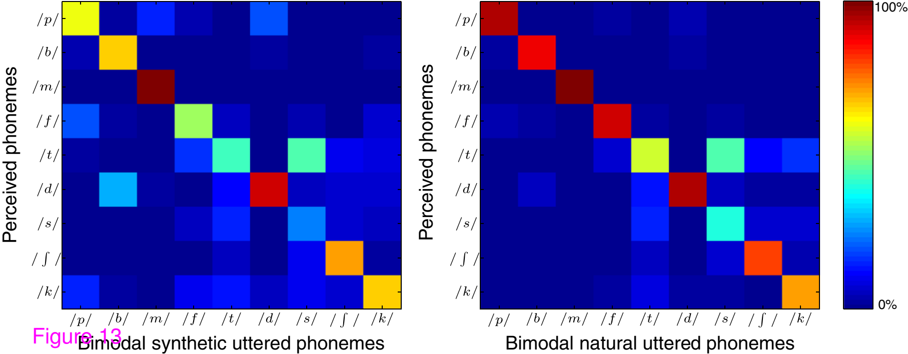

/p/
/b/
/m/
/f/
/t/
/d/
/s/
/ ʃ /
/k/
Perceived phonemes
Bimodal synthetic uttered phonemes
Bimodal natural uttered phonemes
100%
0%

Figure 13

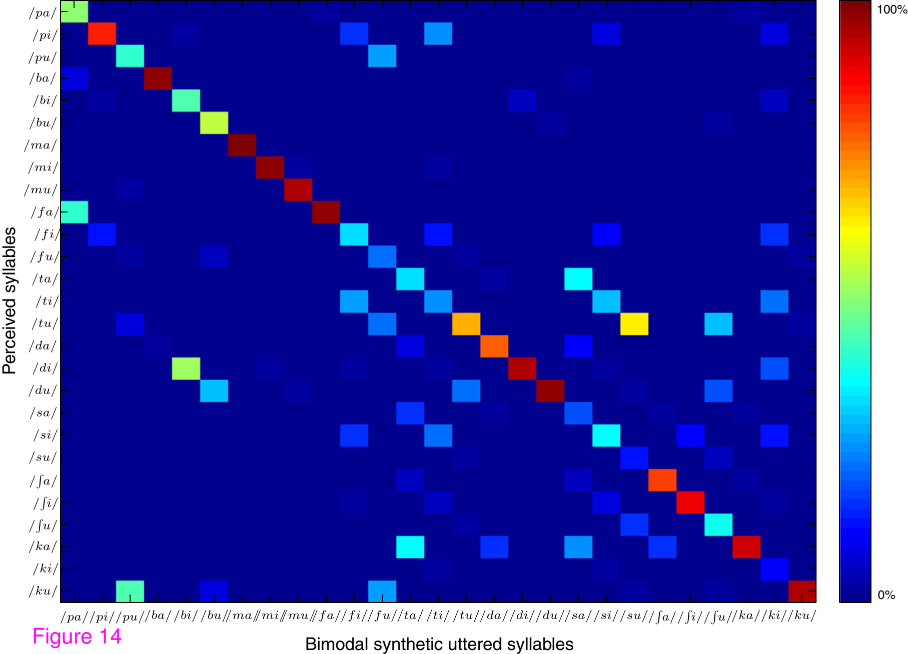

/pa/
/pi/
/pu/
/ba/
/bi/
/bu/
/ma/
/mi/
/mu/
/fa/
/fi/
/fu/
/ta/
/ti/
/tu/
/da/
/di/
/du/
/sa/
/si/
/su/
/ʃa/
/ʃi/
/ʃu/
/ka/
/ki/
/ku/
Perceived syllables
/pa//pi//pu//ba//bi//bu//ma//mi//mu//fa//fi//fu//ta//ti//tu//da//di//du//sa//si//su//ʃa//ʃi//ʃu//ka//ki//ku/
Bimodal synthetic uttered syllables
100%
0%

Figure 14

**Additional files provided with this submission:**

Additional file 1: 1574909649774090_add1.mpeg, 7392K
http://asmp.eurasipjournals.com/imedia/2142917495908969/supp1.mpeg
Additional file 2: 1574909649774090_add2.mpeg, 348K
http://asmp.eurasipjournals.com/imedia/4922732371011276/supp2.mpeg