# LIP MOTION MODELING AND SPEECH DRIVEN ESTIMATION

*F. Lavagetto, S. Lepsøy, C. Braccini and S. Curinga*

DIST - Department of Communications, Computer and Systems Science
University of Genova
Via Opera Pia 13, 16145 Genova - Italy

## ABSTRACT

Recent advances in joint acoustical/visual analysis for model-based lip motion synthesis is presented. The 2D lip motion field is modeled as a linear combination of a low dimensional motion basis computed through Principal Component Analysis (PCA). The vector of PCA coefficients is expressed as a function of a limited set of articulatory parameters which describe the external appearance of the mouth. The acoustical processing estimates these articulatory parameters from the direct analysis of the speech waveform based on a neural processing stage, i.e. through a bank of Time Delay Neural Networks. The achieved results have been subjectively evaluated by visualizing the estimated motion on a wire-frame mouth template presented in synchronization with speech. The experiments carried out so far deal with single-speaker trained TDNNs and with single-speaker PCA, but suitable algorithms for generalizing the techniques are currently under investigation.

## 1. INTRODUCTION

Much work in speech and video processing has over the recent years been directed towards the integration of articulatory, acoustical, and perceptual data. In this paper we consider the so called articulatory parameters often used in phonetics to characterize articulation, and a method is proposed to extract them from the speech signal and to drive active shape models for image synthesis.

In normal face-to-face communication, visual and auditive stimuli are integrated, as there exist rules for the correspondence between articulatory gestures, shape of the vocal tract, and the structure of the acoustic speech signal. This integration of stimuli is characteristic of the *bimodal* nature of speech communication.

Bimodal speech processing concerns the processing of speech where visual and acoustical information are utilized jointly. It has produced significant results in a variety of applications like speechreading systems [1], visual speech recognition [2], lip motion synthesis from speech [3].

In the following sections, the adopted lip motion model is presented and the derivation of model parameters from both visual and acoustical analysis is described along with some experimental results.

## 2. LIP SHAPE MODELING

In this section we give a definition of the lip representation that will be used as well as a description of the active shape models. We represent the lip deformation by $K$ vertices in a two-dimensional wireframe structure of polygons. The total set of vertices can be ordered into a vector of vertical ($v$) and horizontal ($h$) coordinates:

$$x = [v_1 h_1 \cdots v_K h_K]^T \in \mathbf{R}^{2K}.$$

We shall refer to such a vector as the *lip shape* vector and to the $2K$-dimensional space as the *shape space*.

Lip deformation will be modeled as the transformation of a template lip shape vector into another

$$T : \mathbf{R}^{2K} \rightarrow \mathbf{R}^{2K},$$

where the transformation itself is a function of several parameters. A lip shape $x$ will be represented by or synthesized as a transformed template lip shape $x_0$:

$$x = T(x_0).$$

As a model of deformations we use the Active Shape Model (ASM) introduced in [4] for the analysis and location of deformable objects.

The ASM models both rigid motion - scaling, rotation and translation - and nonrigid deformation. Nonrigid deformation is modeled with a set of basis vectors in the shape space; to arrive at a deformed lip shape there is added a linear combination of the deformation basis vectors to the template shape. The ASM expresses a deformation of a template shape $x_0$ by scaling $s$, rotation through an angle $\theta$, translation $t$ and nonrigid deformation by adding a vector in a space spanned by certain vectors $\{p_1, \ldots, p_I\}$, to form the shape $\hat{x}$,

$$\hat{x} = T(x_0) = sR(\theta)\left(x_0 + \sum_{i=1}^{I} \gamma_i p_i\right) + t \qquad (1)$$

where $t$ is the displacement vector and $R$ denotes a matrix that rotates every point around the origin by the same angle $\theta$.

The space spanned by the vectors $\{p_1, \ldots, p_I\}$ will hereafter be referred to as the *nonrigid deformation* space. The vectors will be referred to as *deformation modes*. If we order the basis for the nonrigid deformation space as columns in a matrix

$$P = [p_1 \cdots p_I]$$

and the coefficients in a vector

$$g = [\gamma_1 \cdots \gamma_I]^T,$$

we may write Equation 1 as

$$\hat{x} = sR(\theta)\,(x_0 + Pg) + t. \qquad (2)$$

that we call the **approximation expression**.

The nonrigid deformation space is constructed using a training set of lip shape vectors. Supposing that there are $L$ training vectors, the mean shape vector is used as the template shape vector which is subsequently deformed. With reference to Equation 2, $x_0 = \bar{x}$. The basis vectors $\{p_1, \ldots, p_I\}$ are the *principal components* of the deformations in the training set with respect to the mean shape vector.

## 3. FROM ARTICULATORY PARAMETERS TO ASM COEFFICIENTS

In this section the relation between the articulatory parameters and the coefficients for the nonrigid part of the ASM is discussed. The relation will be expressed as a function that maps articulatory parameters to ASM coefficients.

Letting $J$ and $I$ be the number of available articulatory parameters and of ASM nonrigid deformation modes, respectively, the mapping (assumed linear for simplicity) is of the type

$$F : \mathbf{R}^J \rightarrow \mathbf{R}^I$$

and it will be referred to as the *parameter transform*.

The elements of the matrix $F$ will be computed using the same set of shape vectors as for constructing the nonrigid deformation basis. In addition, to each shape vector $x_l$ there will be assigned a vector $a_l$ of articulatory parameters, forming the **training set**

$$\{(x_l, a_l) : l = 1, \ldots, L\}.$$

The training set also implies coefficient vectors that correspond to the shape vectors adjusted to render unnecessary the rigid part of the ASM,

$$g_l = P^T(x_l - \bar{x}),$$

i.e. they are found by projecting the deviation from the mean shape vector onto the space for nonrigid deformation.

Now the criterion function for constructing the parameter transform can be derived. The coefficient vector $g$ will be approximated by a function of the articulatory parameter vector $a$. The deformations are applied to the mean shape vector $\bar{x}$ to which is associated the parameter vector $\bar{a}$, and in the case $a = \bar{a}$ the coefficients vector should be zero. This is achieved if the argument to the linear parameter transform is $a - \bar{a}$. Hence, we have

$$g \approx F(a - \bar{a}),$$

and we can rewrite the approximation expression to obtain the **synthesis expression**

$$\tilde{x} = sR(\theta)\,(\bar{x} + PF(a - \bar{a})) + t. \qquad (3)$$

The goal is to find a parameter transform $F$ that minimizes the expected synthesis error

$$J(F) = \mathcal{E}\left\{\|x - \tilde{x}\|^2\right\} \qquad (4)$$

It can be demonstrated [5] that when $J(F)$ is estimated over the training set, it can be minimized by the parameter transform:

$$F = \left((U^T U)^{-1} U^T G\right)^T$$

where

$$U = \begin{bmatrix} (a_1 - \bar{a})^T \\ \vdots \\ (a_L - \bar{a})^T \end{bmatrix}, \quad G = \begin{bmatrix} g_1{}^T \\ \vdots \\ g_L{}^T \end{bmatrix}.$$

## 4. ESTIMATION OF THE ARTICULATORY PARAMETERS FROM SPEECH

The ASM coefficients have been computed from articulatory parameters estimated directly from the speech source signal exploiting the intrisic audio-visual correlation.

The speech signal has been sampled at 8 KHz and quantized linearly at 16 bits before being processed through the following steps:

- spectral preemphasis;
- segmentation into non overlapped frames of duration T = 20 ms;
- linear predictive analysis of 10-th order;
- power estimation and computation of the first 12 cepstrum coefficients;
- normalization of the cepstrum coefficients to the range [-1, 1].

The normalized cepstrum vectors are then input to the conversion system based on a bank of TDNNs, each of them trained to provide estimates of a specific articulatory parameter.

In order to understand how many articulatory parameters are necessary to allow a faithful visual synthesis of speech, a very large audio/video database has been collected.

To simplify the extraction of the articulatory parameters from the video frames, the speaker's face was conditioned by means of lipstick and white markers placed in correspondence to the tip of the nose and of the chin. The mouth model which has been employed, sketched in Fig. 1, is defined by a vector of parameters extracted from the frontal view and described in Table 1.

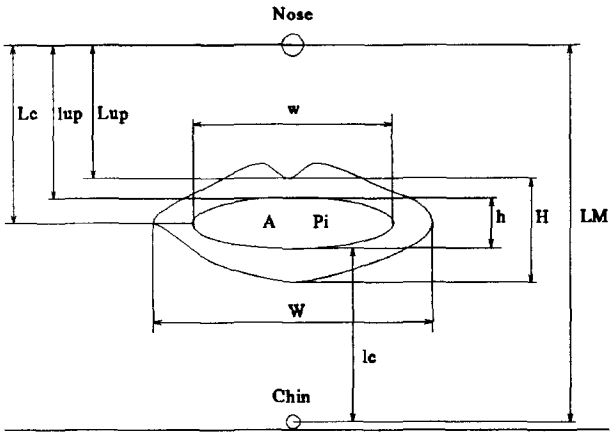


Figure 1: Articulatory parameters extracted from the frontal view

The cross-correlation between these parameters has been estimated in order to identify interdependencies and to provide a basis of independent parameters. From the experimental outcomes it has been decided to use a basis of 9 parameters (W, h, w, Lup, Lc, lup, lc, A, Pe).

A distinct Time-Delay Neural Network [6] has been trained for each independent articulatory parameter using as input the 12-dimensional vectors of the cepstrum coefficients computed from the acoustic corpus. The pattern-target comparison has been done introducing a time delay between them in order to model

| Name | Description |
|---|---|
| H | Vertical opening of outer contours |
| W | Horizontal opening of outer contours |
| h | Vertical opening of inner contours |
| w | Horizontal opening of inner contour |
| Lup | Vertical distance from the nose to the outer contour |
| Lc | Vertical distance from the nose to the corner of the mouth |
| lup | Vertical distance from the nose to the inner contour |
| lc | Vertical distance from the inner contour to the chin marker |
| LM | Vertical distance from nose to the chin marker |
| A | Area of the inner contour |
| Pe | Perimeter of the outer contour |
| Pi | Perimeter of the inner contour |

Table 1: Description of the articulatory parameters used to model the mouth motion

the forward coarticulatory effect: this means that the current output of the network is made as similar as possible to the articulatory parameter corresponding to the acoustic unit displaced few frames back in the past.

As reported in [6], the convergence of the network has been studied with reference to different choices of its parameters, namely the number of neurons, the number of hidden layers, the TDNN memory, the input acoustic representation (spectrum vs cepstrum) and the pattern-target delay. The best results have been obtained experimentally using networks with two hidden layers, composed of 8 and 3 units each, whose memory size is fixed to $D(1) = 2$ (first hidden layer), $D(2) = 3$ (second hidden layer) and $D(3) = 4$ (output layer). The pattern-target delay has been chosen equal to 5 acoustic frames.

## 5. EXPERIMENTAL RESULTS

The methods proposed in the previous sections have been applied to estimate and synthesize the lip movements of a sample speaker, i.e. a young English-speaking woman uttering nonsense consonant-vowel-consonant words from a vocabulary developed to be phonetically balanced. [1] The recorded video corpus is composed of RGB colour images of 352 × 288 pixels. The total sequence takes 100 seconds, and with 20 frames per second there are 2000 frames. The sequence

[1] The vocabulary was developed by the National Association for the Deaf, Ireland.

is divided into 1100 frames (55 seconds) for training and 900 frames (45 seconds) for test.

The linearity assumption for the parameter transform, made in section 3, was tested by comparing the synthesized coefficients with those corresponding to the real shape vectors in the test set onto the nonrigid deformation space.

The coefficients corresponding to lip rounding and horizontal opening were well synthesized. These two modes account for 83% of the total energy (sum of squared norms). The third coefficient was rather badly approximated, which means that one of the modes for vertical opening did suffer. This is evidence that the linearity assumption does not hold for this particular deformation mode.

Figure 2: Corresponding synthetic (left) and real (right) mouth shapes.

A first reason for this is that a linear transformation may be a too simple model for the relation between the articulatory parameters and the ASM coefficients. There may also have been a missing correspondence between measured parameters and extracted contours in our own data. A deeper discussion of this matter can be found in [5]. The texture for the lip region in one of the images was acquired and used for mapping onto wireframe polygons. The wireframe vertices were animated using the first six deformation modes controlled by the 9 articulatory parameters chosen from those in Table 1. The parameters were taken from a subsequence of the test sequence. Fig. 2 shows some examples of the obtained synthetic lips (left) against the original shapes (right).

## 6. CONCLUSIONS

A method has been described that estimates the articulatory parameters directly from the speech signal and converts them into the synthesis coefficients of a non rigid lip model. The articulatory parameters are estimated by means of a suitable analysis of the speech signal performed by a bank of TDNNs. The ASM model allows also the representation of non rigid lip motion, for which a MSE optimal linear mapping from the articulatory parameters to the ASM coefficients has been devised. The experimental results performed so far have confirmed that the approach is more than promising even from a perceptual viewpoint.

## REFERENCES


1. Hennecke, M.E., Stork, D.G., Prasad, K.V., (1995) Visionary Speech: Looking ahead to Practical Speechreading Systems, in "Speechreading by Humans and Machines", D.G. Stork and M.E. Hennecke eds., Springer Verlag, Berlin, pp. 331-349.
2. Luettin, J., Thacker, N.A., Beet, S.W., (1996) Statistical Lip Modelling for Visual Speech Recognition, Proc. 8th EUSIPCO Conf., Sept. 10-13, Trieste, Italy.
3. Curinga, S., Lavagetto, F., Vignoli, F., (1996) Lip Movements Synthesis using Time Delay Neural Networks, Proc. 8th EUSIPCO Conf., Sept. 10-13, Trieste, Italy.
4. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., (1995) Active Shape Models - their Training and Application, in Computer Vision and Image Understanding, Vol. 61, N. 1, pp. 38-59.
5. Lepsoy, S., Curinga, S., (1985) Conversion of Articulatory Parameters into Active Shape Model Coefficients for Lip Motion Representation and Synthesis (submitted to Signal Processing: Image Communication).
6. Lavagetto, F., (1995) Converting Speech into Lip Movements: A Multimedia Telephone for Hard of Hearing People, IEEE Trans. on Rehabilitation Eng., Vol. 3, N. 1, pp. 90-102.