Modeling and Synthesis of Facial Motion Driven by Speech

Payam Saisan[1], Alessandro Bissacco[1], Alessandro Chiuso[2], and Stefano Soatto[1]

[1] University of California, Los Angeles - CA 90095
saisan@ee.ucla.edu, {bissacco,soatto}@cs.ucla.edu
[2] University of Padova - Italy 35131
chiuso@dei.unipd.it

Abstract. We introduce a novel approach to modeling the dynamics of human facial motion induced by the action of speech for the purpose of synthesis. We represent the trajectories of a number of salient features on the human face as the output of a dynamical system made up of two subsystems, one driven by the deterministic speech input, and a second driven by an unknown stochastic input. Inference of the model (learning) is performed automatically and involves an extension of independent component analysis to time-dependend data. Using a shape-texture decompositional representation for the face, we generate facial image sequences reconstructed from synthesized feature point positions.

1 Introduction

Human facial motion carries rich information that we use to interact: We constantly read cues from people's faces, conveying a wide range of useful information often altering our state. While facial motion in isolation is quite interesting, the coupling with the action of speech adds yet another dimension to the problem. Our goal is to understand the dynamic behavior of facial motion as it relates to speech, and infer a model that can be used to generate synthetic sequences of images driven by speech. A great challenge in this task is the evolutionary acuity of human perception to details of the face and facial motion. For a facial model to meet this high standard, we must devise models that can capture subtleties. While there has been remarkable progress in the area of speech content recognition and general facial motion based on speech utterances [10,2,3], there remains an open question of capturing dynamic complexities and interactions between facial motion and speech signals. Such subtleties are encoded largely in the dynamics of facial motion as opposed to static pose geometry and photometry.

The problem is simple to state. We want to collect motion-capture data[1] for an individual, and the associated speech waveform, and from these data build a model that can be used to generate novel synthetic facial motions associated with novel speech segments, for instance for an animated character. However, we want to be able to do this while retaining the "distinctive character" of the

[1] In particular, trajectories of a collection of feature point positions in space.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3023, pp. 456–467, 2004.

individual person in the training set. For instance, if we observe Mr. Thompkins says “happy birthday,” our long term goal is to develop a model that can be used to synthesize novel facial motions that “looks” like Mr. Thompkins’.

The rationale of our approach is based on the fact that facial motion is the result of word utterances combined with physical characteristics of the face that are peculiar to each individual.

2 Relation to Previous Work and Contribution of This Paper

The topic of speech-driven facial animation has been the subject of considerable attention recently. A scheme for modifying emotional attributes of facial motion, such as happiness or anger, associated with utterances is discussed in [7]. In [10] Ezzat et al. propose a variant of the multidimensional morphable model as a representation for images, particularly effective in describing a set of images with local variations in shape and appearance. He uses this representation to develop a statistical interpolation technique, in the space of morphable models, to interpolate novel images corresponding to novel speech segments. In [2] Brand introduces the idea of driving the facial model with a related control signal derived from the speech signal. He introduces a modified hidden Markov model for identification of non-stationary piecewise linear systems. He uses this model to approximate the nonlinear behavior of the face via “quasi-linear” submanifolds. In [3], Bregler et. al propose an image-based method called “Video Rewrite.” This method relies on constructing audiovisual basic building blocks called triphones. It uses a large amount of training data to construct a basis for the entire utterance space. By identifying the correct audio-visual building blocks corresponding to a novel speech utterance and concatenating them it forms image sequences corresponding to novel speech segments. Unlike the past work on constructing generic facial motion synthesizers, we are interested in utilizing the information in speech to capture and drive a facial motion that is realistic and closer to the speaker’s personal dynamic character. Our goal is not to demonstrate a model that spans the entire utterance space, but at this stage to develop the concept and demonstrate its efficacy using only a small set of samples.

Our model decouples the deterministic dynamics driven by speech from the stochastic dynamics driven by samples from a stochastic process with unknown and non-Gaussian distribution. We show how to perform inference of this model, which involves independent component analysis (ICA) [8] applied to a dynamic context. We apply our inference algorithm to a model of a face based on decoupling transformations of the domain of the image from transformation of the intensity values, akin to so-called “active appearance” or “linear morphable models” [9,10,1]. However, unlike traditional active appearance models, we do not require manual selection and registration of interest points, but instead perform the learning automatically. Unlike [17], we do not use a pre-defined grid, but instead we use a geometric structure defined by salient regions of the images where geometric deformations are well-defined.

3 Modeling

In this section we describe a model that is motivated by the considerations above. We first describe the decoupling of appearance and motion, and then the decoupling of speech-driven motion, and noise-driven motion.

3.1 Modeling the Face: Shape and Radiance Decomposition

We make the assumption that a face is a smooth parameterized surface $S : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$, supporting a diffuse albedo $\rho : \Omega \times \mathbb{R} \rightarrow \mathbb{R}^+$, moving and deforming under the action of a group [2] $g(t)$, viewed under perspective projection $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, so that a given point $p = S(x)$ generates an image I at pixel $w(x, t)$ at time t according to

$$I(w(x,t),t) = \rho(x,t) \quad \forall\ x \in \Omega \tag{1}$$

where we have defined the "domain warping" $w(x,t) \doteq \pi(g(t)S(x))$. Without loss of generality we can assume that Ω corresponds to the image-plane at time $t = 0$. Note that the actual shape of the surface, i.e. the quotient $S/g(t)$, cannot be untangled from the deformation $g(t)$ in $w(x,t)$, and from the deformed radiance $\rho(x,t)$ and therefore the "responsibility" of modeling changes in radiance due to shape deformations is shared between the domain warping w and the range transformations ρ. Estimating the two infinite-dimensional functions $w : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^2$ and $\rho : \Omega \times \mathbb{R} \rightarrow \mathbb{R}^+$ in the case of a general scene is very complex, and falls into the general domain of *deformable templates* [13]. Here, we do not have a general scene, but various deformations of the same face due to speech. Therefore, in the spirit of active appearance models [9], we assume that local variability of the domain can be modeled as linear transformations of a number of basis elements:

$$w(x,t) = w_0(x) + W(x)y(t); \quad x \in \Omega,\ t = 1, 2, \ldots \tag{2}$$

where $W = [W_1, \ldots, W_{k_w}]$; $W_i : \Omega \rightarrow \mathbb{R}^2$ are basis elements, $y(t) \in \mathbb{R}^{k_w}\ \forall\ t$ In "active appearance models", one assumes that equation (2) is satisfied *not* on all of Ω, but only at a fixed number of known (often manually selected and registered) "landmark" points $x_1, \ldots, x_l$. Then $W(x_i),\ i = 1, \ldots, l$ can be estimated using principal component analysis (PCA). In [17], x_i are fixed points on a pre-defined grid, so no manual selection of landmarks is necessary. However, whether $w(x,t)$ in (1) can be inferred at a point x depends on the values of I in a neighborhood of $w(x,t)$. If x falls in a region of constant radiance, $w(x,t)$ is not well defined, which can result in unlikely domain deformations being estimated.

In this work, we adopt an intermediate approach, where we evaluate (2) only at photometrically distinct points, modeling the deformation of all and only the points where the deformation can be inferred. However, we rely on

[2] The deformation here is represented by a complex and possibly infinite-dimensional group. We will use a simpler model, which we will elucidate shortly.

the fact that we have a sequence of adjacent views of the image from video to automatically detect and track such photometrically distinct points and maintain point registration. We use a standard point tracker (our implementation of Lucas and Kanade's [20]) to track and obtain trajectories of a number of points on the face and thus the associated shape parameters $y(t)$. The process, including facial image synthesis, is further elaborated in section 5.

In the next subsection we discuss how to model the temporal evolution of these parameters. If the number of available points is small, we could bypass the dimensionality reduction of the deformation w and simply model the trajectory of all the landmark points $\{x_1(t), \ldots, x_l(t) \in \mathbb{R}^2\}_{t=1,\ldots,T}$. In either case, we call the "state" of interest $y(t)$, the latter case corresponding to $W = I$.

3.2 Modeling the Dynamics of the Face

In this section we model the temporal evolution of $\{y(t)\}_{t=1,\ldots,T}$. As we mentioned in the introduction, such evolution compounds the effect of deterministic speech and a more ephemeral input that is not associated with speech characteristics. Here we are interested in *decoupling* these effects.

One possible way is to assume that $y(t)$ is in fact the sum of two components $y(t) = y_d(t) + y_s(t)$, the first generated by a, say, linear system driven by the input sound channels $u(t)$, while the second, $y_s(t)$, generated through a linear system driven by an IID random process $e(t)$ with unknown distribution p_e, independent of $u(t)$. This kind of philosophy has been introduced in the area of subspace identification in [18] and further elaborated upon in [4,6,5].

Assuming that the dynamics of the "deterministic" and "stochastic" models are disjoint, one can give a state space description in decoupled form as follows. We introduce hidden "states" ξ, that we partition into two components: $\xi = [\xi_d^T, \xi_s^T]^T$, a "deterministic" one ξ_d that receives input from the sound channels $u(t)$, and a "stochastic" one ξ_s that receives input from an IID random process $e(t)$ with unknown distribution p_e.

While we have reasons to believe that the dynamics of facial motion can be faithfully modeled with a linear model (faces usually do not exhibit nonlinear behaviors such as limit cycles, bifurcations, or chaos, at least for the majority of individuals), in order to model the subtleties associated to each individual we allow the stochastic input to be drawn from a non-Gaussian distribution p_e. The model we consider, therefore, is in the following decoupled form

$$\begin{aligned} \begin{bmatrix} \xi_d(t+1) \\ \xi_s(t+1) \end{bmatrix} &= \begin{bmatrix} A_d & 0 \\ 0 & A_s \end{bmatrix} \begin{bmatrix} \xi_d(t) \\ \xi_s(t) \end{bmatrix} + \begin{bmatrix} B_d \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ B_s \end{bmatrix} e(t) \\ y(t) &= \begin{bmatrix} C_d \; C_s \end{bmatrix} \begin{bmatrix} \xi_d(t) \\ \xi_s(t) \end{bmatrix} + D_d u(t) + e(t) \end{aligned} \tag{3}$$

where $e(t) \overset{IID}{\sim} p_e$; We assume that the model above is stable and has *minimum phase* ($|\lambda(A_s)| < 1, \quad |\lambda(A_d)| < 1, \quad |\lambda(A_s - B_s C_s)| < 1$, where λ denotes the largest eigenvalue), and that $e(t)$ is a (strict sense) white process [3]. Further-

[3] I.e. $e(t)$ and $e(s)$ are indepedent for $t \neq s$

more, we assume that there exists a (square invertible) matrix D so that the components of

$$v(t) \doteq D^{-1}e(t) = [v_1(t), \ldots, v_{k_w}(t)]^T \tag{4}$$

are *independent* with density function $q_i(\cdot)$[4]. In the next section we argue that there are procedures to chose the dimension of the states ξ_d, ξ_s, but we shall not discuss this point in the paper. Note that we assume that the dynamics are decoupled (off-diagonal blocks of the transition matrix are zero). This is in the spirit of the so-called Box-Jenkins model, well known in the system identification literature [16]. The goal of the inference process (learning) is, given a sequence of measured trajectories $\{y(t)\}_{t=1,\ldots,T}$, to estimate the states $\{\xi_d(t), \xi_s(t)\}$, the model parameters A_d, A_s, B_d, B_s, C_s, C_d, D_d, the mixing matrix D and the non-Gaussian density of the stochastic input q. While the first part (identification of a model in decoupled form) has been studied in the system identification literature [18,6,5], dealing with (and estimating) a non-Gaussian driving noise is a non-standard task, which we discuss in the next section.

Once the model is identified, we can generate synthetic sequences by feeding the model with a speech input, and samples from the density q, as we explain in section 5.

4 Inference

In this section we discuss how to identify the model parameters and estimate the states of the model (3). Despite the linear structure, the model does not fall in the standard form suitable for applying off-the-shelf system identification algorithms, due to (a) the decoupled structure of the input-to-state relationship and (b) the non-Gaussian nature of the stochastic input. We will address these problems separately in the following subsections.

4.1 Combined Identification of the Model

In this section we concentrate on the identification of the model (3), following the approach proposed in [18,6,5]. Under a technical assumptions called "absence of feedback" (see Granger [12]) the stochastic processes y_d and y_s, called the *deterministic* and the *stochastic component* of y, defined by the conditional expectations

$$y_d(t) \doteq E[y(t) \mid u(t), u(t-1), ..., u(t-k), ...] \qquad y_s(t) \doteq y(t) - y_d(t) \tag{5}$$

are uncorrelated at all times [18]. It follows that $y(t)$ admits an orthogonal decomposition as the sum of its deterministic and stochastic components

$$y(t) = y_d(t) + y_s(t) \qquad E[y_s(t)y_d(\tau)^T] = 0 \quad \text{for all } t, \tau.$$

[4] Note that, if $y(t)$ is a full-rank purely non-deterministic process $e(t)$ has the same dimension k_w as $y(t)$.

Note that y_d is actually a *causal* linear functional of the input process, and is hence representable as the output of a causal linear time-invariant filter driven only by the input signal u. Consequently, $y_s(t)$ is also the "causal estimation" error of $y(t)$ based on the past and present inputs up to time t. Its input-output relation has the familiar form $y = F(z)u + G(z)v$ with "stochastic" and "deterministic" transfer functions $F(z) = C_d(zI - A_d)^{-1}B_d + D_d$ and $G(z) = I + C_s(zI - A_s)^{-1}B_s$.

Up to this point there is no guarantee that combining a state space realization of $F(z)$

$$\begin{aligned} \xi_d(t+1) &= A_d\xi_d(t) + B_d u(t) \\ y_d(t) &= C_d\xi_d(t) + D_d u(t) \end{aligned} \tag{6}$$

and one of $G(z)$

$$\begin{aligned} \xi_s(t+1) &= A_s\xi_s(t) + B_s e(t) \\ y_s(t) &= C_s\xi_s(t) + e(t) \end{aligned} \tag{7}$$

yielding (3) results in a minimal model (i.e. with the minimum number of state components).

In most practical cases the stochastic and deterministic dynamics will be completely different, and hence (3) will be minimal.

A subspace identification procedure based on this decomposition has been introduced in [18] and later refined and analyzed in a series of papers by the same authors [4,6,5] and can be summarized as follows. Using available data $\{y(t), u(t), t = 1, .., T\}$:

1. Estimate the deterministic component $\hat{y}_d(t) \doteq E[y(t) \mid u(1), u(2), ..., u(T)]$. (see [18,4,6,5] for details)
2. Use a standard "deterministic" subspace identification technique to identify the system parameters A_d, B_d, C_d, D_d. (see [6,5] for details)
3. Estimate the stochastic component $\hat{y}_s(t) \doteq y(t) - \hat{y}_d(t)$
4. Prefilter the stochastic component with a filter constructed from the identified deterministic system to compensate for a certain distortion due to the fact that only finite data are available (see [4] for details).
5. Use the prefiltered data as an input to the algorithm in [21] to estimate the stochastic parameters A_s, C_s, B_s.

The subspace procedures used in step (2.) and (5.) provide also order estimation techniques which allow to suitable choose the dimension of the states ξ_d and ξ_s, we refer to [4] for details.

4.2 Isolating the Stochastic Part: Revisiting "Dynamic ICA"

From the identification step we obtain a minimum-phase realization of the stochastic component of $y(t)$:

$$\begin{aligned} \xi_s(t+1) &= A_s\xi_s(t) + B_s D v(t) \\ y_s(t) &= C_s\xi_s(t) + D v(t). \end{aligned} \tag{8}$$

where $v(t)$, defined in equation (4), has independent components.

A *predictor* $\hat{y}_s(t|t-1)$ for the system output at time t is a (in general non-linear) function of the data up to time $t-1$, $\hat{y}_s(t|t-1) = f(y_s(t-1),\ldots,y_s(t-k),\ldots)$ that is designed to approximate the output $y_s(t)$ according to some criterion. The optimal predictor, in the sense for instance of minimum variance of the estimation error $y_s(t)-\hat{y}_s(t|t-1)$, is the conditional mean $\hat{y}_s(t|t-1) = E[y_s(t)|y_s(t-1),..,y_s(t-k),..]$. Under our assumptions (i.e. $v(t)$ strictly white and $(A_s - B_sC_s)$ stable) the predictor is just given by the inverse system of (8), which is named the "innovation model". The process $e(t) = Dv(t)$ is the "innovation process", i.e. the (optimal) one-step-ahead prediction error:

$$\begin{aligned} \hat{\xi}_s(t+1) &= (A_s - B_sC_s)\hat{\xi}_s(t) + B_sy_s(t) \\ e(t) &= y_s(t) - C_s\hat{\xi}_s(t) = y_s(t) - \hat{y}_s(t|t-1). \end{aligned} \tag{9}$$

At the same time, we want to enforce the constraint that the components of $v(t) = D^{-1}e(t)$, are independent; this can be done by minimizing the relative entropy (Kullback-Liebler divergence) between the joint density of $v(t)$ and the product of the densities $q_i(\cdot)$ of its components $v_i(t) \doteq D^{-1}_{.i}e(t)$:

$$\min_{D,q_i} K\left(|\det(D)|p_e(y(t)-\hat{y}(t|t-1))\Big|\Big|\prod_{i=1}^{k_w} q_i(D^{-1}_{.i}e(t))\right) \tag{10}$$

where $D^{-1}_{.i}$ denotes the $i-th$ row of the matrix D^{-1} and $K(p||q) \doteq \int \log\frac{p}{q}dP(x)$. This problem can be considered a dynamic extension of independent component analysis (ICA), a problem that has been addressed both in the literature of blind deconvolution using higher-order statistics [11] and in the learning literature [14, 22]. In particular, [11] estimates the parameters of a non-minimum phase system via blind deconvolution based on high-order cumulants.

Our assumption that the innovation are temporally strictly white allows to solve the dynamic ICA problem rather easily. Of course a more general model would not assume that the optimal prediction is linear or alternatively that the (linear) one step prediction errors are independent. The recent work [22] addresses the problem above, both for the case of non-linear models and for linear models, by using gradient descent algorithms. In the case of minimum-phase models as in (8), this approach do not fully exploit the structure of the (linear) problem. Therefore such a gradient procedure, when successful, cannot do better than a simple algorithm [19] that consists in a closed-form algorithm for identifying the model parameters, followed by a static ICA to whiten the components of the input. A similar approach has been advocated in [14].

In fact, since the system is assumed to be minimum phase and the inputs $e(t)$ temporally strictly white, as we have argued above, the optimal predictor is linear and depends only on second order properties on the process $y_s(t)$. The parameters A_s, C_s, B_s, can be recovered using linear system identification techniques. Subspace identification procedures as the ones previously described solve this problem and are particularly suited to work with high dimensional data (i.e. k_w large).

After the innovation model has been estimated, standard (static) ICA can be used to estimate a mixing matrix D and the density function $q(\cdot)$ from the residuals $e(t) = y_s(t) - \hat{y}_s(t|t-1)$. This method was first proposed in [19] as being suboptimal. As we have argued, with the hypothesis we make here, it is actually optimal.

The reader may ask what happens if the assumptions made in (8) are not satisfied. If the model is non-linear, then we know no better than running a large optimization problem in the fashion of [22]. If the model is non-minimum phase, our solution will yield the closest minimum-phase model, in the sense of minimum variance. Alternatively one can identify the model using high-order cumulants as in [11].

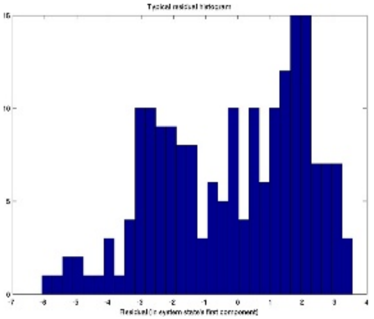

Fig. 1. Typical example of a residual histogram (sample approximation of q). Although the sample pool (number of frames) is small, the non-Gaussian nature of the distribution is clear.

5 Experiments

Face data were obtained using a 60Hz camera and tracking points on the lower region of the face for 200-300 frames. An implementation of the Shi-Tomasi feature tracker [20,15] was developed for this purpose.

We modeled face images using shape and radiance elements as in [9]. The shape element $s = (x_1, x_2, ..., x_l) \in \mathbb{R}^{2l}$ is defined by vertex coordinates (tracked points) of an n-point triangular mesh encompassing the face. Associated with every $s(t)$ is the supporting face albedo (texture), $\rho(x,t)$, such that $I(x,t) = \rho(x,t)$ where I is the face image at pixel x. To obtain sensible configurations of points to encompass the lower part of the face around the mouth and to reduce the number of outliers, we guided the feature selection by providing an image mask defining the regions to select features from. Figure 2 shows the configuration of tracked points on the subject's face. For every utterance sequence we obtained a training data set $s(t)$ and corresponding $\rho(x,t)$. Speech data was extracted from the synchronized audio signal. We used 256 periodogram coefficients as representation for speech segments corresponding to individual video frames, and we PCA reduced the dimensionality to arrive at $u(t) \in \mathbb{R}^4$. The choice of dimension here was a design parameter adjusted for best results.

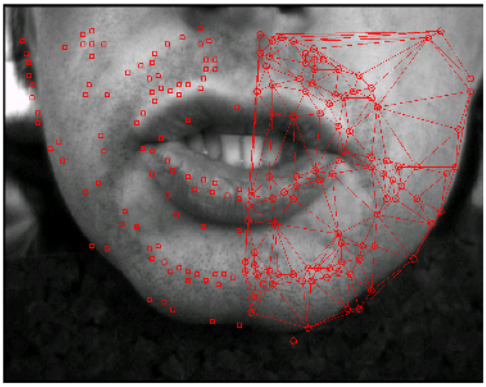

Fig. 2. Standard tracking schemes [20] were used to track feature point positions over 200-300 frames, sampled at 60 frames per second. About 200 points were selected in the first frame and tracked throughout the sequence.

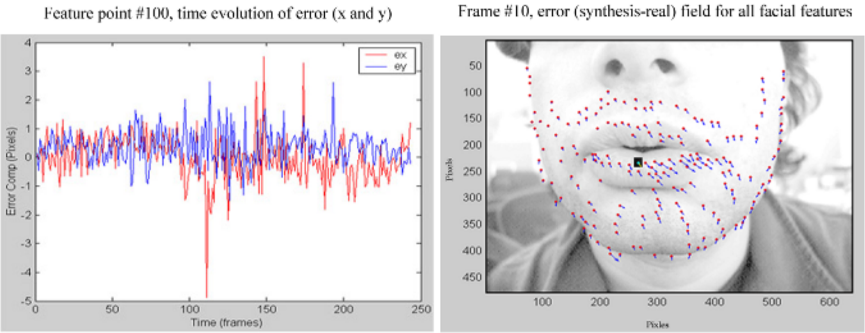

Fig. 3. Typical examples of error plots, feature position discrepancies between synthesis and actual data obtained via cross validation. The right figure is the the time evolution of error (discrepancy between synthesized feature motion vs. actual motion) for feature point number 100, a typical point near the center where fast and error prone motions occur. The left is the error vector field (synthesis-data) for all the points for a typical frame. The error for all feature points remained small for other frames in the sequence as depicted in this example.

Given $s(t)$ data we obtained the PCA reduced shape parameters $y(t)$ representing the output of the system. Following the inference procedure of section 4.1 we first identified the deterministic system parameters A_d, B_d, C_d, D_d using $y(t)$ as output and $u(t)$ as input. Then we identified the stochastic subsystem parameters as outlined in 4.2. As part of this step we get the non-Gaussian histograms corresponding to independent components of $v(t) = D^{-1}e(t)$ which is later used to generate, by sampling from the distribution, the random input driving the stochastic subsystem.

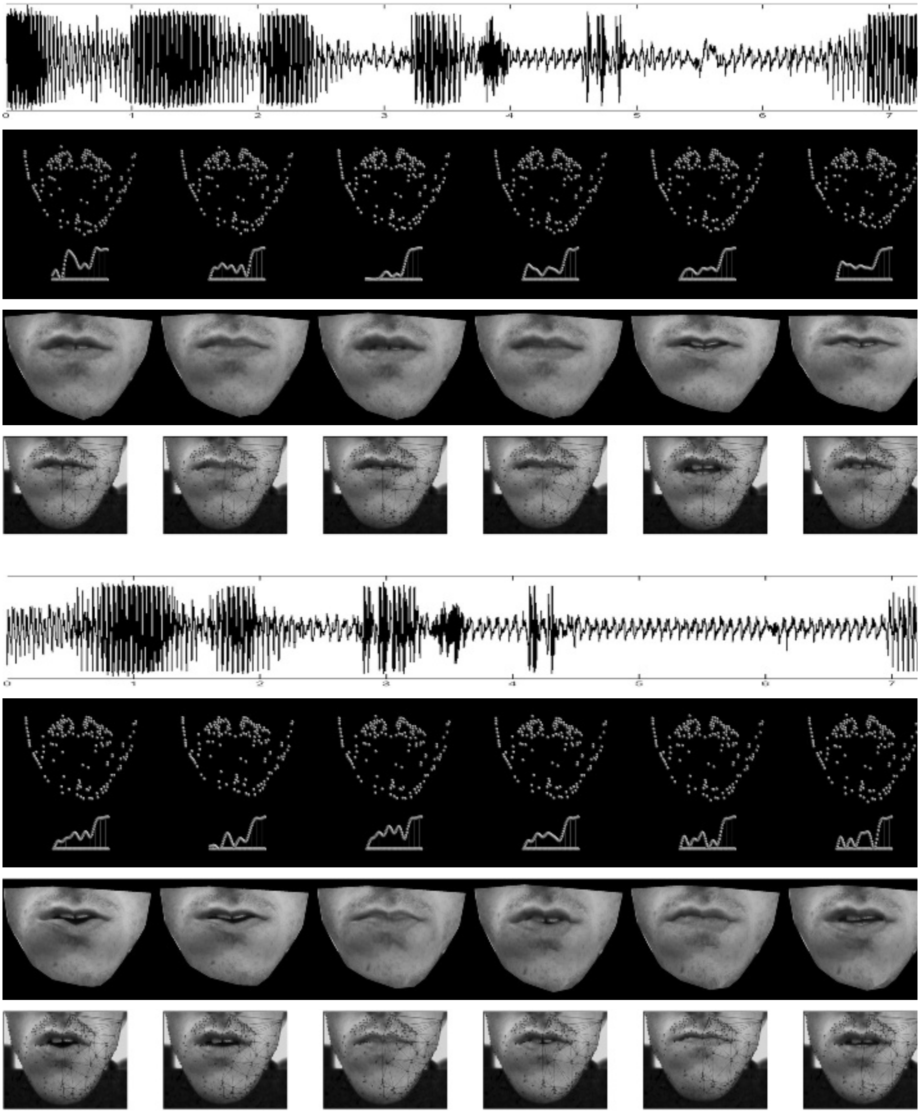

Fig. 4. An example of facial motion driven by a novel speech input. The subject is uttering the quote "live long and prosper". Cross validation technique was used where first half of the video is utilized for learning system parameters, and the speech part of the second half of the video is used for synthesis and validation. Row one is the speech signal, sampled at 44100 Hz. Row two is the system output, synthetic feature point positions. Row three is the full textured reconstruction of the face by way of identifying the "closest" feature configuration in the data-set and morphing its corresponding face image into the shape described by synthesic face pose. Row 4 is the actual facial motion associated with the data used for control proposes in cross-validation.

Given the identified model parameters and novel speech segment $u_n(t)$ we evolved the system (3) forward in time to obtain corresponding synthetic facial motion trajectories. This involved feeding $u_n(t)$ to the deterministic component of the system and drawing random samples from the non-Gaussian histogram of q to drive the stochastic component of the system. Note that here $u_n(t)$ corresponds to the same speaker and utterance as in the data, but it is a novel instance of it. For testing purposes we used only half of the data segments for training. The other half was used to extract the speech segment $u_n(t)$.

At the end we used the synthesized facial shapes, $s_n(t)$, to construct facial image sequences. For a given shape $s_n(t)$ we identified the closest (L_2 norm) shape s_i in the training data and morphed its corresponding albedo ρ_i onto $s_n(t)$. Facial image I at pixel x was computed according to $I(x) = \rho_i(\overline{w}(x, s_i, s_n(t))$ where we have defined the piecewise affine (backward) warp function $\overline{w} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as $\overline{w}(x, s_i, s_n) = A(x, s_i, s_n)x + b(x, s_i, s_n)$, with $A \in \mathbb{R}^{2x2}, b \in \mathbb{R}^2$. $\overline{w}$ maps pixel coordinate x within the triangular grid of s_n to point $\overline{w}(x, x_i, x_n)$ within the corresponding triangular grid of s_i.

The identification of the correct shape s_i from data to match $s_n(t)$ is, of course, highly non-trivial, particularly for systems designed to include the entire span of utterances. Such schemes would require construction of a basic alphabet for the space of utterances in the image space; visemes and other variants have been devised for this purpose and there are existing techniques for identifying viseme sequences corresponding to arbitrary speech waveforms. But in our case this is sufficient for demonstrating the efficacy of the modeling process which is mainly on the geometry of the face.

Motivated by the ultimate prospect of a real-time system, we relied on graphics texture mapping to achieve morphing of the matched albedos onto shapes of synthesized faces. That is, by creating a mesh, in this case 2D, using the shape vectors we mapped the matched albedos in the data onto novel facial shapes. The technique is computationally efficient and benefits from graphics hardware for texture mapping. The resulting dynamics was faithful to original utterances and reconstructed images exhibit no blurring artifacts[5].

6 Conclusions

We presented a method for modeling facial motion induced by speech. We used a representation for the face where geometric and photometric elements are decoupled. We modeled the dynamics of the geometric (shape) component using a linear dynamical system made up of two parts, a deterministic component driven by the speech waveform and a stochastic part driven by non-Gaussian noise. In our initial stage of development we show examples of the model at work using a set of various utterances, including digits and famous quotes. With this small set, we showed experimental results demonstrating the efficacy of our model in capturing the complexities of time dependent and multi-modal data.

[5] Sample movies of synthesized sequences can be downloaded from `http://www.ee.ucla.edu/$\sim$saisan/Face.html`

Acknowledgments. This work is supported by AFOSR F49620-03-1-0095, NSF ECS-0200511, CCR-0121778, and ONR N00014-03-1-0850.

References

1. V. Blanz, T. Vetter. A morphable model for synthesis of 3d faces. *Proceedings of ACM SIGGRAPH*, 187–194, 1999.
2. M. Brand. Voice Puppetry *Proceedings of ACM SIGGRAPH 1999*, 21–28, 1999.
3. C. Bregler, M. Covell, and M. Slaney Video Rewrite: Driving Visual Speech with Audio *Proceedings of ACM SIGGRAPH* , 353–360, 1997.
4. A. Chiuso and G. Picci. Subspace identification by orthogonal decomposition. In *Proc. 14th IFAC World Congress*, volume I, pages 241–246, 1999.
5. A. Chiuso and G. Picci. Subspace identification by data orthogonalization and model decoupling. *submitted to Automatica*, 2003.
6. A. Chiuso and G. Picci. Asymptotic variance of subspace methods by data orthogonalization and model decoupling. In *Proc. of the IFAC Int. Symposium on System Identification (SYSID)*, Rotterdam, August 2003.
7. E. Chuang, C. Bregler Facial expression space learning. *To appear in Pacifica Graphics*, 2002.
8. P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.
9. T. F. Cootes, G. J. Edwards and C. J. Taylor, Active Appearance Models. In *Proc. 5th European Conference on Computer Vision*, Freiburg, Germany, 1998
10. T. Ezzat., G. Geiger, T. Poggio. Trainable Videorealistic Speech Animation *Proceedings of ACM SIGGRAPH 2002*, 388–398, 2002.
11. B. Giannakis. and J. Mendel. Identification of nonminimum phase systems using higher order statistics. *IEEE Trans. Acoustic Speech and Signal Processing*, 37(3):360–377, 1989.
12. C. W. J. Granger Economic processes involving feedback In *Information and Control*, 6, 1963, pp.28-48/
13. U. Grenander Elements of Pattern Thoery. The Johns Hopkins University Press, 1996
14. A. Hyvärinen. Independent component analysis for time-dependent stochastic processes. 1998.
15. H. Jin, P. Favaro and S. Soatto. Real-time Feature Tracking and Outlier Rejection with Changes in Illumination. In *Proc. of the Intl. Conf. on Computer Vision*, July 2001
16. L. Ljung System indentification: theory for the user *Prentice-Hall, Inc*, ISBN 0-138-81640-9, 1986.
17. I. Matthews and S. Baker Active Appearance Models Revisited. In *International Journal of Computer Vision*, 2004.
18. G. Picci and T. Katayama. Stochastic realization with exogenous inputs and "subspace methods" identification. *Signal Processing*, 52:145–160, 1996.
19. P. Saisan and A. Bissacco Image-based modeling of human gaits with higher-order statistics. In *Proc. of the Intl. Workshop on Dynamic Scene Analysis*, Kopenhagen, June 2002.
20. J. Shi and C. Tomasi Good Features to Track *CVPR*, 1994.
21. P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.
22. L. Zhang. and A. Cichocki Blind deconvolution of Dynamical Systems : A State-Space Approach. *Proceedings of the IEEE. Workshop on NNSP'98*, 123–131, 1998.