Visual Coding and Tracking of Speech Related Facial Motion

Lionel Reveret, Irfan Essa

Visual Coding and Tracking of Speech Related Facial Motions

Lionel Reveret Irfan Essa

GVU Center / College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280, USA

Abstract

We present a new method for video-based coding of facial motions inherent with speaking. We propose a set of four Facial Speech Parameters (FSP): jaw opening, lip rounding, lip closure, and lip raising, to represent the primary visual gestures in speech articulation. To generate a parametric model of facial actions, first a statistical model is developed by analyzing accurate 3D data of a reference human subject. The FSP are then associated to the linear modes of this statistical model resulting in a 3D parametric facial mesh that is linearly deformed using FSP. For tracking of talking facial motions, the parametric model is adapted and aligned to a subject's face. Then the face motion is tracked by optimally aligning the incoming video frames with the face model, textured with the first image, and deformed by varying the FSP, head rotations, and translations. Finer details of lip and skin deformation are modeled using a blend of textures into an appearance model. We show results of the tracking for different subjects using our method. Finally, we demonstrate the facial activity encoding into the four FSP values to represent speaker-independent phonetic information and to generate different styles of animation.

1. Introduction and Motivation

In the context of face-to-face communication, speech is more than the transmission of an acoustical signal. The production of speech sounds is related to very specific and stable geometrical configurations of the lips and the jaw. Human beings are constantly exposed to both the acoustical stimuli and their visual correlates on the face. We perceive, and are very sensitive to, the spatio-temporal coherence between the sounds of speech and the facial gestures that are served to partially "shape" those sounds [6, 30]. Even animations of talking faces are subjected to this ontologic fact in order to convey a believable perception [20]. In this paper, we present a method to extract and encode the facial deformation associated with speech by spatio-temporal analysis of the video stream.

The complexity of the non-rigid deformation of facial movements coupled with the lack of robust features motivates the use of parameterized motion models to regularize the automatic analysis of face images. Such models have been used to track head movements, recognize expressions, and measure details up to the level of quantifying eyebrow raises and lip curls. Only a few simple attempts have addressed the coding and the automatic analysis of speech motion. This can be partly attributed to the lack of an existing experimental specification similar to the well-established Facial Action Coding System (FACS) proposed by Ekman and Friesen [11] for encoding motion of facial expressions. Previous works in automatic lip reading have attempted the recognition of a closed vocabulary (letters, digits, isolated words) or features tracking [9, 19, 23], but rarely a robust and high-level motion recovery has been addressed as it has been for expressions [8, 12, 3]. Some recent work on coding facial motions, called Facial Animation Parameters (FAPs), which are now included in the MPEG4 specifications, do model lip and mouth shapes [33, 29, 10]. However, the emphasis of this work is still aimed at animation and low-bitrate transmissions [31]. In fact, some of the above-mentioned contributions and a survey report [22] suggest the importance of building an encoding system that is more suitable for modeling visual speech. This is specifically the goal of our research effort.

In this paper, we show that an accurate model of speech motion learned from data of an expert subject can be re-used to track other subjects' face motion after a morphological (geometric, structural) adaptation. This model implements a high-level encoding of speech motion, which aids in the automatic visual recognition of non limited vocabulary (*i.e.*, not restricted to the learning set), as well as photo-realistic and non photo-realistic facial animation. We demonstrate the ability of this model to encode visual speech action parameters from video tracking of lips and face motion of talking subjects.

Our coding of facial motion for speech movement is based on four degrees of freedom which have been qualitatively described in the phonetic literature [26]. We demonstrate its capabilities to be used for tracking long sequences

of lip and face movements in a model-based approach. We improve on the tracking by adding the a texture-based approach, which provide increases robustness.

2. Related Work

There is considerable work in the area of face processing from video. Most of it concentrated on model-based tracking of face movement. Here we undertake a brief exposition of this motivating work that aids in the development our specific model for speech gestures. We specifically concentrate on earlier work on model-based tracking and on methods for extracting motion information from video.

Model-based tracking: DeCarlo and Metaxas [8] have successfully demonstrated the use of a parameterized face model to track movement of the head, smiling and mouth opening for different subjects. This approach uses a hand-designed model of face motion, with one single control parameter for the opening of the mouth. In the case of speech production, lips and face deform in a complex way, which cannot be represented with only one degree of freedom. The authors mention a need for better parametric representation of speech motion in their paper. The DeCarlo and Metaxas method of tracking adds a stronger model to extend the Black and Yacoob [3] approach, where simple affine motion models were used to measure deformations. Black and Yacoob relied on FACS model to model facial expression.

Physical models of faces have been proposed for analysis the facial motion as they allow for more degrees of freedom [12, 32]. However, the modeling is mainly focused on solving the tracking of canonical facial expression and does not, at present, model the motion of speech production.

Basu *et al.* [1] have addressed the motion of lip movement in speech production. In this work, a model of lip motion is learned from video for each subject from the tracking of ink markers on the lip surface. After the learning phase, this model allows for accurate tracking of outer lip contours feature, but does not implement a general coding of lip motion. About 10 degrees of freedom are necessary for each subject, which could result in instability in the optimization procedure.

Some recent techniques on analysis and synthesis of faces with speech have shown significant promise. For example, Video Rewrite [5] is an impressive technique that generates facial animations by reordering existing video frames. The choice of frames to play is determined by analyzing the audio track to extract phonemic information and its relationship to training video data. Voice Puppetry [4] is yet another impressive technique that claims to generate facial motion using the raw audio signal. It achieves this by learning a facial control model by analyzing video and audio of real facial behavior, automatically incorporating vocal and facial dynamics such as co-articulation. Both the Video Rewrite and Voice Puppetry techniques are however bound by needs of extensive data, with a related training phase on acoustical signal. We base our approach on the hypothesis that the *morphological variability of facial motion between different speakers is easier to solve (and scales better) than acoustical normalization.*

Model registration from images: Traditionally, optical flow has been used to provide pixel level information to align motion of the model onto the image brightness flow. The model-based approach in this case consists of regularizing the brightness consistency equation of the optical flow, into a model-based formulation from the a priori model of the face movement [8, 12, 3, 2, 18].

Some approaches show that a texture-based formulation of object tracking can be proposed as an alternative to optical flow. The Active Blob technique [27] implements a texture-based tracking of any deformable object with closed boundaries. Using statistical modeling of shape and texture, Active Appearance Models [7, 24] have been used to model and register differences in facial morphology. The texture presents a higher robustness than optical flow, as it is not subjected to error accumulation [28, 16, 15]. In addition, from the perspective of real-time implementation, recent developments in 3D graphic hardware for texture rendering, makes available high performance texture-mapping at a low cost.

3. Modeling of speech motion

3.1 The original model

Traditionally, Lipsynched animation relies on a set of lips and face shapes corresponding to the production of each phoneme (acoustical units of speech). These shapes are usually called *visemes*, for visual phonemes. Reveret *et al.* [26] present a detailed statistical analysis of 3D data of lip and face of one expert phonetician subject pronouncing all of the visemes. This statistical analysis allows to reduce the geometrical redundancy of the visemes and implements a model of lip and facial movements for speech. Following the same method, we collect 3D data using a 3D reconstruction of markers glued on the face of an expert phonetician. The video from three calibrated cameras is recorded, and the points were manually identified to extract a 3D reconstruction. In total, 34 face/lip shapes were captured (10 vowels, 8 consonants in 3 different vocalic context), each shape measured by 198 points in 3D. The resulting model showed that statistically, most of the variance of all the 34 visemes could be reconstructed with only 6 linear modes and a mean shape. Furthermore, these modes have been interpreted as phonetically pertinent gestures, consistent with [26]: (1) the opening the jaw; (2) the lip rounding,

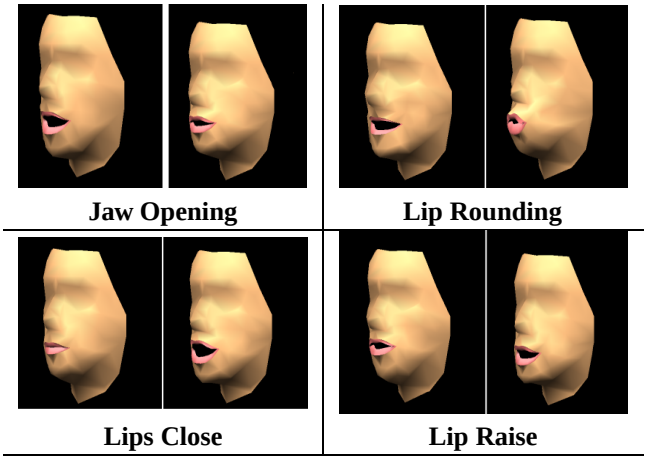

Figure 1: Our resulting 3D model and the extreme variations along the 4 FSPs (± 3 standard deviations from the mean shape). Opening of the jaw (FSP1), rounding of the lips (FSP2), closure of the lips (FSP3), raising of the lips (FSP4).

used to separate rounded vowel like [u] and spread vowel like [i]; (3) the closure of the lip for bilabial stop consonant like [p] [b] [m]; (4) lip raising for labio-dental fricatives consonants like [f] [v]; (5) advance of the jaw; and (6) a remaining motion due to the raising of the pharynx.

Using this spatial model of facial action allows us to generate a 3D model with actions represented as a linear combination of the 6 modes, $\phi_1, \ldots, \phi_6$, controlled by parameters that we introduce here as Facial Speech Parameters (FSP), $\mathbf{a} = \{a_1, \ldots, a_6\}$.

The last two parameters listed above have very limited variations, especially for the frontal views. In practice, we also observed that the last two parameters resulted in some instability for the automatic estimation from video. Consequently, we have chosen to ignore them for most of the analysis in this paper and focused on the automatic extraction of the first four FSP parameters.

Consider $\mathbf{X} = \{x_1, y_1, z_1, \ldots, x_n, y_n, z_n\}$ that describes a 3D geometric model, and μ as its mean shape, then we have a deformation model:

$$\mathbf{X}(\mathbf{a}) = \mu + \sum_{i=1}^{4} a_i \phi_i = \mu + \phi \mathbf{a}, \qquad (1)$$

which can be controlled by varying the FSPs (a).

The Figure 1 shows our resulting 3D model and the extreme variations along the 4 FSPs (± 3 standard deviations from the mean shape).

3.2. Aligning the model to different subjects

The procedure described above provides a detailed model of facial movements at the cost of a time-consuming hand labeling of markers on several shapes. To reduce this step, we introduce a method to align the morphology of the initial model on a new subject, while keeping the same description of motion learned from the reference subject. This hypothesis is based on the observation that, despite difference in morphology, any human vocal track is subjected to the same spatial constraints and therefore will deform in a similar way, including for the face, motion of the jaw and lips. This hypothesis is similar to the modeling of DeCarlo and Metaxas [8] in the sense that morphology and gesture are separately parameterized. In our case, we benefit from a detailed model of facial deformation for speech, learned from real data of a human subject.

Our normalization can be formulated as an update of the mean shape in equation 1, the remaining FSP modes being kept identical for the new subject. Having,

$$\mathbf{X_{ref}}(\mathbf{a}) = \mu_{ref} + \sum_{i=1}^{4} a_i \phi_i, \qquad (2)$$

from 1 for the reference subject, we model the new subject as,

$$\mathbf{X_{new}}(\mathbf{a}) = \mu_{new} + \sum_{i=1}^{4} a_i \phi_i, \qquad (3)$$

whose 3D mesh is controlled by the same FSP parameters. This normalization implies to find the numerical values of the new mean shape μ_{new}.

To process this update, the 3D model of the reference subject in a rest position is mapped onto the face of the new subject by feature alignment. These features are represented into a simplified model of the face in Figure 2, with each node corresponding to a specific node in the original 3D model. The mapping is performed by searching for the rotation, translation and scaling factors that best match the 2D front view projection of the 3D model nodes corresponding to the user specified features.

Once this first alignment has been made, local alignment of the mesh vertices is done. A Radial Basis Functions (RBF) relaxation of the displacements interpolates the displacement of the remaining vertices not covered in the simplified mesh [14].

Figure 2 shows the results of the alignment of the reference model to a new subject.

For both subjects, reference and new, the rest position is coded by the same FSP configuration $\mathbf{a_{rest}}$, which sets the initial model into a position where jaw and lips are closed (FSP_1, FSP_3), with a neutral spreading of the lips (FSP_2) and no raising of the upper lip (FSP_4).

Consequently, given $\mathbf{X_{ref}}(\mathbf{a_{rest}})$ being the rest position of the reference subject and $\mathbf{X_{new,rest}}$ being the result of the morphologic adaptation process described above for the new subject in a similar rest position, we obtain the mean

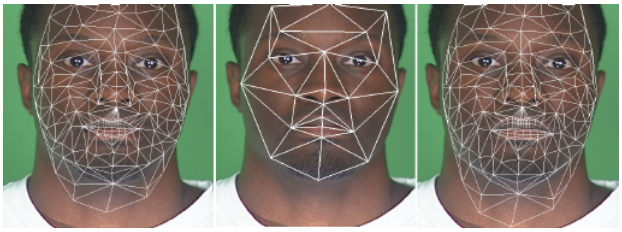

Figure 2: The original model for the reference model; the hand label features set on the new subject; the result of the original model aligned on the specified features.

μ_{new} for the new subject simply by substitution. Per Equation (1), we have

$$\mathbf{X}_{ref}(\mathbf{a}_{rest}) = \mu_{ref} + \phi \mathbf{a}_{rest}. \quad (4)$$

Then using Equation (1) for $\mathbf{X}_{new,rest} = \mathbf{X}_{\mathbf{new}}(a_{rest})$, we get

$$\mu_{new} = \mathbf{X}_{new,rest} - \phi a_{rest}. \quad (5)$$

To validate the articulatory hypothesis (*i.e.,* the usage of the same FSP modes for different subjects), we hand labeled three different speakers, doing 6 lip and shapes configuration, while uttering [a], [i], [u], [p] in [apa] and [f] in [afa]. The FSP configuration is recovered from the features location by an optimizing procedure that minimize the distance between the labeled features and the projection of the corresponding model points.

We obtain the following results, showing a pertinent repartition of the shapes according to the FSP interpretation even after the morphological adaptation (Figure 3). The jaw parameter (FSP1) isolate the [a] shape (wide opening), the protrusion parameter (FSP2) separates [u] from [i] shapes, the lip closure parameter (FSP3) separate the vowels from consonants that require a joining of the lips and finally the lip raising parameter (FSP4) separate the [p] consonants from the [f] consonants.

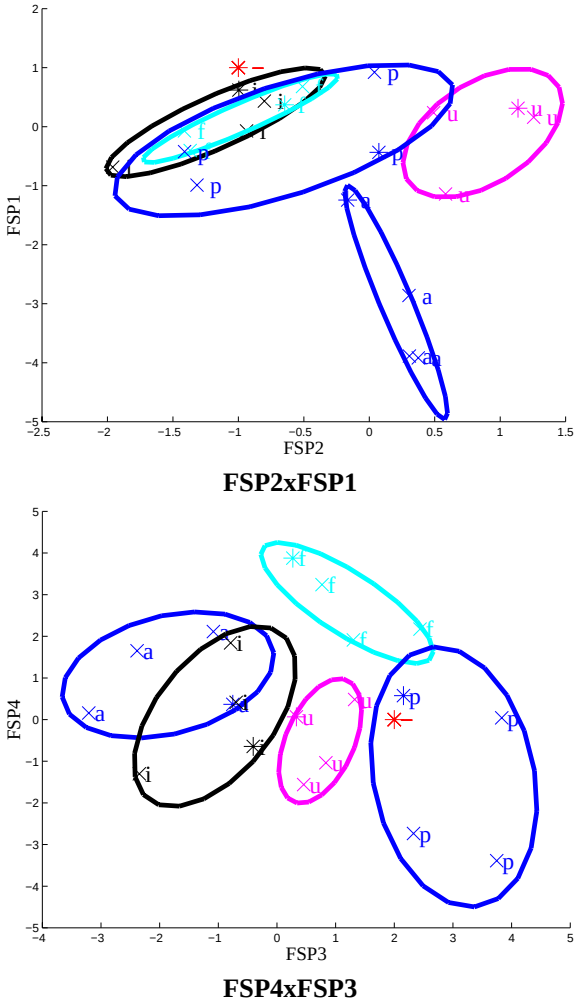

Figure 3: Plots showing variations in FSP amongst our subjects. The value of the FSP parameters of the reference speaker are represented with stars, while the other 3 new speakers are represented with crosses. They are identical for the rest position by construction (see section 3) and differ for all the other shapes. This aids in the validating our model.

4. FSP registration from texture

4.1 Objective function

As mentioned in the introduction, registration from texture presents an interesting alternative to optical flow, as it is not heavily penalized by risk of drift. For any new face, taking the initial image for the morphological alignment allows us to set a texture correspondence for the model. Now the texture follows subsequent deformation of the model and provide a synthetic image of the face. The tracking consists of finding the four numerical value of the FSP, plus translation and rotation that minimizes the difference between the projection of the textured model and the incoming image to analyze. We present this mathematically as follows:

For a set of position, rotations, and FSP parameters, let's introduce $p = [r, t, a]$, I is the raw image to analyze, $\hat{I}(p)$ the image synthesized by the textured model and $\|\ldots\|$ is the Euclidean norm on the RGB components of the image. The objective function is defined as :

$$E(p) = \frac{1}{n}\sum_{i=1}^{n} e_i(p)^2, \quad (6)$$

$$e_i(p) = \|\hat{I}(p)(x_i, y_i) - I(x_i, y_i)\|, \quad (7)$$

where (x_i, y_i) defines the screen position of the pixel i in the image of the textured model and in the original image. The

n pixels considered are only those covered by the model projection.

We improve the robustness of this objective function by using a robust norm instead of the Euclidean norm, in order to reject outliers. We use the Geman and McClure robust error norm [13] parameterized by a threshold σ:

$$\rho(x,\sigma) = \frac{x^2}{\sigma + x^2}. \tag{8}$$

The objective function to minimize is now:

$$E(p) = \frac{1}{n}\sum_{i=1}^{n} \rho(e_i(p),\sigma). \tag{9}$$

4.2 Levenberg-Marquardt optimization

The Levenberg-Marquaardt optimization solves a non-linear least square minimization and therefore is well suited to model-based tracking by texture alignment as formulated above [24, 27]. The Levenberg-Marquardt algorithm requires the first and second derivative of the function to minimize with respect to every parameter, *i.e.,* the 3 rotations, the 3 translation and the 4 FSP parameters.

For the first derivative and the second derivatives, we have:

$$\frac{\partial E}{\partial p_j} = \frac{1}{n}\sum_{i=1}^{n} \rho'(e_i(p),\sigma)\frac{\partial \hat{I}(p)}{\partial p_j} \tag{10}$$

$$\begin{aligned}\frac{\partial^2 E}{\partial p_j \partial p_k} &= \frac{1}{n}\sum_{i=1}^{n} \rho''(e_i(p),\sigma)\frac{\partial \hat{I}(p)}{\partial p_j}\frac{\partial \hat{I}(p)}{\partial p_k} \\ &+ \rho'(e_i(p),\sigma)\frac{\partial^2 \hat{I}(p)}{\partial p_j \partial p_k}\end{aligned} \tag{11}$$

The partial $\frac{\partial \hat{I}(p)}{\partial p_j}$ with respect to a particular model parameter p_j is approximated by perturbing the parameter by a small δ, warping that model, and then measuring the resulting change in the residual error. The hardware texture mapping capability is very valuable for gradient calculations here [27].

The optimal value of the parameters p is then up-dated by iteratively taking the step δp that solves the linear equation formed from the approximation of the Hessian $\mathbf{H}$ and the gradient $\mathbf{g}$:

$$(\mathbf{H} + \lambda \mathbf{1})\delta p = -g, \tag{12}$$

$$\mathbf{g} = [\frac{\partial E}{\partial p_j}]_{j=1\ldots 10}. \tag{13}$$

As mentioned in [25], the second derivatives of the image can be omitted in the formulation of the Hessian:

$$\mathbf{H} = [\frac{\partial^2 E}{\partial p_j \partial p_k}]_{j=1\ldots 10, k=1\ldots 10}. \tag{14}$$

The λ coefficient in Equation (12) serves to stabilize the inversion of the Hessian matrix when it is rank deficient.

Figure 4: Example of textures blending. On the left, the original image. On the center, the result of the 3D model aligned on this shape, textured with the image corresponding to the rest position. On the right, the effect of texture blending, showing that the required wrinkles appear on the surface of the lips, while the texture corresponding to the face in a rest position has a low blending coefficient.

5. Multi-texturing

Using a single texture for tracking is quite limited as it make a erroneous link between the lighting conditions and the geometric deformation of the model: as the model geometry is changing, the surface normals are changing as well but the lighting appearance is not. It results in unrealistic location of the lighting distribution, which is important for the optimization algorithm to align texture. One way to alleviate this problem is to estimate the lighting orientation of the environment and simulate on the geometry of the model. However, in presence of complex mixture of lighting sources (ambient, directional, etc.), this process is difficult to automate.

In LaCacia *et al.* [16], a model of lighting variation is learned off-line to allow an illumination-independent decomposition of texture images. We propose here an approach based on texture blending.

In addition to cope with the lighting problem, textures blending allowed us to simulate small detail of skin deformation such as wrinkles, which are naturally appearing in the production of speech movements but are not geometrically represented by the sparse geometric mesh.

The following sections explain how the multi-texture is formulated and how it can be integrated in the optimization process, as well as the choice of the basis of texture images.

Linear alpha-blending formulation

For a given subject, the six reference textures are chosen, corresponding to [a], [i], [u], [p] in [apa] and [f] in [afa]. The 3D model is aligned onto the images using the same procedure of features selection described in Section 3. Once the 3D mesh is aligned onto each image, the textures are blended into a linear class [24, 7].

$$\hat{I}(p) = \sum_k \omega_k(a) I_k, \tag{15}$$

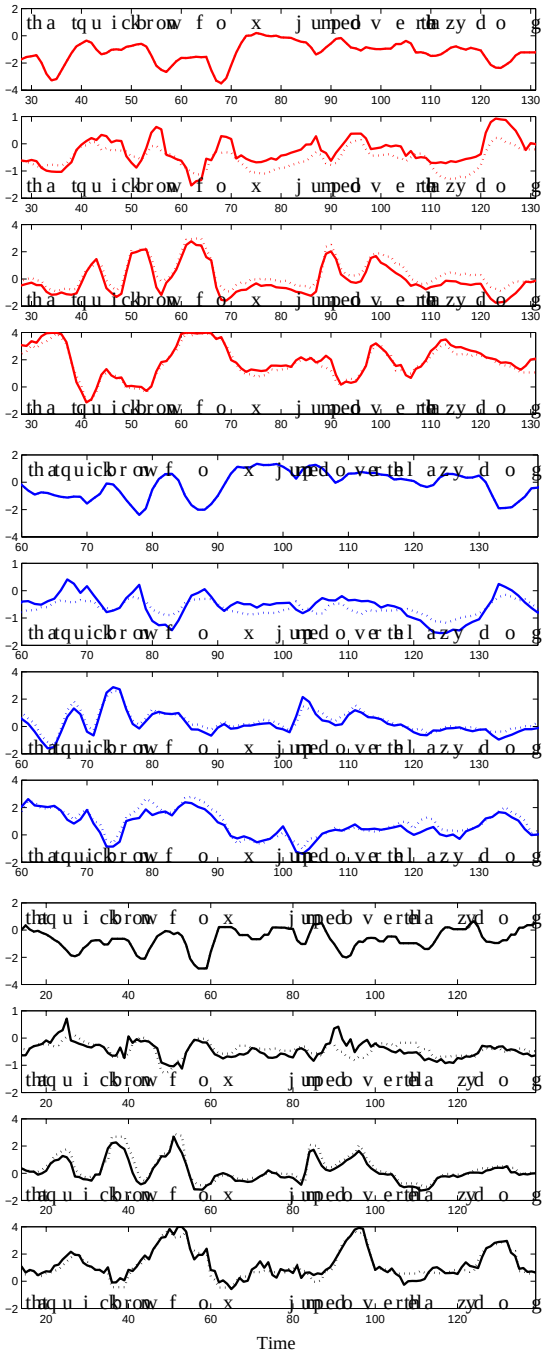

Figure 5: Results of tracking from video the four FSP for three subjects, uttering the same sentence. The dashed line correspond to the tracking with one single texture. The plain line correspond to the tracking with a multi-textures blending. Subject 1: RED, Subject 2: BLUE, and Subject 3: BLACK.

$$\omega_k(a) = e^{-\lambda_k \|X(a) - X_k\|^2}, \qquad (16)$$

where $\|X(a) - X_k\|^2$ is the sum of the Euclidean distances over all the vertices between the current shape $X(a)$ and the reference shape X_k.

The alpha coefficients are chosen as a set of radial basis function of the mean square distance between the vertices of the current 3D mesh and the vertices of the 3D mesh corresponding to the reference texture. The coefficient ω_k have been experimentally evaluated.

In the tracking procedure, the multi-texture showed an improvement in the detection of rounded shapes (“q[U]ick, br[OW]n”) by creating stronger minima of the error function at high value of FSP2 parameters, which corresponds to the rounding of the lips that occurs for this class of vowel.

6. Experiments and Results

FSP parameters extraction

We have tested the tracking of several different subjects uttering the same sentence: “That quick brown fox jumped over the lazy dog.” The subjects have been filmed with frontal faces with a somewhat controlled lighting (the multi-texturing aids with lighting variation as a person speaks).

The figure 7 displays the overall quality of the tracking for 6 important frames in the sequence of one subject. We have included an MPEG1 video `491-1.mpg` to show the tracking in action.

The goal of the FSP description is first to provide a model that constrains the variation of the facial motion to a subspace specific to speech gesture. This approach allows us to introduce robustness into the parameter estimation. The risk of a drift and loss of the tracker is very low as the model remains on the constrained subspace.

In addition to this robustness in tracking, the claim of this approach is to bring a motion description that is independent of the speaker. The Figure 5 shows the results of the extraction of the four FSP for each subject separately. It illustrates the influence of using the blending of several textures, instead of a single texture. As expected, we observe an emphasis on the rounding parameter FSP2 for rounded vowels. However, using several textures seems to introduce some instability in the trajectory of the extracted parameters. This could be attributed to the introduction of more local minima, as the texture of the model is able to change over time.

All our subjects exhibited a stable behavior when we compared linguistically meaningful information across different speakers. Here are some of the our observations based on our experiments.

- the first FSP (opening of the jaw) shows a drop on the pronunciation of the [o] in “fox” and “dog”, corresponding to the jaw opening.

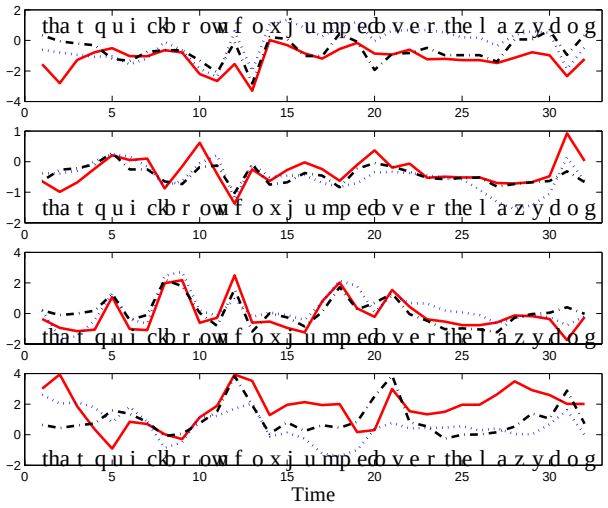

Figure 6: Results for three subject, time-aligned on the phonemes of the same uttered sentence. Subject 1: RED, Subject 2: BLUE, and Subject 3: BLACK.

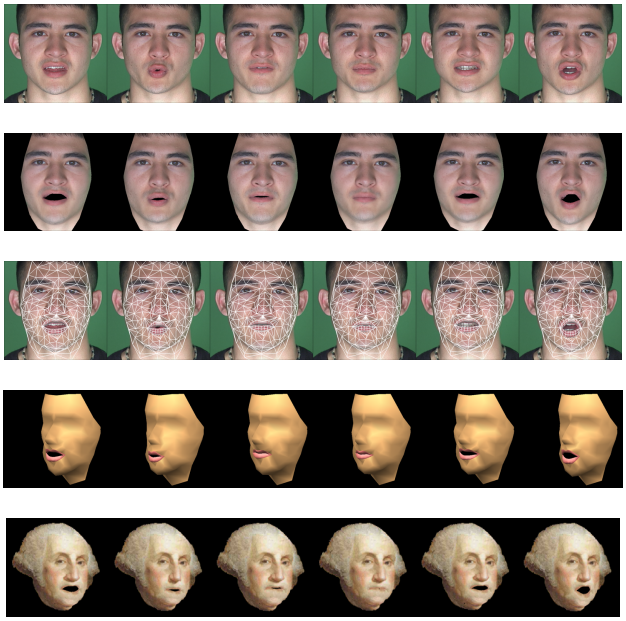

Figure 7: The results at 6 key frames for one subject. Top to bottom: original sequence, textured model, wire-frame on top of model, shaded/animated model and still picture animation from the data of this subject. See `491-2.mpg`.

- the second FSP (lips rounding) reaches maxima on the expected rounded vowels [u] and [o]. The use of multiple textures (plain line) allows a better detection of these vowels, compared to the single texture case (dashed lines)
- the third FSP (lips closure) appears for the pronunciation of stop consonants ([b] in "brown" and [p] in "jumped"). This parameters shows high values for [f] and [v] consonants. However, these consonants are separated form the [p] and [b] consonants thanks to the fourth FSP (lips raising).

In Figure 6, we have combined the results for every subjects in a time-aligned representation. The time-alignment was done manually using the audio stream. There do exist systems to allow for such time alignment automatically in the speech community.

Application to facial animation

The texture-based model used for the tracking by alignment on images provides a photo-realistic facial animation solution (Figure 7 [bottom]). In addition to morphing different shapes, the blending of different views would allow a better full 3D representation as in [24].

As an extension of this, using the procedure for aligning the model on a face shape, we have used the extracted FSP from video to animate the picture showed in Figure 7 (MPEG1 video `491-2.mpg`). As the FSP encodes natural gesture of speech production, this results in possibilities of high quality facial animation from one image only.

Finally, we have implemented an animation of 3D NURBS-based characters from the FSP parameters extracted from video. Instead of defining a lip and face shape for each phonemes like in a traditional lisynching process, we use a shape corresponding to the extreme variation of each FSP. Figure 8 shows an example of morphing targets suggested for the animation from the FSP extracted from video.

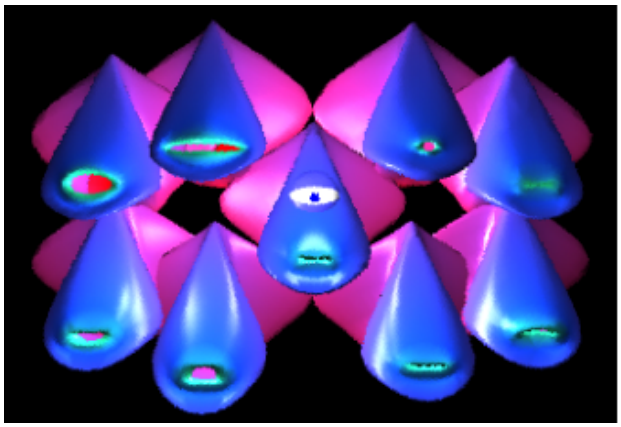

Figure 8: Morphing targets associated to the extremes positive and negative variations of the four FSP for a NURBS-based computer graphic character.

7. Summary and Conclusions

We present a novel method for extracting detailed facial movements of a speaker from video. We learn a detailed 3D model of facial motion for speech movements from a careful 3D labeling of a reference subject. This model is then linearly controlled by 4 parameters, that we introduced as FSP for Facial Speech Parameters. We show the use of these parameters as a coding of talking faces and implement them as the models to encode speech actions from video. We introduce a model-based tracking approach with texture registration and blending. A normalization of the face morphology allowed to use this model to track other subject and extract consistent motion information. We demonstrate the use of this coding for animation.

The relationship of our coding to phonetic description is perhaps very exciting. We feel that such coding via FSP could be used as well to provide efficient visual cues for audio-visual speech recognition and bring robustness to the automatic speech recognition system. The main advantage of our FSP coding for video tracking relies on the fact that it constrain the complex behavior of facial movement of speech to only 4 degrees of freedom. However, this modeling currently does not cover motion variation due to expression that could occur with speaking. One of the natural extension of this work will be to investigate how the FSP coding could be extended to cope with facial expression and still preserves a stable behavior for tracking from video.

References

[1] S. Basu, N. Oliver, A. Pentland, "3D Modeling and Tracking of Human Lip Motions", *Proc. of ICCV'98*, Bombay, India, Jan. 4-7, 1998

[2] S. Basu, S. and I. Essa. "Motion Regularization for Model-based Head Tracking.", *Proceedings of International Conference on Pattern Recognition*, Vienna, Austria, August 1996

[3] M. J. Black and Y. Yacoob. Tracking and recognizing facial expressions in image sequences, using local parameterized models of image motion. Technical Report CAR-TR-756, Center of Automation Research, University of Maryland, College Park, 1995.

[4] M. Brand. "Voice Puppetry." *In Proceedings of ACM SIGGRAPH 1999*, pp 21-28, August 1999.

[5] C. Bregler, M. Covell, and M. Slaney. "Video Rewrite: Driving visual speech with audio". In Proc. ACM SIGGRAPH '97,1997.

[6] R. Campbell, B. Dodd, and D. Burnham, (Eds.) *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech.* Psychology Press Ltd., East Sussex, UK. 1998.

[7] T. Cootes, G. Edwards, C. Taylor, "Active Appearance Models", *Proc. of the ECCV'98, Vol. 2, pp.484-498, 1998*.

[8] D. DeCarlo, D. Metaxas, "Optical Flow Constrains on Deformable Models with Applications to Face Tracking", in *International Journal of Computer Vision*, 2000.

[9] S. Dupont and J. Luettin. "Audio-Visual Speech Modelling for Continuous Speech Recognition", In *IEEE Transactions on Multimedia* Vol2, No 3, p141-151, 2000.

[10] P. Eisert and B. Girod. Analyzing facial expression for virtual conferencing. *IEEE Computer Graphics & Applications*, 18(5), September – October 1998. ISSN 0272-1716.

[11] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Consulting Psychologists Press Inc., 577 College Avenue, Palo Alto, California 94306, 1978.

[12] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[13] S. Geman, D.E. McClure, "Statistical methods for tomographic image reconstruction", *Bull. Int. Statistic Institute*, LII-4, 5-21, 1987.

[14] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1998.

[15] T.S. Jebarra, A. Pentland, "Parametrized Structure from Motion for 3D Adaptive Feedback Tracking of Faces", in *Proc. of CVPR'96*, CVPR, 1996.

[16] M. La Cacia, S. Sclaroff, V. Athitsos, "Fast, Reliable Head Tracking under Varying Illumination: An Approach based on Registration of Texture-Mapped 3D Models", in *IEEE Transactions on PAMI*, Vol. 22, No. 4, pp. 322-336, April, 2000.

[17] J. J. Lien, T. Kanade, J. F. Cohn, C. C. Li, and A. J. Zlochower. Subtly different facial expression recognition and expression intensity estimation. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1998*, pages 853–859, 1998.

[18] K. Mase. Recognition of facial expressions for optical flow. *IEICE Transactions, Special Issue on Computer Vision and its Applications*, E 74(10), 1991.

[19] K. Mase and A. Pentland. Lipreading by optical flow. *Systems and Computers*, 22(6):67–76, 1991.

[20] D. W. Massaro, M. M. Cohen, J. Beskow, and R. A. Cole, "Developing and evaluating conversational agents." In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (Eds.) *Embodied conversational agents.* Cambridge, MA: MIT Press, 2000.

[21] F. I. Parke and K. Waters. *Computer Facial Animation*. AK Peters, 1996.

[22] C. Pelachaud, N. Badler, and M. Viaud. Final Report to NSF of the Standards for Facial Animation Workshop. Technical report, National Science Foundation, University of Pennsylvania, Philadelphia, PA 19104-6389, 1994. Available from http://www.cis.upenn.edu/.

[23] E. Petajan. Automatic lipreading to enhance speech recognition. In *Computer Vision and Pattern Recognition Conference*. IEEE Computer Society, 1985.

[24] F. Pighin, R. Szeliski, D. H. Salesin, "Resynthesizing Facial Animation through 3D Model-Based Tracking", in *Proc. of the ICCV*, 1999.

[25] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computing", Cambridge University Press, second edition, 1992.

[26] L. Reveret, G. Bailly, P. Badin, "MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation", in *Proc. of the 6th Int. Conference of Spoken Language Processing, ICSLP'2000*, Beijing, China, Oct. 16-20, 2000.

[27] S. Sclaroff, J. Isidoro, "Active Blobs", in *Proc. of the ICCV*, pp. 1146-1153, Mumbai, India, January, 1998.

[28] A. Schödl, A. Haro, I. Essa, "Head Tracking Using a Textured Polygonal Model", in *Proc. of the 1998 Workshop on Perceptual User Interfaces*, 1998.

[29] A. Smolic, B. Makai, and T. Sikora. Real-time estimation of long-term 3-d motion parameters for snhc face animation and model-based coding applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(2):255, March 1999.

[30] D. G. Stork and M. E. Hennecke (Eds.). *Speechreading by Humans and Machines*. Models, Systems, and Applications Series: NATO ASI Series, Vol. 0, Springer Books, 1996.

[31] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication, Special Issue on MPEG-4*, vol. 15, pp. 387-421, Jan. 2000.

[32] D. Terzopoulos, K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", in *IEEE Transactions on PAMI*, Vol. 15, no. 6, pp. 569-579, 1993.

[33] H. Tao, H.H. Chen, W. Wu, and T.S. Huang. "Compression of mpeg-4 facial animation parameters for transmission of talking heads". *IEEE Transactions on Circuits and Systems for Video Technology*, 9(2):264, March 1999.

[34] K. Waters and T. Levergood, “An automatic lip-synchronization algorithm for synthetic faces”, *Proc. of the Multimedia Conference*, pages 149-156, San Francisco, California, September 1994. ACM