Emphatic Visual Speech Synthesis

Javier Melenchón, Elisa Martínez, Fernando De La Torre, and José A. Montero

***Abstract*—The synthesis of talking heads has been a flourishing research area over the last few years. Since human beings have an uncanny ability to read people's faces, most related applications (e.g., advertising, video-teleconferencing) require absolutely realistic photometric and behavioral synthesis of faces. This paper proposes a person-specific facial synthesis framework that allows high realism and includes a novel way to control visual emphasis (e.g., level of exaggeration of visible articulatory movements of the vocal tract). There are three main contributions: a geodesic interpolation with visual unit selection, a parameterization of visual emphasis, and the design of minimum size corpora. Perceptual tests with human subjects reveal high realism properties, achieving similar perceptual scores as real samples. Furthermore, the visual emphasis level and two communication styles show a statistical interaction relationship.**

***Index Terms*—Audiovisual speech synthesis, emphatic visual-speech, talking head.**

I. Introduction

ACCESS to the increasing availability of digital information stored in computers [1] must be carried out through output transducers (e.g., loudspeakers and screens). In order to obtain this information in an effortless way, the interaction process should be as simple and natural as possible. The human face is our most common interface. Starting from a very young age [2], the face is used in most of our social interactions [3]. In fact, there is a specific region of the brain dedicated to recognize and analyze faces [4]. It has been shown that we perceive faces as bimodal signals [5], [6] of audiovisual information; bimodality helps the global understanding of a spoken message [7], specially with low signal-to-noise ratios. Previous work has shown that access to digital information would be easier if human faces could be used in the process, e.g., synthesizing talking faces ([8], [9]). There are a variety of applications that use virtual faces such as virtual avatars, low bit rate videoconference, visual telephony for the hard of hearing, and tools for speech therapists. In this last case, more exaggerated movements of the lips would be desirable to increase their teaching capabilities. Here, the control of visual emphasis in the synthesis could be helpful.

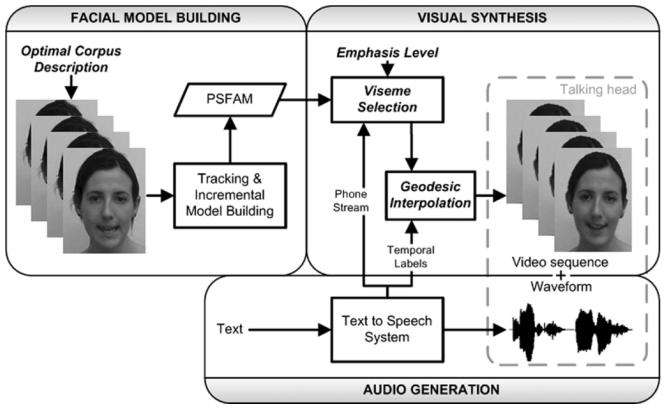

Fig. 1. System diagram has three blocks: facial model building, visual synthesis, and audio generation. Bold italic font identifies specific contributions. Facial model building consists of optimal corpora description and a simultaneous tracking with incremental construction of the person-specific facial appearance model. Visual synthesis comprises a geodesic interpolation procedure and a viseme selection algorithm, allowing visual emphasis control. Finally, the audio generation block uses a text-to-speech system to drive the visual synthesis and provides the speech waveform of the talking head.

Parke *et al.* [10] introduced the first talking head using rough 3-D faces. Since then, talking heads have evolved into more realistic faces with high expressive power and personalization capabilities [11], [12]. However, no solution has yet been proposed to provide both expressiveness and personalization features together. This paper aims to fulfil this gap by synthesizing realistic faces with different levels of visual emphasis and reducing personalization efforts. The realism includes visual emphasis; in particular, it is the level of exaggeration of visible articulatory movements of the vocal tract, ranging from hypo-articulated behaviors to hyper-articulated ones. These effects are shown when speaking and their proper emulation can provide additional realism to the synthesis process.

A. Contributions and Similar Research Works

The proposed audiovisual speech synthesis scheme is a sample based and text-driven approach. Fig. 1 illustrates the main contributions of the paper: realism, emphatic synthesis, and personalization:

- A geodesic viseme interpolation algorithm with visual unit selection allows the achieving of the high realism synthesis without any manual intervention. The proposed interpolation is a nonlinear scheme which provides good photorealism. The visual unit selection is inspired by the work of [13] for speech synthesis to obtain credible videorealism. The proposed viseme interpolation can be seen as a simplified version of that of Cohen and Massaro [14],

Manuscript received January 15, 2008; revised October 07, 2008. Current version published February 11, 2009. This work has was supported in part by the Spanish Ministry of Education and Science under Grant TEC2006-08043/TCM. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Helen Meng.

J. Melenchón is with the Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya, 08018 Barcelona, Spain (e-mail: jmelenchonm@uoc.edu).

E. Martínez and J.A. Montero are with the GPMM—Grup de Recerca en Processament Multimodal, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, 08022 Barcelona, Spain (e-mail: elisa@salle.url.edu; montero@salle.url.edu).

F. De La Torre is with the Robotics Institute, Carnegie Mellon University, Pittsburgh PA, 15213 (e-mail: ftorre@cs.cmu.es).

Digital Object Identifier 10.1109/TASL.2008.2010213

using image representation instead of parameter articulators, linear interpolation instead of decaying exponentials and taking into account only neighborhood visemes instead of broader sets.
- A parameterization of visual emphasis, based on the visual unit selection and order statistics.
- The design of minimum size corpora to reduce the length of the personalization procedure. This design 1) provides the short uniform distributed corpus needed by the visual unit selection, 2) finds meaningless word sets with optimal visual coverage and minimum size, and 3) uses parallel genetic algorithms (PGA) [15] to cope with its intrinsic nonlinear specifications. Optimal visual coverage is understood as uniform distributions of visual appearances and their transitions. The facial model uses person-specific facial appearance models (PSFAM) [16] and it is efficiently learned automatically with the incremental computation of eigenspaces [17].

Perceptual tests have been carried out with human subjects, revealing nonsignificant $(p > 0.15)$ photorealism differences between real and synthetic talking heads. Moreover, a significant statistical interaction $(p < 0.01)$ has been detected between the visual emphasis level and the two different communication styles (speaking and singing).

This work is closely related to the Ph.D. thesis of Cosatto [18] and Ezzat *et al.* [19]. However, there are several differences. In [18], visual selection was used for creating transitions between consecutive visemes, while in this work, transitions are created with the geodesic interpolation, guided by some target visemes given by the visual unit selection. Furthermore, this paper uses a specific corpora designed under visual constraints. Instead of tracking individual facial features like in [18], our tracking and synthesis mechanism uses a holistic representation of the face to provide better inter-feature consistency in the synthesis. Principal component analysis (PCA) is used in this work and in [18] to reduce data dimensionality; however, lip parametrization is used to guide the animation in Cosatto's work, while PCA encoding is used here instead, taking into account a richer description of the visual appearance of the face (e.g., tongue, teeth) in the animation procedure. With respect to the work of Ezzat *et al.* [19], this paper broadens it with the addition of emphasis, the alignment process and viseme grouping concepts. To build the PSFAM automatically, rigid transformations are used instead of optical flow, incremental singular value decomposition is taken into account instead of expectation maximization PCA, synthesis is achieved with a discrete model, there is no need of warping procedures and, finally, Castilian Spanish is used instead of English.

The paper is organized as follows. Section I introduces the paper, its novelties in Section I-A and related work about talking heads and visual emphasis in Sections I-B and C, respectively; Section II describes the visual synthesis scheme, providing some phonetic concepts in Section II-A, describing the PSFAM in Section II-B, explaining the geodesic viseme interpolation in Section II-C, the visual unit selection in Section II-D, and the visual speech emphasis in Section II-E; Section III describes the PSFAM training, specifying how to obtain a new PSFAM in Section III-A and the proposed visual corpora description algorithm in Section III-B; experimental results are given in Section IV, including perceptual evaluation in Section IV-B and the performance of the corpora description algorithm in Section IV-A; finally, concluding remarks with future work guidelines are stated in Section V.

B. Talking Heads

Since the pioneering work in facial animation [10] and audiovisual speech perception [20], much effort in research has made considerable progress in the strive for audiovisual speech synthesis. Researchers have considered various approaches to translate phonetic or auditory information to visual speech. These attempts have mostly dealt with two key aspects: the relationship between input data (e.g., text, audio), visual features, and the facial animation model.

The relationship between input data and visual features can be text-driven like in [11], [18], [19], and [21] or speech-driven like in [22], [23], whether the visual appearance is built from symbolic phonetic input or from auditory information directly. Speech-driven approaches offer poorer performance than text-driven ones but do not require the phonetic transcription of the message needed by text-driven approaches, since they can obtain synthetic visual output from speech waveforms.

The facial animation models can be classified into 3-D facial models [10], [11], [22], [23] or sample-based ones [18], [19], [21], [24]. The former usually render 3-D polygonal deformable meshes, while the latter directly specify the radiance level of each pixel in the image, without 3-D structures. A more detailed and accurate appearance is offered by 3-D models; however, they have reduced motion control [9]. Static realism or photorealism is best achieved with sample based models. For dynamic realism or videorealism, both approaches excel in different kinds of movements: 3-D models for rigid motions and sample based models for nonrigid ones. Despite the large amount of research in talking heads, little work has addressed the hyper or hypo articulated mouth movements [25], [26] (they are reviewed in Section I-C).

A recorded corpora is required when constructing sample based models [9]. Recorded images must be processed in order to ease their recovery when synthesizing output video sequences and typical corpus recording times are about 10–20 min. They consist of a set of short words and/or short meaningful sentences [18], [19], [21], and usually require expert performance. The word and sentence selection criteria are barely detailed in the literature and most are only based on auditory constraints, e.g., the proposal of Cosatto [18] following a corpus design for audio speech synthesis [27]. Instead, global procedures that include visual features could be used. In this paper, we use the optimal corpora design baselines of Black [28] for the unit selection speech synthesis framework [13] and transfer it to the proposed synthesis model, obtaining optimal visual coverage corpora with minimum size.

C. Visual Emphasis

We refer to visual emphasis as the level of exaggeration of visible articulatory movements of the vocal tract. Lasseter [29] defines exaggeration as "accentuating the essence of an idea via the design and the action." In the image synthesis field, design

refers the static appearance, while action relates to movement. In this paper, the term exaggeration does not negate reality, following Redman's definition [30].

When dealing with talking faces, visual emphasis has been used in two main fields. From the design point of view, visual emphasis has been used in the caricaturization of facial images (e.g., Rautek *et al.* [31]). From the action point of view, visual emphasis appears in hyper and hypoarticulation visual effects. They relate facial appearances with stress or emotions when speaking and increase or decrease the intelligibility of the spoken message [32].

The relationship between facial appearance and stress in speech was first noticed by Ekman [33], when showing a correlation between the eyebrow movements of a speaking person and the stress shown through her/his speech. This idea has been later enforced [34]–[37]. Specifically, [34] related eyebrow movement to pitch variations, [35] concluded that stress was best perceived by the visual channel, [36] studied particular articulators involved with stress, and [37] investigated different facial regions. Moreover, facial movements have an influence on the perceived message and its related emotion. The work of [38] shows that some speech styles can vary the perceived realism of the message, and [25] found that emotion is better interpreted with exaggerated movement expressions than with slowed down facial movements. Following a similar line, Beskow *et al.* [26] observed that articulators experienced accentuated movements for focal pronunciations. Therefore, it is interesting to include visual emphasis in talking heads synthesis. Very few attempts have been made along this line. Aylett [39] presented a method for synthetic hyperarticulated auditory speech and Hill *et al.* [25] used a synthetic nonrealistic talking head in their experiments.

II. Visual Synthesis

The visual synthesis framework is described here. Section II-A states some preliminary phonetic concepts. Section II-B describes the use of person-specific facial appearance models (PSFAM) [16]. Given an input text, a PSFAM and an emphasis level (α between 0 and 1), the visual synthesis algorithm generates a synthetic video (see Fig. 1) with the help of a text-to-speech engine (TTS) [40]. The TTS provides three elements: a set of temporal labels, which drives the geodesic interpolation of Section II-C to find realistic image sequences for each transition regardless of its duration; a phone stream, taken as input by the visual unit selection of Section II-D, to find the set of transitions; and a waveform, which is attached to the synthesized video to form an audiovisual signal. The emphasis level and the visual unit selection allow the parameterization of visual emphasis, stated in Section II-E.

A. Phonetic Concepts

This section describes basic phonetic terminology used in this paper such as *phoneme*, *phoneme class*, *viseme*, *viseme class*, *sound unit*, and *visual unit* (Fig. 2). Phonemes are the minimal meaning units [41] and each one can be expressed acoustically with a variety of different sound units. Visemes are the minimal indistinguishable sets of visual units [42]. Ideally, there should be a one-to-one relationship between phonemes and visemes, since every sound needs a particular and distinct configuration of the vocal tract. However, indistinguishableness effects appear when noise and inherent occlusions are added in our perception process [43], so a many-to-many relationship is actually used instead. Two or more similar (in a sense of perceptually indistinguishable) phonemes can belong to the same phoneme class, and two or more similar visemes can be represented by the same viseme class. Phoneme classes are mainly defined by manner and voicing speech features, while viseme classes are determined by the place of articulation, being more similar for internal articulation ones [43].

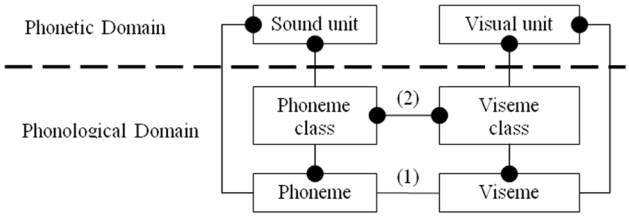

Fig. 2. Relationship among phonemes, phoneme classes, visemes, viseme classes, sound units, and visual units. Cardinality of many is represented with a circle termination and single cardinality without it. The diagram is divided into phonological and phonetic domains. Audiovisual relationship is given by (2); however, it could be represented ideally by (1), if occlusion and noise were not present.

It is worth mentioning that although phonemes are strictly an abstract entity, they generally appear in the literature instead of the more specific term *allophone*, or similar sounds tied to the same phoneme. From now on in this paper, the word phoneme will be used instead of allophone, to ease the task of the reader to follow the terminology of existing literature.

B. Person-Specific Facial Appearance Model

A PSFAM is a parameterized appearance model that has been previously used to represent the face [16], [44], based on modular eigenspaces (ME) [45]. This model has been selected because of its modular, generic and compact data representation as well as its full compatibility with the proposed geodesic interpolation, visual selection, and emphatic control algorithms of Sections II-C–E, respectively.

The PSFAM is built from a training set of N images (see notation[1]) $\mathbf{I} \in \Re^{P\times M}$ of P rows and M columns and consists of two elements. The first one assumes the face segmented into R facial regions, representing the location information with R binary images $\boldsymbol{\pi}^r \in \Re^{P\times M}$, each one with Q_r pixels of value 1, being Q_r the number of pixels of the facial region of $\boldsymbol{\pi}^r$; moreover, $\sum_{r=1}^{R} \boldsymbol{\pi}^r = \bar{\mathbf{1}} \in \Re^{P\times M}$ is another binary image defining the whole face region. The second element contains R low-dimensional subspaces $\mathbf{B}^r \in \Re^{Q_r\times L}$, one for each facial region, representing the appearance information with rank L. Any texture data vector $\mathbf{t}_n^r \in \Re^{Q_r\times 1}$ corresponding to the facial appearance of region r in frame n can be obtained as a linear combination of the columns of $\mathbf{B}^r$, with weights in vector $\mathbf{w}_n^r \in \Re^{L\times 1}$. Texture vectors $\mathbf{t}_n^r$ are composed with binary images $\boldsymbol{\pi}^r$

[1]Bold capital letters denote a matrix $\mathbf{A}$, bold lowercase letters a column vector $\mathbf{a}$. $\mathbf{a}_i$ represents the ith column of matrix $\mathbf{A}$. All non-bold letters represent scalar variables or functions if they are followed by $()$. Matrices, vectors, and scalars related to the jth facial region are denoted by superscript $^{\cdot j}$. $\mathbf{A} \circ \mathbf{B}$ denotes the Hadamard (point wise) product between two matrices of equal dimension.

through function v to obtain images $v(\mathbf{t}_n^r, \boldsymbol{\pi}^r) = \mathbf{F}_n^r \in \Re^{P\times M}$, where $\mathbf{F}_n^r = \mathbf{I}_n \circ \boldsymbol{\pi}^r$, i.e., images with original pixel values of $\mathbf{I}_n$ in nonzero pixels of $\boldsymbol{\pi}^r$. Then

$$\sum_{r=1}^{R} v(\mathbf{B}^r \mathbf{w}_n^r, \boldsymbol{\pi}^r) = \sum_{r=1}^{R} \mathbf{I}_n \circ \boldsymbol{\pi}^r = \mathbf{I}_n \circ \boldsymbol{\Pi} = \mathbf{F}_n \quad (1)$$

where $\mathbf{F}_n \in \Re^{P\times M}$ is an image with the whole face of $\mathbf{I}_n$ displayed over a black background (assuming 0 as black). The face remains in the same position in both images.

In this piece of work, PSFAM is applied with two differences: the first element of vectors $\mathbf{w}_n^r$ becomes 1 and the first column of $\mathbf{B}^r$ becomes the mean appearance $\overline{\mathbf{m}}^r \in \Re^{Q_r\times 1}$ of the facial region r through all images, so

$$\mathbf{B}^r \mathbf{w}_n^r = \mathbf{U}^r \mathbf{c}_n^r + \overline{\mathbf{m}}^r \quad (2)$$

where $\mathbf{U}^r \in \Re^{Q_r\times K}$, and $\mathbf{c}_n^r \in \Re^{K\times 1}$ for $K = L - 1$. In this paper, $\mathbf{U}^r$ contains the eigenspace of the appearance of facial region r, centered around the mean information $\overline{\mathbf{m}}^r$. This eigenspace is obtained from the texture information matrix $\mathbf{T}^r \in \Re^{Q_r\times N}$, with the texture vectors $\mathbf{t}_n^r$ in its columns. The singular value decomposition (SVD) [46] is applied to $\mathbf{T}^r$

$$\mathbf{T}^r = \mathbf{U}^r \boldsymbol{\Sigma}^r (\mathbf{V}^r)^T + \overline{\mathbf{m}}^r \cdot \mathbf{1}^T \quad (3)$$

where the columns of $\boldsymbol{\Sigma}^r(\mathbf{V}^r)^T \in \Re^{K\times N}$ contain the projected texture vectors $\mathbf{c}_n^r \in \Re^{K\times 1}$ and $\mathbf{1} \in \Re^{N\times 1}$ is a row vector of N ones. Although $L = K - 1$ is defined as the rank of matrix $\mathbf{T}^r$, lower values can be considered, e.g., 20, with no loss of perceptual quality. Textures were extracted from 320×240 facial images.

Finally, in order to reduce artifacts of $\mathbf{F}_n$ in its facial region transitions (see Fig. 3), blending effects are introduced in pixels near two or more facial regions. A smooth transition is included in the edges of images $\boldsymbol{\pi}^r$, making them no longer binary (nevertheless, their sum $\boldsymbol{\Pi}$ still remains binary).

C. Geodesic Interpolation

Let $\mathbf{S}^r = \{\mathbf{s}_1^r \ldots \mathbf{s}_M^r\}$ be an ordered set of M textures of region r corresponding to a transition between two projected texture vectors $\mathbf{c}_i^r = (\mathbf{U}^r)^T(\mathbf{s}_1^r - \overline{\mathbf{m}^r})$ and $\mathbf{c}_f^r = (\mathbf{U}^r)^T(\mathbf{s}_M^r - \overline{\mathbf{m}^r})$. Remember that each texture $\mathbf{s}_m^r$ is related to a facial region appearance $\mathbf{F}_m$ through function v. Transition $\mathbf{S}$ is considered videorealistic in this work if consecutive pairs of facial regions $\mathbf{F}_n$ and $\mathbf{F}_{n+1}$ are similar: the more similar they are, the more videorealistic the transition will be. Moreover, each facial region $\mathbf{F}^m$ must be photorealistic, i.e., it must be within its appearance subspace, approximated by matrix $\mathbf{U}^r$; however, $\mathbf{U}^r$ is a generic description of the appearance subspace of region r and PSFAM assumes Gaussian data distributions. The real subspace is, in fact, unknown. Nevertheless, a nonuniform sampled version of this subspace is found in the real projected texture vectors $\boldsymbol{\Sigma}^r(\mathbf{V}^r)^T$ (3) of the training sequence. These vectors can be used to find approximations of the desired transition. The idea is to find a path from $\mathbf{c}_i^r$ to $\mathbf{c}_f^r$ which lies near them, keeping within the subspace. In fact, the nearer the trajectory , the more probable it will be inside the real subspace. One way to follow this geodesic constraint is to pass through the columns of $\boldsymbol{\Sigma}^r(\mathbf{V}^r)^T$ (the real projected texture vectors) with small jumps. The shortest path algorithm of Dijkstra [47] and powers of the Euclidean distance have proven to be useful in this case. To use them, graph notation is required.

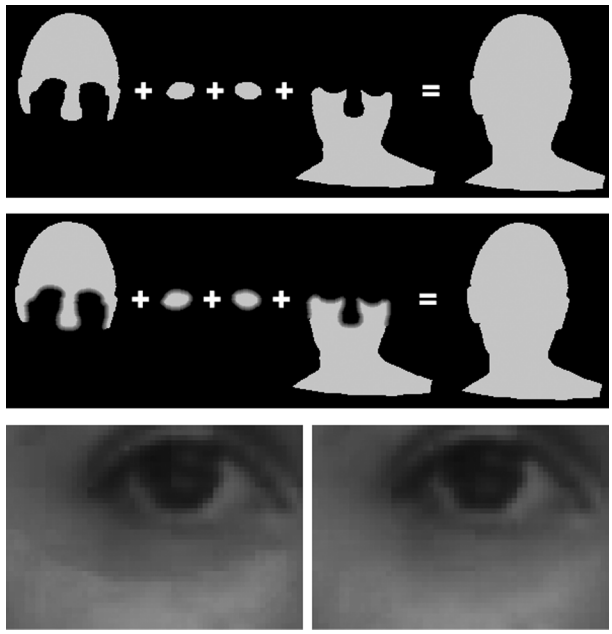

Fig. 3. Smooth region merging. From top to bottom and left to right: unweighted mask set; weighted mask set; visual artifacts under the eye due to illumination changes; reduced visual artifacts thanks to smoothing transitions between facial regions using the weighted mask set.

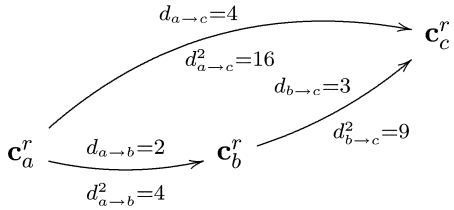

Fig. 4. Distance between real projected texture vectors. Euclidean distance is represented by d. It can be checked that $d_{a\to b} + d_{b\to c} = 5 > 4 = d_{a\to c}$ and $d^2_{a\to b} + d^2_{b\to c} = 13 < 16 = d^2_{a\to c}$. In the former, making one large jump results in less distance than with two smaller ones. In the latter, an opposed result is obtained due to the use of the power of 2.

Let $\mathbf{G}^r$ be a fully connected graph with the real projected texture vectors as nodes. Let also $\|\mathbf{c}_k^r - \mathbf{c}_l^r\|^\beta$ be the distance between two random nodes, with $\|\cdot\|$ as the Euclidean norm.

$$\mathbf{G}^r(k,l) = \|\mathbf{c}_k^r - \mathbf{c}_l^r\|^\beta. \quad (4)$$

When $\beta = 1$, the shortest path between nodes is a straight line. When $\beta > 1$, the shortest path is likely to pass through additional texture vectors (Fig. 4) because large jumps are penalized by the power β. In fact, higher penalizations are obtained as β grows. In this paper, a value of two for β is chosen, to reduce the computational cost. To find the shortest paths in graph $\mathbf{G}^r$, the Dijkstra algorithm [47] is used.

When the shortest path is obtained, the geodesic distance between real projected texture vectors $\mathbf{c}_i^r$ and $\mathbf{c}_f^r$ can be determined. To obtain the desired images of the transition $\mathbf{S}^r$, the

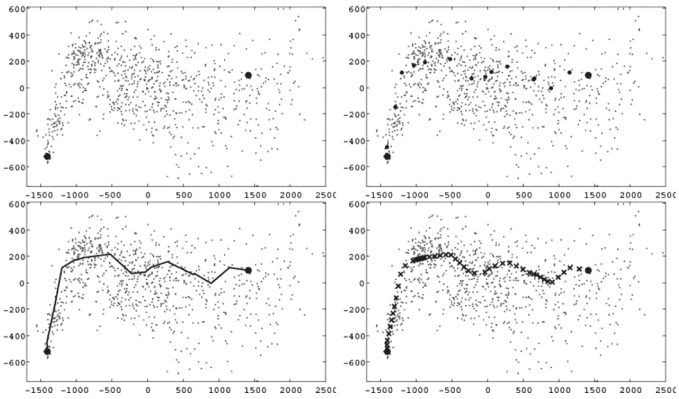

Fig. 5. Geodesic interpolation. Subspace samples are noted by red dots, while blue objects are related to interpolated information. Top: initial and final points to be interpolated are marked with large blue dots in both images; the shortest path between them is obtained when passing through the little blue dots of the right figure. Bottom: the trajectory corresponding to the shortest path is shown in the left figure, while its sampled version is drawn in the right one. The squared Euclidean distance is taken as the distance metric in this example.

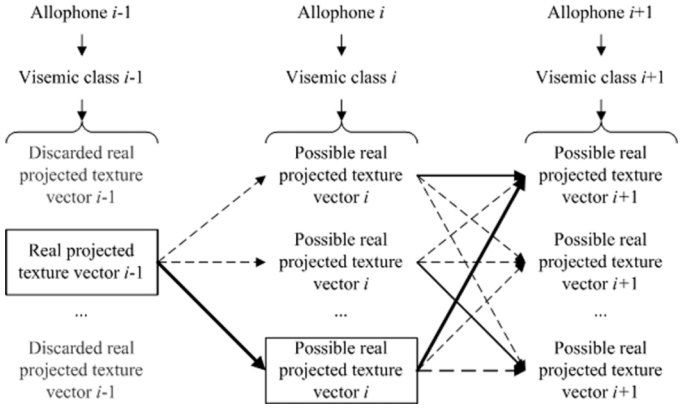

Fig. 6. Visual unit selection process of real projected texture vector i, given a sequence of three phonemes. From top to bottom: the three phonemes, the three corresponding viseme classes, and their related real visual units. The real visual unit $i-1$ has been selected by a previous iteration, discarding the other possibilities. The nearest real visual unit $i+1$ to each real visual unit i is marked with a solid line and the shortest path between real visual unit $i-1$ and $i+1$ is shown with a thick solid one. Real visual units are therefore known for $i-1$ and i, but real visual unit $i+1$ will be selected next.

shortest path must be sampled into M points $\hat{\mathbf{c}}_k^r$. Each sample point $\hat{\mathbf{c}}_k^r$ can be obtained as the linear interpolation of the closest two real projected texture vectors. Therefore, $\hat{\mathbf{c}}_k^r$ may not correspond to any observed texture vectors, becoming some kind of virtual projected texture vectors. In fact, this is very likely to happen, with the exception of $\mathbf{c}_i^r$ and $\mathbf{c}_f^r$, respectively (Fig. 5).

It must be noted that since no temporal information is kept from the training data, the shortest path may not produce the right timing of the transition. This effect has been approximated with the assumption that the duration between sample points is proportional to the distance between them.

To further improve the photorealism of synthetic images, a Gaussian blur over the edge pixels of $\mathbf{\Pi}$ (between the face region and the background) is applied in each synthetic image $\mathbf{F}_k$ obtained from $\hat{\mathbf{c}}_k^r$ for every region to simulate the point spread function effects over their boundary pixels [48].

D. Visual Selection Algorithm

A new visual selection algorithm is proposed to improve the videorealism performance of this framework and to model visual speech emphasis (as detailed in Section II-E). While the previous Section II-C stated the specific transitions between two real projected texture vectors, this one explains how to select the appropriate set of real projected texture vectors among those of the training set. To do so, higher level information symbols are needed, e.g., visemes for the mouth region. The idea is to have several real projected texture vector candidates for each high-level information symbol, so the more suited vector can be selected for every symbol when a symbol stream is specified (Fig. 6). The higher the variety of real projected texture vectors, the better the performance of the selection process. In order to have maximum visual variance with minimum recording time, an optimal corpora description algorithm is proposed in Section III-B.

When the selection process is used to synthesize the mouth region, a mapping between viseme classes to real projected texture vectors (which is one to many, see Fig. 2) is used. If the symbol stream consists of phonemes, an extra mapping between phonemes and viseme classes (which is many to one, as can be deduced from Fig. 2) is needed beforehand. To select the best real projected texture vector among the candidates for a given symbol, the unit selection technique used in concatenative speech synthesis [13] can be adapted here, using only the concatenative cost. Assuming the exponentially decaying influence of phonemes in coarticulation effects [14], only the next and previous phonemes for a given one are considered; given the current phoneme a_i, the next one a_{i+1} and the previously selected real projected texture vector $\mathbf{c}_{i-1}$, an error concatenation function (E_c) can be defined to find the current one $\mathbf{c}_i$

$$E_c(\mathbf{c}_i, \mathbf{c}_{i+1}) = D(\mathbf{c}_{i-1}, \mathbf{c}_i) + D(\mathbf{c}_i, \mathbf{c}_{i+1}) \tag{5}$$

where $D(a,b)$ represents the geodesic distance between real projected texture vectors a and b, found with the geodesic interpolation method of Section II-C; $\mathbf{c}_{i+1}$ identifies a real projected texture vector related to viseme class v_{i+1}, which is related to the phoneme a_{i+1}; finally, $\mathbf{c}_i$ is the real projected texture vector related to the current phoneme a_i through the viseme class v_i. Note that $\mathbf{c}_i$ and $\mathbf{c}_{i+1}$ can take different values while $\mathbf{c}_{i-1}$ is fixed. Taking the minimum value of E_c for each $\mathbf{c}_i$, a vector of concatenation costs $\mathcal{E} = \min_{\mathbf{c}_{i+1}} E_c$ is obtained. The index of its first-order statistic can be used to determine the value of $\mathbf{c}_i$ which minimizes both $\mathcal{E}$ and E_c. A viterbi algorithm [49] with a trace back depth of two can implement this approach (Fig. 6).

E. Emphatic Control

Examining the previous vector $\mathcal{E}$, its order statistics are related to the similarity of consecutive real projected texture vectors. While the first order statistic selects the more similar consecutive real projected texture vectors, i.e., $\min_{\mathbf{c}_i} \min_{\mathbf{c}_{i+1}} E_c$, the last one corresponds to the least similar sequence, i.e., $\max_{\mathbf{c}_i} \min_{\mathbf{c}_{i+1}} E_c$. Varying this order statistic can be therefore used to force transitions between very different real projected texture vectors, obtaining some sort of exaggerated or *emphasized visual transitions*. For mouth appearances, they contain

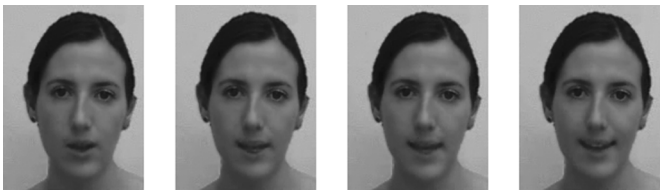

Fig. 7. Emphasis examples of four videos uttering the word \nina\. Each frame shows the uttering of \i\ with different emphasis level α of 0.0, 0.3, 0.7, and 1.0, from left to right.

the same viseme class sequence but are represented by real projected texture vectors which may not be close to each other (Fig. 7), i.e., same viseme classes but different visual units. This fact does not nullify realism, because there is still the videorealism given by the geodesic interpolation and the photorealism provided by the subspace $\mathbf{U}^r$. Further experimental results (Section IV-B) show that mouth synthesis realism does not only depend on the selected order statistic, but also on the communication style given by the audio mode of the global message.

The synthetic sequence with the least mouth movements attached to any phoneme set is obtained with a first-order statistic. The sequence of real projected texture vectors are as similar as possible, therefore generating a transition of a mouth that tries to hypoarticulate, i.e., to move as little as possible (becoming a short path in the subspace). If an expressive mouth is desired, i.e., a mouth with exaggerated visual speech movements, an i-order statistic must be used, obtaining greater exaggerated lip movements for larger values of i (see Section IV). This is due to the sequence of different real projected texture vectors, which are not visually similar among them (therefore generating longer paths in the subspace), therefore synthesizing a mouth that tries to hyperarticulate. Note that since viseme timings are given by the TTS (or also by a labeled sound waveform), the duration of transitions between real projected texture vectors remains unaffected by the selected order statistic.

As a result, visual speech emphasis can be regulated through a simple scalar value, which sets the specific sequence of real projected texture vectors to interpolate with the geodesic interpolation of Section II-C. To obtain a common measure for every viseme class, a normalized value α w.r.t. the index of the higher order statistic is used instead of raw i-order statistics. Scalar α becomes 0 for the first-order statistic and 1 for the largest one in each viseme class.

III. Facial Model Building

The PSFAM [16] can be automatically learned by the simultaneous alignment and subspace definition process described in Section III-A. It only needs a video file showing a frontal view of a real face while uttering a set of specific words of the consonant-vowel-consonant-vowel (CVCV) form, given in Section III-B. Parallel genetic algorithms [15] have been selected to find this minimum size word sets given the high nonlinearities of its definition.

TABLE I
Proposed Viseme Classes for Castilian Phonemes

1 - \a\	7 - \p\, \b\, \B\, \m\
2 - \e\	8 - \T\, \D\
3 - \i\, \I\	9 - \r\, \R\, \s\, \z\, \n\, \l\, \t\, \d\
4 - \o\	10 - \\, \tS\, \L\, \J\
5 - \u\, \U\	11 - \k\, \g\, \G\, \ × \, \N\
6 - \f\, \M\	12 - silence

A. Simultaneous Alignment and Subspace Definition

Face alignment is necessary to obtain a meaningful PSFAM (Section II-B) from a training video sequence [16]. This work is based on the well-known Lucas–Kanade tracking algorithm [50] and the subspace constancy assumption proposed in [51]. The latter needs a subspace centered around a specific mean information to allow changes of appearance in the tracking process. To obtain them, the singular value decomposition (SVD) is used as in [16], but following a scheme like those of [17], [52], and [53], which compute the SVD incrementally and take changes of the mean information into account. Consequently, a simultaneous and causal performance is obtained for tracking and subspace learning purposes, achieving an automatic process for aligning the training data. The former extracts the motion information from the object to be tracked using the learned subspace, while the latter provides an update on the appearance subspace with the registered object.

B. Optimal Corpora Description

In order to obtain good performance results in the synthesis process, with the geodesic interpolation of Section II-C and the visual selection process of Section II-D, the training sequence must show the widest range of possible mouth appearances. In this paper, a method for designing reduced size corpora for visual synthesis is proposed using parallel genetic algorithms (PGA), therefore avoiding time consuming recording activities attached to large corpora. Experimental results of Section IV show the good performance of this kind of corpora when used with the proposed algorithms.

The appearance variability is based on one of the conclusions of Summerfield's work [43]. He studied the visual similarity of English phonemes and found that the main discriminating visual feature among consonants was their place of articulation. This assumption is used in this work to cluster consonant phonemes into viseme classes through their place of articulation. Taking into account the 24 Castilian Spanish phonemes, six places of articulation can be considered [54]: bilabial, labio-dental, linguodental, palatal, alveolar, and velar. Regarding vowels, there are five in Castilian Spanish: \a\, \e\, \i\, \o\, and \u\. Consequently, the Castilian visemes can be classified into 12 viseme classes (five vowels, six consonants, and silence) as can be seen in Table I. Semivowels \ɪ\ and \ʊ\ are grouped with \i\ and \u\, respectively.

All visemes and their transitions must be observed in the uttered word set and their distributions must be as uniform as possible. The word set template consists of x disyllable words of the form CVCV since CV transitions account for more than 80% in Castilian Spanish [55] and are easy to pronounce, thus

reducing mispronunciation errors. They are pronounced in isolation and have no meaning, so users must pay attention when reading and uttering them, providing implicitly better visual articulation. They have stress in the first syllable, to provide some kind of the exaggerated appearances needed for the emphasis synthesis. Consonant clusters are not used because they increase the size of the corpus and can make word pronunciation difficult, thus increasing also mispronunciation errors. Possible syllables are: \ma\, \fa\, \θa\, \na\, \ɲa\, \×a\ \me\, \fe\, \θe\, \ne\, \ɲe\, \×e\ \mi\, \fi\, \θi\, \ni\, \ɲi\, \×i\ \mo\, \fo\, \θo\, \no\, \ɲo\, \×o\ \mu\, \fu\, \θu\, \nu\, \ɲu\, \×u\. In x words, there are $2x$ consonants, $2x$ vowels, x V-silence, x C-silence transitions, and $3x$ CV transitions (order is assumed not important). This leads to $2x/5$ instances of each vowel, $x/3$ instances of each consonant, $x/5$ of V-silence transitions, $x/6$ of C-silence transitions, and $x/10$ of CV ones. The smallest x to achieve uniform distributions of all elements is the $\mathrm{lcm}(3,5,6,10) = 30$ words.

In order to evaluate the suitability of different word sets, let $\mathbf{n}_{V\to sil}$ be a five-element vector containing the number of V-silence transitions, let $\mathbf{n}_{sil\to C}$ be a six-element vector containing the number of silence-C transitions and let $\mathbf{n}_{V\leftrightarrow C}$ be a 30–element vector with the number of transitions between each VC or CV combination (order is ignored). Finally, let $\mathcal{F}$ be the cost function and defined as follows:

$$\mathcal{F} = \frac{n_V + n_C}{11 + 5\sigma_f + 6\sigma_s + 30\sigma_t} \tag{6}$$

where the number of different vowels and consonants are represented by n_V and n_C, respectively; σ_f is the variance of vector $\mathbf{n}_{V\to sil}$, σ_s is the variance of vector $\mathbf{n}_{sil\to C}$, and σ_t, the variance of vector $\mathbf{n}_{V\leftrightarrow C}$. Constants 5, 6, and 30 are used to obtain the sums of squared differences (SSD) and 11 is included to restrict the values of $\mathcal{F}$ between 0 and 1.

The optimal solution ($\mathcal{F} = 1$) is not unique; however, the search space is extremely large. Therefore, exhaustive techniques cannot solve it in a reasonable amount of time and others have to be used. In this case, genetic algorithms [56] are selected due to the high dimensionality and nonlinearity features of this search problem; furthermore, PGA are chosen because they show increased performance and avoid premature convergence [15]. The meaningless word set used in this study is \×iθa ×ami ×anu maɲu ×eɲe nu×a fu×o θemi me×o moθa fu×i ɲimo feθi ɲafo fuma θuɲo noɲe nefa θemu ɲifi naθu moɲu mune ninu ×oθu ɲafe ɲina θi×e fone θofi\.

IV. Evaluation

Two experiments have been carried out to evaluate the contributions of this work: first, a brief analysis of the PGA performance to obtain optimal sets of words (see Section IV-A); second, a set of perceptual tests (see Section IV-B) to evaluate visual emphasis and realism features. Objective measurements can be given for the former; however, qualitative evaluation is needed in the latter since perceptual features like realism are very hard to obtain quantitatively [57].

Regarding the objective comparison among different proposals, there is an increasing interest in benchmarking them [58], [59], though these proposals only analyze their own related work, and it is difficult to find qualitative objective comparisons based on perceptual studies between different talking head systems.

A. Corpus Results

The recorded corpus was a 154-s sequence of 720×576 pixels and 25 frames per second, using 67 MB. The talker was a 22–year-old girl and her face and shoulders were recorded while speaking in front of the camera. She was asked to utter the CVCV words set of Section III-B once, in neutral speaking (not emphatic), and in an isolated way. She was also told to stress the first syllable of each word. The uniformly distributed word set was obtained with a PGA following the suggestions of [15]. The final configuration has been a fully connected 4-deme network with a migration rate of 0.01 and a total population of 2400 with 300 bits each. Elitism has been taken into account, as well as multipoint crossover and roulette selection; mating and mutation probabilities have been set to typical values of 1 and 0.001, respectively. The algorithm takes less than two hours to find a new word set 99% of the time with 2 MB of RAM in a 3.2-GHz Pentium IV processor. Note that this memory usage is related to obtain the word set used to record the corpus, not to the corpus itself.

B. Perceptual Tests

Three tests have been performed on 95 people from both sexes and with ages ranging from 14 to 56 years. They were not told previously about the contents of the tests to help obtain unbiased evaluations. Each test consisted of a set of video sequences. All videos were shown at 25 fps with a size of 336×256. Audio information is played at 8000 Hz in the videos that need an auditory context in the evaluation process. Since videos with real images were provided in the tests, real speech audio was included to avoid biased results due to different voices in the case that TTS speech was included in the synthetic videos. These audio streams were time labeled to allow synchronization with its synthetic versions. Mean opinion scores (MOS) have been used to measure the particular features of each video for each test, providing scores between 1 and 5 for each video. Control points were introduced to measure the consistency of each evaluator. They consisted of repeating videos randomly within each test (every video was repeated once) and consistency was obtained as the mean absolute difference between the scores given to repeated videos. The 50% most consistent evaluators were selected in this study (obtained consistency of < 0.8 points over 5). Specific values for each test are provided in Table II and Fig. 9. The synthetic videos were built with $\beta = 2$. The value of α varies in each test. Kolmogorov–Smirnov tests [60] have checked for similar distributions needed by Kruskal–Wallis tests [60], a kind of nonparametric analysis of variance tests which have been used due to the non-Gaussianity of the data.

The first test was made to check for the achieved visual realism of the synthesis. No audio was played in this test. Two sets

Fig. 8. Three frames uttering the same phoneme \i\ in a real video frame (left); a frame synthesized with the proposed method (center), and the same frame synthesized without region smoothing and with only one visual unit per viseme (right).

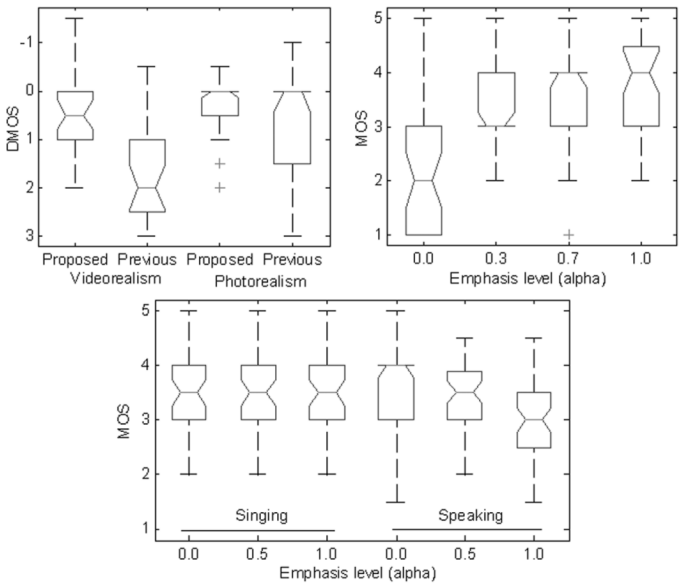

Fig. 9. Box plots of first (top left), second (top right), and third (bottom) tests. In the first test, *previous* results are referred to videos without selection algorithm for dynamic videos (related to videorealism) and videos without blending for static ones (related to photorealism).

TABLE II
PERCEPTUAL TESTS RESULTS

First test Realism		Second test Emphasis		Third test Mode Singing		Speaking
	DMOS	α	MOS	α	MOS	MOS
V	0.48	0.0	2.14	0.0	3.61	3.63
VWS	1.81	0.3	3.31	0.5	3.45	3.35
P	0.37	0.7	3.64	1.0	3.56	2.99
PWB	0.65	1.0	3.78			

of videos were shown: dynamic and static videos. The former set was used to evaluate videorealism, while the latter, to evaluate photorealism. Scores between 1 (minimum) and 5 (maximum) were given for each video. Dynamic videos were real, synthetic (V) and synthetic without the visual selection algorithm (VWS) of Section II-D (Fig. 8). They last 6 s each and were displayed in pairs real-V and real-VWS, uttering first the same three nonsense CVCV words ("nefa, cemu,ñifi") and then other three nonsense CVCV words ("ñafo, fuma, zuño"). Differential MOS (DMOS) were obtained for each pair from the difference between the two obtained MOS of each pair (inspired by recommendation ITU-R BT.500). Similarly, static videos were also real, synthetic (P), and synthetic without the blending effects (PWB) of Section II-B. Static images of stressed \i\ and unestressed \a\ were shown for real, P and PWB videos during 3 s each. Real-P and real-PWB were also displayed in pairs to obtain DMOS for each one. Synthetic videos were built with $\alpha = 0.2$. The visual selection algorithm improves the perceived realism from 1.81 (real-VWS DMOS) to 0.48 (real-V DMOS), with a very significant difference ($p < 0.00$). Synthetic videos (V) obtain a nonsignificant difference w.r.t. real instances at. 01 level (since $p < 0.03$). The proposed blending effects enhance photorealism from 0.65 (real-PWB DMOS) to 0.37 (real-P DMOS), also with a very significant difference ($p < 0.00$). Synthetic videos (P) achieve a nonsignificant difference ($p > 0.15$) with respect to the real sequences.

The second test evaluated whether α could be used to produce different synthetic videos varying only the visual emphasis effects described in Section II-E. The same sentence of 6-s duration was uttered in all four synthetic videos, with α values of 0.0, 0.3, 0.7, and 1.0. The sentence had unmeaningful content in the first half and meaningful content in the second one. The videos had real audio information, to give an auditory context to the speaker of the sounds being uttered. Reference synthetic video examples with extreme values of α (0 and 1) were given to normalize the evaluators' scores between 1 (minimum α) and 5 (maximum α). Significant differences at .05 level are revealed among pairs of α values, but not between 0.3 and 0.7 ($p > 0.54$). This result is translated by taking into account only three values in the last test. Moreover, there is a strong linear correlation (> 0.99) between the obtained MOS and $\ln(\alpha + 0.02)$. Note that the different emphatic content has been generated with a neutral corpus with stressed words.

The last test analyzed the general realism of the synthesis for different values of α and for two different communication styles: talking and singing. When singing, high α values are expected to be more helpful than when talking, because of the more exaggerated movements a singer can express. Evaluators were asked to assign a value from 1 (low realism) to 5 (high realism) to each video. Three α values (0.0, 0.5, and 1.0) for talking and singing videos give six synthetic videos. Real audio was used to be able to distinguish between talking and singing. Talking videos included the same 6-s sentence of the second test. Singing videos included a 13-s sung English verse to focus the attention on the movements and not on the content. Kruskal–Wallis reveals no significant differences ($p > 0.62$) for different emphasis levels when singing. However, it shows significant differences ($p < 0.01$) between emphasis levels when speaking, with higher realism for hypoarticulated movements (low α). The same realism is detected between singing and hypoarticulated talking ($p > 0.28$). Furthermore, an interaction is detected in this experiment ($p < 0.01$) between the emphasis level and communication style, e.g., maximum realism perception is given for low emphasis when speaking but its value is of little or no importance when singing.

V. CONCLUDING REMARKS

This paper proposes a novel method for automatic and realistic facial synthesis with emphatic control. To achieve realistic results, the visual synthesis relies on a novel viseme

selection algorithm based on a geodesic interpolation method. The viseme selection algorithm and a parameterization of visual emphasis with an α scalar provide the emphatic control. Moreover, a design of short, uniformly distributed corpora is proposed to reduce personalization efforts. PGA is used to obtain specific corpora description instances. Experimental results conclude that the new viseme selection algorithm obtains good videorealism (a DMOS of 0.48 out of 5 w.r.t. the real sequence) and photorealism (a DMOS of 0.37 out of 5 w.r.t the real images) levels. Experimental results also show a significant interaction between the perceived emphasis and two communication styles (speaking and singing); this fact would express the importance of visual emphasis in the perceived realism of a synthetic face.

Although we show promising results, future work needs to address 3-D changes in pose in the PSFAM and to include more communication styles like crying or screaming. Comparing the proposed reduced corpora design and the time-consuming ones of other works can also be carried out to determine if they can shorten this step without significant realism loss. Intelligibility tests with different emphasis levels are also planned.

REFERENCES

[1] E. André, "The generation of multimedia presentations," in *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, R. Dale, H. Moisl, and H. Somers, Eds. New York: Marcel Dekker, 2000, pp. 305–327.

[2] J. Morton and M. Johnson, "Conspec and conlern: A two-process theory of infant face recognition," *Psychol. Rev.*, vol. 98, no. 2, pp. 164–181, 1991.

[3] R. A. Hinde, *Non-Verbal Communication*. Cambridge, U.K.: Cambridge Univ. Press, 1975.

[4] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[5] D. Massaro, "Auditory visual speech processing," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 1153–1156.

[6] T. Chen, "Audiovisual speech processing: Lip reading and lip synchronization," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–31, Jan. 2001.

[7] D. Massaro, J. Beskow, M. Cohen, C. Fry, and T. Rodriguez, "Picture my voice: Audio to visual synthesis using artificial neural networks," in *Proc. AVSP*, Santa Cruz, NM, 1999, pp. 133–138.

[8] G. Bailly, "Audiovisual speech synthesis," in *Proc. ETRW Speech Synth.*, 2001 [Online]. Available: citeseer.ist.psu.edu/bailly03audiovisual.html

[9] J. Ostermann and A. Weissenfeld, "Talking faces—Technologies and applications," in *Proc. ICPR'04: Pattern Recognition, 17th Int. Conf.*, Washington, DC, 2004, vol. 3, pp. 826–833.

[10] F. Parke, "Computer generated animation of faces," in *Proc. ACM'72: ACM Annual Conf.*, New York, 1972, pp. 451–457.

[11] Y. Pei and H. Zha, "Transferring of speech movements from video to 3D face space," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 1, pp. 58–69, Jan. 2007.

[12] Y. Chang and T. Ezzat, "Transferable videorealistic speech animation," in *Proc. SCA'05: 2005 ACM SIGGRAPH/Eurographics Symp. Comput. Animation*, New York, 2005, pp. 143–151.

[13] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, GA, 1996, vol. 1, pp. 373–376.

[14] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," *Models and Techniques in Computer Animation*, pp. 139–156, 1993.

[15] E. Cantu-Paz, *Efficient and Accurate Parallel Genetic Algorithms*. Norwell, MA: Kluwer, 2000.

[16] F. D. la Torre and M. Black, "Robust parameterized component analysis: Theory and applications to 2d facial appearance models," *Comput. Vis. Image Underst.*, vol. 91, no. 1-2, pp. 53–71, 2003.

[17] P. M. Hall, D. R. Marshall, and R. Martin, "Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition," *IVC*, vol. 20, no. 13–14, pp. 1009–1016, Dec. 2002.

[18] E. Cosatto, "Sample-based talking-head synthesis," Ph.D. dissertation, Swiss Federal Inst. of Technol., Lausanne, Switzerland, 2002.

[19] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 388–398, 2002.

[20] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[21] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," *Proc. SIGGRAPH*, pp. 353–360, 1997.

[22] J. Beskow, "Talking heads: Models and applications for multimodal speech synthesis," Ph.D. dissertation, Dept. of Speech, Music, Hearing, KTH, Stockholm, Sweden, 2003.

[23] R. Gutiérrez-Osuna, P. Kakumanu, A. Espósito, O. García, A. Bojórquez, J. Castello, and I. Rudomin, "Speech-driven facial animation with realistic dynamics," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 33–42, Feb. 2005.

[24] T. Beier and S. Neely, "Feature-based image metamorphosis," in *Proc. 19th Annu. Conf. Comput. Graphics Interactive Techniques (SIGGRAPH)*, Chicago, IL, 1992, pp. 35–42.

[25] H. Hill, N. Troje, and A. Johnston, "Range- and domain-specific exaggeration of facial speech," *J. Vision*, vol. 5, no. 10, pp. 793–807, 2005.

[26] J. Beskow, B. Granström, and D. House, "Visual correlates to prominence in several expressive modes," in *Proc. Interspeech*, Pittsburg, PA, 2006, pp. 1272–1275.

[27] A. Syrdal, "Development of a female voice for a concatenative text-to-speech synthesis system," *Current Topics Acoust. Res.*, pp. 169–181, 1994.

[28] A. Black, "Perfect synthesis for all of the people all of the time," in *Proc. 2002 IEEE Workshop Speech Synth.*, 2002, pp. 167–170.

[29] J. Lasseter, "Principles of animation as applied to 3D character animation," *Comput. Graphics*, vol. 21, pp. 35–44, 1987.

[30] L. Redman, *How To Draw Caricatures*. New York: McGraw-Hill, 1984.

[31] P. Rautek, I. Viola, and M. Gröller, "Caricaturistic visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 12, no. 5, pp. 1085–1092, Sep./Oct. 2006.

[32] B. Lindblom, "Explaining phonetic variation: A sketch of the h and h theory," *Speech Production and Speech Modelling*, vol. 55, pp. 403–439, 1990.

[33] P. Ekman, *About Brows: Emotional and Conversational Signals*. Cambridge, U.K.: Cambridge Univ. Press, 1979, pp. 169–248.

[34] C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and f_0 variations," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 2175–2179.

[35] M. Swerts and E. Krahmer, "Congruent and incongruent audiovisual cues to prominence," in *Proc. Speech Prosody Conf.*, Nara, Japan, 2004, pp. 69–72.

[36] R. Scarborough, P. Keating, M. Baroni, T. Cho, S. Mattys, A. Alwan, E. Auer, and L. Bernstein, "Optical cues to the visual perception of lexical and phrasal stress in English," in *Proc. Int. Conf. Speech Prosody*, 2006, pp. 217–220.

[37] E. Krahmer and M. Swerts, "Testing the effect of audiovisual cues to prominence via a reaction-time experiment," in *Int. Conf. Spoken Lang. Process.*, Pittsburg, PA, 2006, 2006, paper 1288-Mon3A3O.4.

[38] G. Pourtois, D. Debatisse, P. Despland, and B. de Gelder, "Facial expressions modulate the time course of long latency auditory brain potentials," *Cognitive Brain Res.*, vol. 14, no. 1, pp. 99–105, 2002.

[39] M. Aylett, "Synthesising hyperarticulation in unit selection tts," *Proc. Interspeech*, pp. 2521–2524, 2005.

[40] R. Guaus and I. Iriondo, "Diphone based unit selection for catalan text-to-speech synthesis," in *Workshop Text, Speech, Dialogue*, Brno, Czech Republic, 2000, pp. 277–282.

[41] X. Huang, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[42] C. Fisher, "Confusions among visually perceived consonants," *J. Speech Hearing Res.*, vol. 11, pp. 796–804, 1968.

[43] A. Summerfield, *Hearing by Eye: The Psychology of Lip-Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987, ch. Some preliminaries to a comprehensive account of audio-visual speech perception, pp. 3–51.

[44] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, Freiburg, Germany, 1998, pp. 581–695.

[45] T. Jebara, K. Russell, and A. Pentland, "Mixtures of eigenfeatures for real-time structure from texture," in *Proc. Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 128–138.
[46] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins Univer. Press, 1996.
[47] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.
[48] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE Signal Process. Mag.*, vol. 13, no. 3, pp. 43–64, May 1996.
[49] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.
[50] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.
[51] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.
[52] J. Melenchón, L. Meler, and I. Iriondo, "On-the-fly training," in *Proc. AMDO*, Palma de Mallorca, Spain, 2004, pp. 146–153.
[53] J. Lim, D. Ross, R. L. ans, and M. H. Yang, "Incremental learning for visual tracking," *Adv. Neural Inf. Process. Syst.*, pp. 793–800, 2005.
[54] A. Ríos, *La Transcripción Fonética Automática Del Diccionario Electrónico de Formas Simples Flexivas Del Español: Estudio Fonológico en el Léxico*. Barcelono, Spain: Univ. Autónoma de Barcelona, 1999.
[55] M. de Vega, C.Álvarez, and M. Carreiras, "Estudio estadístico de la ortografía castellana: La frecuencia silábica," *Cognitiva*, vol. 4, no. 1, pp. 75–114, 1992.
[56] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston, MA: Addison-Wesley, 1989.
[57] J. Li, G. Chen, Z. Chi, and C. Lu, "Image coding quality assessment using fuzzy integrals with a three-component image model," *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 1, pp. 99–106, Feb. 2004.
[58] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," *MIT AIM*, 2003, 2003–003.
[59] S. Ouni, M. Cohen, H. Ishak, and D. Massaro, "Visual contribution to speech perception: Measuring the intelligibility of animated talking heads," *EURASIP J. Audio, Speech, Music Process.*, vol. 2007, Article ID 47891.
[60] D. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton, FL: Chapman & Hall/CRC, 2007.

Javier Melenchón received the B.Sc. and M.Sc degree in multimedia, the B.Sc. and M.Sc. degrees in computer science, and the Ph.D. degree, all from Enginyeria La Salle in Universitat Ramon Llull, Barcelona, Spain, in 2000, 2001, 2002, 2004, and 2007, respectively.

He worked as an Associate Professor in the Department of Communications and Signal Theory, Enginyeria La Salle, from 2001 to 2007 and joined the Multimodal Processing Research Group in the same institution in 2005. Since 2007, he has been working as an Associate Professor in the Estudis d'Informàtica, Multimèdia i Telecomunicació in Universitat Oberta de Catalunya, Barcelona, Spain. His main research interests are audio, image and multimodal signal processing, as well as machine learning.

Elisa Martínez received the B.Sc. degree in telecommunication engineering, the M.Sc. degree in electrical engineering, the Ph.D. degree from Enginyeria La Salle in Universitat Ramon Llull (URL), Barcelona, Spain, in 1993,1995, and 2000, respectively, the M.Sc. degree in computer vision from the Universitat Autònoma de Barcelona in 1998, and the M.S. degree in project management from URL in 2002.

She is a Professor in the Communications and Signal Theory Department, URL, and the head of the Multimodal Processing Research Group in the same institution. Her major research interests are computer vision, including structure from motion, visual guidance of mobile robots, and multimodal analysis/synthesis.

Fernando De la Torre received the B.Sc. degree in telecommunications and the M.Sc. and Ph.D. degrees in electronic engineering from Enginyeria La Salle in Universitat Ramon Llull, Barcelona, Spain, in 1994, 1996, and 2002, respectively.

In 1997 and 2000, he became an Assistant and an Associate Professor, respectively, in the Department of Communications and Signal Theory at the Enginyeria La Salle. Since 2005, he has been a Researcher at the Robotics Institute, Carnegie Mellon University (CMU), Pittsburgh, PA. His research interests include machine learning, signal processing, and computer vision, with a focus on understanding human behavior from multimodal sensors. He is directing the Human Sensing Lab (http://humansensing.cs.cmu.edu) and the Component Analysis Lab at CMU (http://ca.cs.cmu.edu). He has coorganized the first workshop on component analysis methods for modeling, classification, and clustering problems in computer vision in conjunction with CVPR-07, and the workshop on human sensing from video in conjunction with CVPR-06. He has also given several tutorials at international conferences on the use and extensions of component analysis methods.

José A. Montero received the B.Sc. degree in industrial engineering from Escola Universitària d'Enginyeria Tècnica Industrial de Barcelona, Universitat Politècnica de Catalunya, the M.Sc. degree in electrical engineering, and the Ph.D. degree from Enginyeria La Salle, Universitat Ramon Llull (URL), Barcelona, Spain, in 1992, 1998 and 2008, respectively.

He has been an Associate Professor in the Communications and Signal Theory Department, URL, since 1998.