

# LIP ANIMATION OF A PERSONALIZED FACIAL MODEL FROM AUDITORY SPEECH

*Javier Melenchón, Ignasi Iriondo, J.Claudi Socoró, Elisa Martínez, Lourdes Meler*

La Salle School of Engineering, Ramon Llull University  
Pg. Bonanova 8, 08022 Barcelona, Spain  
{jmelen, iriondo, jclaudi, elisa}@salleURL.edu

## ABSTRACT

This paper proposes a new method for lip animation of personalized facial model from auditory speech. It is based on Bayesian estimation and person specific appearance models (PSFAM). Initially, a video of a speaking person is recorded from which the visual and acoustic features of the speaker and their relationship will be learnt. First, the visual information of the speaker is stored in a color PSFAM by means of a registration algorithm. Second, the auditory features are extracted from the waveform attached to the recorded video sequence. Third, the relationship between the learnt PSFAM and the auditory features of the speaker is represented by Bayesian estimators. Finally, subjective perceptual tests are reported in order to measure the intelligibility of the preliminary results synthesizing isolated words.

## 1. INTRODUCTION

Transmission of multimedia information has increased in the last years over the PSTN (Public switched telephone network) with the introduction of new devices with advanced media capabilities (images, video, ...). Optimal transmission of video content is necessary to avoid congestion collapse until new telecommunications infrastructures are introduced [1].

In the context of talking heads, many works are being done to exploit the correlation between the visual and the auditory information in audio-visual speech production [2]; one of their main results is the prediction and synthesis of the face appearance from acoustic information. Recent advances in this field lead to animate a talking head from speech. However, actual results are not person specific yet and most systems use intrusive tracking systems to extract the facial appearance.

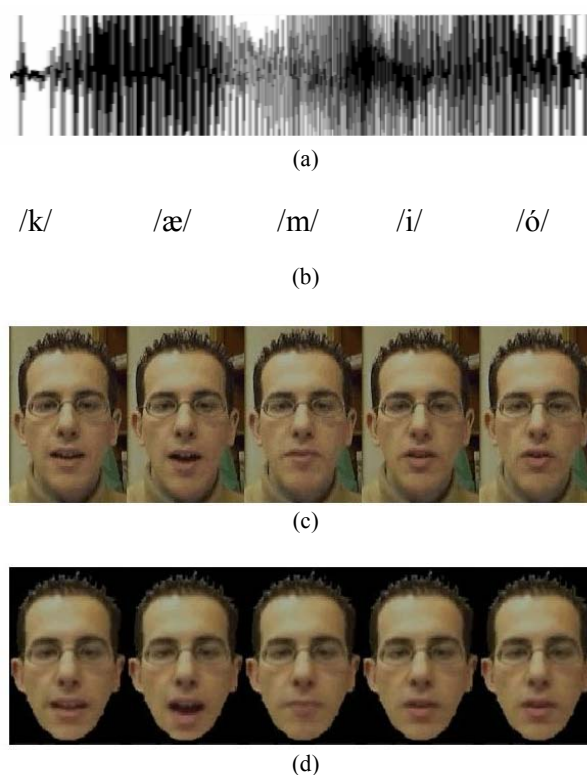


Figure 1. (a) Waveform of the utterance /kæmió/. (b) Phonetic transcription. (c) Real image sequence (d)

Facial animation driven by speech has been tried to be solved using different techniques as artificial neural networks (ANN) [3,4,5,6], Hidden Markov Models [5,7,8,9,10,11,12] (HMM), and direct mappings [13,14,15,16]. In ANN, output visual parameters are derived from acoustic information using multiperceptron layers [3,4,5,6] and vector quantization in [3]. In HMM, a joint audio-visual model is learnt from the user in order to predict visual appearance from auditory information; Hidden Markov Models Inversion is presented in [12],

making HMM more robust to noisy acoustic environments; dynamics are included in [8], enhancing the videorealism of the synthetic sequence. These two methods, ANN and HMM, have a computational expensive cost, mainly in the training step. To overcome this computational drawback, direct mappings are used, although they lose contextual information [13,14,15]. However, the algorithm presented in [16] is based in a bijective correspondence of audio-visual pairs, associating each video frame with the acoustic features from past and future audio frames.

In this work, an alternative method for the prediction of person specific visual appearance from auditory speech is presented and it is based on Bayesian estimators. Visual and auditory information representation is shown in section 2 and their relationship is explained in section 3, as well as the different evaluated estimators. The paper also provides experimental results and subjective tests in section 4. Finally, some concluding remarks about the whole work are done in section 5.

## 2. ANALYSIS

The input for the analysis module consists on an audio-visual sequence of 25 frames/s and a waveform sampling rate of 16KHz. Auditory and visual information will be obtained from this sequence and will be used in section 3 in order to form the probability density functions for the Bayesian estimators.

### 2.1. Auditory information

The auditory information is extracted from the waveform attached to the recorded sequence. This waveform is broken into  $M$  short 50% overlapping frames  $v_i$  of 40ms of audio. L prediction coefficients (LPC) are extracted from each frame  $v_i$  to form each  $\tilde{a}_i = (\tilde{a}_i^1 \ \tilde{a}_i^2 \ \dots \ \tilde{a}_i^L)^T$ . A smoothing operation on each coefficient over all frames is done to reduce gaussian noise, thus obtaining  $\hat{a}_i = (\hat{a}_i^1 \ \hat{a}_i^2 \ \dots \ \hat{a}_i^L)^T$ . Finally, it has been shown that the static relationship between auditory and visual configurations of speech can only account for approximately 65% of the variance in facial motion [17]. To add context information, the audio parameters for the  $i$ -th frame are defined as  $a_i = (\hat{a}_i^T \ \hat{a}_{i-1}^T \ \dots \ \hat{a}_{i-N+1}^T)^T$ , a concatenation of  $N$  successive smoothed vectors of LPC  $\hat{a}_i$ .

### 2.2. Visual information

The visual information is extracted from the recorded image sequence using the registration algorithm presented in [18]. This algorithm takes as input the recorded image sequence and a set of masks and returns a set of orthonormal bases  $B$  (PSFAM) and a matrix of coefficients  $C$  with columns  $c_i$ . Every image  $I_i$  is represented as a linear combination of the columns of  $B$  with  $P$  coefficients  $c_i$ .

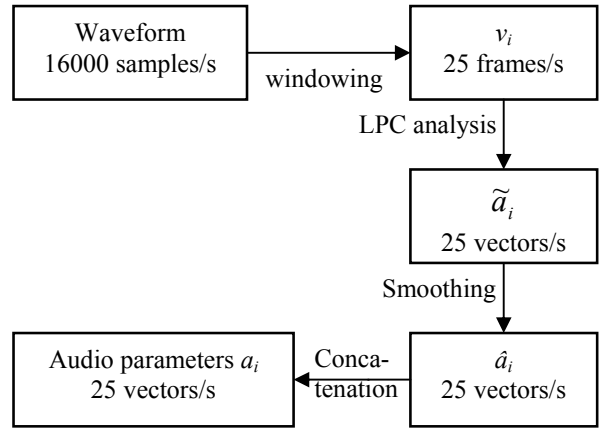


Figure 2. Auditory information analysis

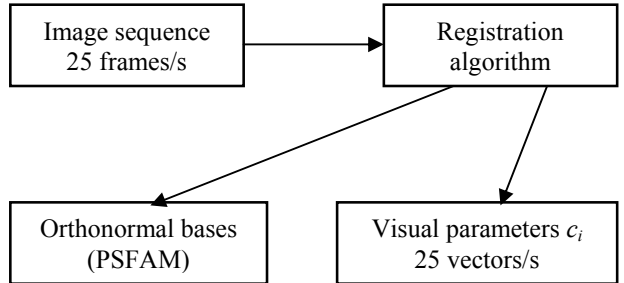


Figure 3. Visual information analysis

## 3. ESTIMATORS

With regard to the relationship between phonemes and the visual appearance of phonemes (visemes), some authors have claimed a many-to-one relationship [19] and others have stood for a many-to-many one [8]. In this work, the first assumption has been taken in order to simplify the estimation process. Thus, a given sound can only have a unique viseme, but a viseme can correspond to different sounds. The problem to be solved is reduced to find an injective application  $f$ :

$$\begin{aligned} f: \mathfrak{R}^{LN} &\rightarrow \mathfrak{R}^P \\ a_i &\rightarrow c_i \end{aligned} \quad (1)$$

In this section, different Bayesian estimators are presented. Besides, linear methods as Kalman estimation were first taken into account, but very poor results were obtained, revealing the non-linearity nature of the problem.

### 3.1. Bayesian estimation

Three different estimators have been considered: maximum likelihood (ML), maximum a posteriori (MAP) and minimum mean squared error (MMSE) [20]. Bayesian estimation methods used in this paper require the definition of viseme classes  $\theta_k$ . These ones are determined by vector quantification of all the columns  $c_i$  of matrix  $C$ , detailed in section 2.2, by means of k-means clustering. Each viseme class  $\theta_k$  has a probability of occurrence  $p(\theta_k)$  and a centroid  $c_k$ , which is the representative viseme of that class. A priori probabilities  $p(a|\theta_k)$ , modeled as Gaussians  $N(a_k, A_k)$ , can be obtained knowing some correspondences between auditory ( $a_i$ ) and visual ( $c_i$ ) parameters. Eight centroids  $c_k$  have been considered given the acceptable output quality of the quantified image sequence. When the priori probabilities are established for each class, these methods will estimate the most probable viseme class  $\theta_j$  for a given vector of audio parameters. This class is associated to the centroid  $\hat{c}_j$ , which corresponds to the mouth appearance that will be synthesized.

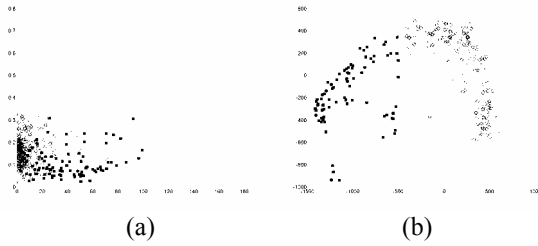


Figure 4. Two phoneme (a) and viseme (b) classes. Viseme classes are obtained by means of vector quantification. Phoneme classes are constructed using viseme classes and auditory and visual parameters correspondences.

### 3.2. ML estimator

The most probable viseme class  $\theta_j$  is given by the maximum a priori probability:

$$\hat{\theta}_j = \arg \max_{\theta_k} \{p(a_j | \theta_k)\} \quad (2)$$

### 3.3. MAP estimator

The most probable viseme class  $\theta_j$  is given by the maximum a posteriori probability:

$$\hat{\theta}_j = \arg \max_{\theta_k} \{p(\theta_k | a_j)\} = \arg \max_{\theta_k} \{p(\theta_k)p(a_j | \theta_k)\} \quad (3)$$

Note that ML method is equivalent to the MAP method assuming equiprobability between classes.

### 3.4. MMSE estimator

The most probable mouth appearance identified by  $\hat{c}_j$  is given by the mean of the a posteriori probabilities of all classes:

$$\begin{aligned} \hat{c}_j &= \sum_{k=1}^L c_k \cdot p(\theta_k | v_j) = \sum_{k=1}^L c_k \cdot \frac{p(\theta_k)p(v_j | \theta_k)}{p(v_j)} = \\ &= \frac{1}{p(v_j)} \sum_{k=1}^L c_k \cdot p(\theta_k)p(v_j | \theta_k) \end{aligned} \quad (4)$$

## 4. EXPERIMENTAL RESULTS

An image sequence of 40 seconds is recorded with a frame rate of 25Hz. In this sequence, the user is asked to speak twenty isolated words. Due to memory limitations, the registration algorithm used in [18] cannot work with image sequences of more than 1000 frames. More work is been done in order to work with longer sequences.

The probability density functions described in section 3 are calculated using the data of this recorded sequence. The auditory information has a context of three audio windows ( $N=3$ ) and 9 LPC's, resulting in 18 element vectors  $a_i$ . The visual information obtained from the registration algorithm detailed in [18] consists on 57 element vectors  $c_i$ . Vector quantification is done using eight centroids, resulting in eight viseme classes.

To evaluate the synthesis of this sequence using the three estimation methods we measure the Euclidean distance between predicted ( $\hat{c}_j$ ) and real ( $c_j$ ) visual parameters.

Best results are obtained with MMSE estimator. An example of the synthesis results can be seen in figure 1 in the utterance of the Catalan word 'camió' with phonetic transcription /kæmió/.

Intelligibility tests have been done with gaussian noise and real office noise with SNR ratios of -12dB and -9dB.

The original sequence helped to understand an additional 24,2% of the whole words, while, the synthesized sequence improved it in a 14,3%. If only the vowel sounds are taken into account, the original sequence achieves a 17,3% and the synthesized one reaches a 13,8%. This fact is due to a poor training set of images, which has not allowed greater number of centroids, which would have modeled sounds in more detail.

## 5. CONCLUDING REMARKS

A new method for lip animation from auditory speech has been proposed. This method learns the specific appearance of the user in a non-intrusive manner by means of the registration algorithm detailed in [18]. In order to map auditory to visual information, vector quantification and different Bayesian estimators are used. Best results are obtained with MMSE estimator. The performance of the synthesized sequence is almost the same (80%) of the real sequence in understanding vowel sounds and a 60% of the real sequence in understanding whole words. More work has to be done in order to increase the length of the training sequence to achieve better quality in synthesis. Future works will include studies like [21] regarding to data-fusion techniques for exploiting the non-gaussian correlation in auditory and visual information.

## 6. REFERENCES

- [1] N. Osifchin, "An infrastructure for telecommunications power in a new era in public networking", TELESCON'00, Dresden, Germany, pp. 23 -28, 7-10 May 2000.
- [2] D.W. Massaro, "Auditory visual speech processing", 7th European Conference on Speech Communication and Technology (Eurospeech'01), Scandinavia, Aalborg, Denmark, September 2001.
- [3] P. Hong, Z. Wen, T.S. Huang, "Real-time Speech Driven Avatar with Constant Short Time Delay", Proc. Int. Conf. on Augmented, Virtual Environments and 3D Imaging, Greece, 2001.
- [4] D.W. Massaro, J. Beskow, M.M. Cohen, C.L. Fry, T. Rodríguez "Picture My Voice: Audio to Visual Synthesis using Artificial Neural Networks", Proc. from AVSP'99, Santa Cruz, USA, 1999,
- [5] E. Agelfors, J. Beskow, B. Granström, M. Lundeberg, G. Salvi, K.E. Spens, T. Öhman, "Synthetic visual speech driven from auditory speech", Proc. AVSP'99, Santa Cruz, USA, 1999.
- [6] F. Lavagetto, "Converting Speech into Lip Movements: A Multimedia Telephone for Hard of Hearing People", IEEE Trans. On Rehabilitation Engineering, Vol. 3, No 1, pp. 90-102, 1995.
- [7] Y. Huang, X. Ding, B. Guo, H.Y. Shum, "Real-time face synthesis driven by voice", Proc. of Computer-Aided Design and Computer Graphics, Kunming, PRC, Aug., 2001.
- [8] M. Brand, "Voice Puppetry", ACM SIGGRAPH, ISBN: 0-201-48560-5, pps 21-28 , August 1999
- [9] K. Kakihara, S. Nakamura, K. Shikano, "Speech-to-face movement sintesis based on HMM", Proc. of ICME'00, New York, 2000.
- [10] E. Yamamoto, S. Nakamura, K. Shikano, "Lip movement sintesis from speech based on Hidden Markov Models", Speech Communication, Vol. 26, pp. 105-115, 1998.
- [11] T. Chen, "Audiovisual Speech Processing: Lip Reading and Lip Synchronization", IEEE Signal Processing Magazine, 2001.
- [12] K. Choi, Y. Luo, J.N. Hwang, "Hidden Markov Model Inversion for Audio-to-Visual Conversion in an MPEG-4 Facial Animation system", Journal of VLSI Signal Processing, Vol. 29, pp. 51-61, 2001.
- [13] T.A. Faruque, A. Kapoor, R. Kate, N. Rajput, L.V. Subramaniam, "Audio driven facial animation for audio-visual reality", Proc. of ICME'01, Japan, 2001.
- [14] S. Morishima, "Multi-modal translation system using face tracking and lip synchronization", Proceedings of the Third International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science and e-Medicine on the Internet, Jan. 2002, L'Aquila Italy
- [15] M.D. Bondy, E.M. Petriu, M.D. Cordea, N.D. Georganas, D.C. Petriu, T.E. Whalen, "Model-based Face and Lip Animation for Interactive Virtual Reality Applications", Proc. ACM Multimedia 2001, Ottawa, Sept. 2001
- [16] A. Exposito, R. Gutierrez-Osuna, P. Kakumanu, O.N. Garcia, "Optimal Data-Encoding for Speech-Driven Facial Animation", WSU Report, WSU-CS-02-04.
- [17] H. Yehia, P.Rubin, E.Vatikiotis-Bateson, "Quantitative Association of Vocal-tract and Facial Behaviour", Speech Communications Vol. 26, No. 1-2, pp. 23-43, 1998.
- [18] J. Melenchón, F. De la Torre, I. Iriondo, F. Alías, E. Martínez, Ll. Vicent, "Text to Visual Synthesis with Appearance Models", Proc. of ICIP'03, Barcelona 2003.
- [19] R. Gutierrez-Osuna, P.K. Kakumanu, A. Exposito, O.N. Garcia, A.Bojorquez, J.L.Castillo, I.J.Rudomin, "Speech-driven facial animation with Realistic Dynamics", Advances in Neural Information Processing Systems, Denver, November, 2000.
- [20] T.K. Moon, W.C. Stirling, "Mathematical Methods and Algorithms for Signal Processing", Prentice-Hall, U.S.A., 2000.
- [21] J.W. Fisher III, T. Darrell, W.T. Freeman, P. Viola, "Learning joint Statistical Models for Audio-Visual Fusion and Segregation", Submitted to IEEE Transactions on Multimedia.