FROM RAW IMAGES OF THE LIPS TO ARTICULATORY PARAMETERS : A VISEME-BASED PREDICTION

L. Revéret
Institut de la Communication Parlée
Université Stendhal / INPG, BP25 38040 Cedex 9 Grenoble, France.
Tel. +33 4 76 82 41 28, FAX: +33 4 76 82 43 35, E-mail: reveret@icp.grenet.fr

ABSTRACT

This paper presents a method for the extraction of articulatory parameters from direct processing of raw images of the lips. The system architecture is made of three independent parts. First, a new greyscale mouth image is centred and downsampled. Second, the image is aligned and projected onto a basis of artificial images. These images are the eigenvectors computed from a PCA applied on a set of 23 reference lip shapes. Then, a multilinear interpolation predicts articulatory parameters from the image projection coefficients onto the eigenvectors. In addition, the projection coefficients and the predicted parameters were evaluated by an HMM-based visual speech recogniser. Recognition scores obtained with our method are compared to reference scores and discussed.

1. INTRODUCTION

There are two main approaches to the processing of mouth images in automatic lipreading. The stochastic approach makes wide use of learning techniques providing image features poorly interpretable. The articulatory approach aims at measuring as accurately as possible anatomical and/or geometrical parameters which can be interpreted in phonetic terms. Although the method here proposed involves image processing techniques generally used in the stochastic approach, it is articulatory-oriented indeed. It gives some description of a mouth shape in phonetic terms with regard to phonetically labelled visemes [2]. It also gives a reliable evaluation of geometric parameters of the lips that could not be automatically measured on natural lips without prior make-up.

2. THE DIFFERENT APPROACHES TO AUTOMATIC LIPREADING

There are two main classes of systems: those model-based and those image-based.
In the model-based systems, a geometrical model of the lip contours (external and/or internal) is applied directly to the input image of the speaker's lips. Splines [5] and polynomial equations [11] are commonly used for liptracking.

In the image-based approach, the whole set of image pixels is processed. Various techniques have been used: Colour transformation of the image texture in order to extract lips area [6, 9]; optical flow analysis [8], etc.

Both model-based and image-based methods can benefit from information reduction through multidimensional analysis such as Principle Component Analysis (PCA). In the model-based approach, the model deformation may be limited to the main variation modes of its control parameters (Active Shape Models [7], B-spline model [5]). In the image-based approach, PCA has shown its powerfulness to dramatically decrease image information since the pioneering works by [10] in face recognition (eigenfaces), later applied to raw images of the mouth [3, 4].

3. OUR VISEME-BASED APPROACH

Working on a corpus consisting of 786 repetitions of sentences "c'est pas $V_1C_1V_2C_1V_1z$", [2] identified 23 classes of lip shapes as representatives of expressionless speech production by one male speaker of French (Figure 1). In this corpus, the lips of the speaker were carefully made-up in blue so that several articulatory parameters could be accurately measured by a chromakey system [6] on a set of phonetically labelled images from front and profile.

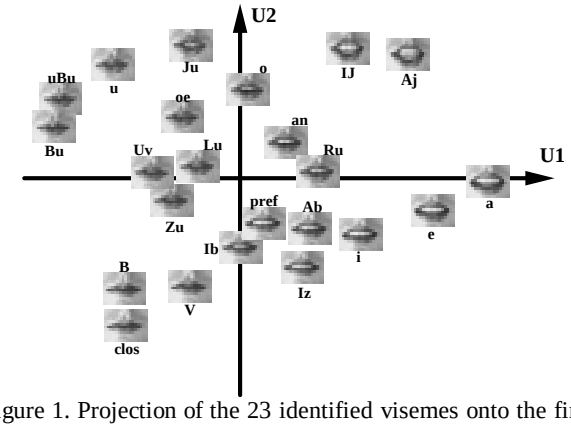

Figure 1. Projection of the 23 identified visemes onto the first two components of a PCA performed on lip parameters.

The identification of the 23 visemes results from a set of multidimensional analyses (clustering, correspondence, and discrimination) performed on the articulatory parameters measured. These 23 visemes include

coarticulated vowels and consonants in various phonetic contexts. They can be interpreted as an optimal coverage of the labial space, as defined by geometrical lip parameters. Of course, there is no unique link between such a viseme and a phoneme which is only defined as a phonological symbol. For instance, the /a/ uttered in a /z/ context is classified into the so-called viseme [i] that also includes a steady /i/.

4. IMAGE PROJECTION

From the 23 visemes identified, 23 representative greyscale images centred on the lips have been selected and downsampled to 32x24=768 pixels. As the viseme images came from different sequences, they have been spatially aligned to the reference image of the prephonatory viseme to suppress head movements. For this, a gradient descent algorithm in X and Y was used to optimise the correlations between the reference viseme and the others. The 23 adjusted images provided a first basis of $\Re^{768}$ vectors on which any new lip image could be projected and thus described by only 23 coordinates. Though a set of 23 images is already a limited training set compared to similar image-based approaches [3], a PCA applied to these 23 images showed that some redundancy still exists: The first four (resp. eleven) eigenvectors account for 80% (resp. 95%) of the total variance. We call eigenvisemes the 22 artificial images corresponding to the resulting eigenvectors. Any subset of these eigenvisemes provide a new vector basis to perform image projection on an orthonormal image space of lower dimension.

Some similarities in structure were observed between the space defined by the eigenvectors from the PCA applied on the *lip parameters* measured on the visemes and the space defined by the eigenvectors from the PCA applied on the *images* of the visemes. The correlation between visemes coordinates in the "parameter space" and in the "image space" along the first (resp. the second) eigenvector is r=.91 (resp. r=.87).

Figure 2 shows the image representation of the first two eigenvisemes. An articulatory interpretation of these eigenvisemes can be proposed. The first eigenviseme can be considered as an opposition between a closed shape and an open one. The second eigenviseme may be seen as an opposition between a spread shape and a rounding one. This is confirmed by the distribution of the 23 visemes projected on the two first eigenvectors. Due to the correlations mentioned above, the projections are similar in the images space and lip parameters space. The first axis opposes /Aj/, /a/, /i/ (open) to /B/, /Bu/, /V/ (closed) for example (see Figure 1). The second one opposes /i/, /Iz/, /Ib/ (spread) to /Ju/, /o/, /Aj/ (rounded).

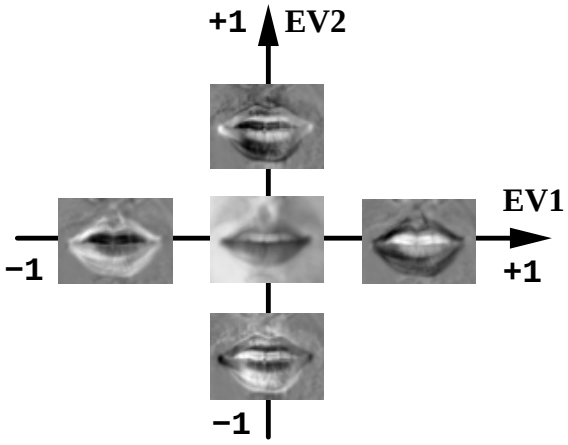

Figure 2. The first two eigenvisemes. The image at the center corresponds to the average image of the 23 visemes images. Along the eigenvectors, darker areas represent a decrease in luminance (lips vs. skin and teeth), and vice-versa.

Another interpretation of the eigenvectors can be made from those calculated with the PCA applied on the lip parameters (Figure 3). An increase of both height and width parameters can be seen along the first eigenvector, whereas the outer height increases and the outer width decreases along the second eigenvector.

	A	B	S	A'	B'
U1	**0.37**	**0.37**	0.39	0.16	0.33
U2	0.07	0.21	0.14	**–0.45**	**0.31**

Figure 3. The first two eigenvectors in lip parameters space

These similarities allow us to hypothesize that our lip parameters could be predicted from the coordinates of any lip image along the 22 eigenvisemes, or a subset thereof.

5. PARAMETER PREDICTION

We have implemented a multilinear interpolation to predict the geometrical lip parameters from the coordinates of an image along the image space defined by the eigenvisemes. Prior to that, the geometrical parameters must be accurately measured on the 23 images of the visemes to serve as interpolation points.

In the general case of a multilinear interpolation, the prediction of a vector Y (e.g., the 11 lip parameters) from a vector X (e.g., a 32x24 image) can be formulated as

$$Y = \Sigma_{i=1...m}\, \varphi_i(X)\, W_i = W\, \Phi(X)$$

where W_i are unknown vector coefficients and $(\varphi_i)_{i=1...m}$ is a basis of m functions (here we use the m projection onto the first m eigenvisemes). From the m coordinates of the 23 visemes on the m eigenvisemes and their 23 associated sets of measured lip parameters, the values of the 11 parameters allow to define W.

$$P = W\,\Phi(V) = W\,(E^t V) \Rightarrow W = P\,(E^t V)^+$$

where P is the 11x23 matrix of eleven lip parameters measured on the 23 visemes, V is the 768x23 matrix of the greyscale images of the visemes, E is the 768x(m) matrix of the greyscale images of m eigenvisemes, and $(E^t V)^+$ is the pseudo-inverse of $(E^t V)$.

Before projection, a new greyscale image is centred on the mean image of the 23 visemes by subtracting the average vector. It is then aligned by minimising the difference between the original image and its reconstruction from the basis [10] to reduce the loss of information introduced by the projection onto the eigenvisemes.

Finally, the vector of parameters Y can be estimated from an image X, as

$$Y = W\,(E^t X) = P\,(E^t V)^+\,(E^t X)$$

6. EVALUATION

To evaluate our system, we used the same corpus as [2] where the speaker's lips were made up in blue. The corpus was made of 486 sentences corresponding to 9 repetitions of 54 sentences. This allowed us to compare the accurate measurement of the chromakey system and the prediction from our system, even though our system is greyscale based and thus requires no special make-up. An example of parameter prediction is shown in Figure 4. Inner contour area and outer contour width prediction are compared to their actual value accurately measured with a chromakey technique along an /azyzaz/ sequence of 760 ms (38 frames).

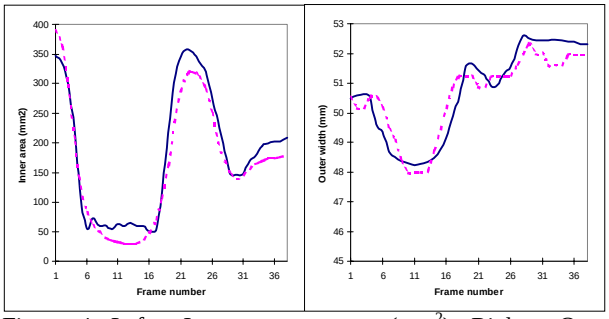

Figure 4. Left : Inner contour area (mm^2); Right : Outer contour width (mm); measured parameter is in dot line, predicted in full line.

The capacity of our parameters prediction to render relevant speech information was tested by an HMM-based automatic lipreader [1]. Four training/test conditions were tested: CK/CK, CK/PP, PP/PP, PC/PC, where CK stands for "chromakey-measured" parameters, PC for principal component on images, and PP for predicted parameters. A jack-knife technique was used to increase the number of training/test conditions with 7 repetitions in the training set and 2 repetitions in the testing set. Five permutations were finally tested in each experimental condition.

6.1. Results in the CK/CK condition

The purpose of this test was to obtain reference scores from the accurate measurement system to serve as a baseline for the others. Seven frontal parameters were used in this condition: Inner height, width and area; outer height and width; upper and lower lip area. Over the 5 permutations, the mean score was 77.8 % and ranged between 72.% and 82.4 %.

6.2. Results in the PC/PC condition

This test evaluates to which extent projections on eigenvisemes are representative of speech variation (Figure 5).

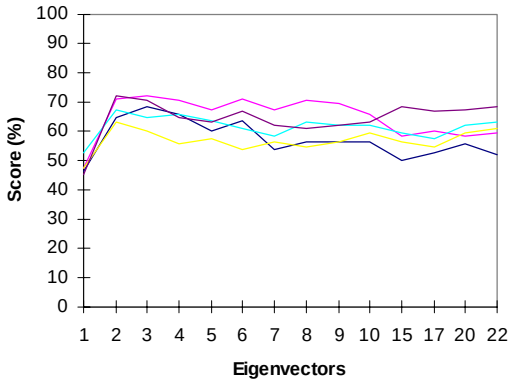

Figure 5. Recognition scores in the PC/PC condition. The X-axis shows the number of first eigenvectors used.

The first two eigenvisemes alone perform 67.8% of correct recognition on average. It is the highest mean recognition score. The first PC alone leads to only 50% correct identification, and more than two eigenvectors tends to lightly decrease the performance too. This result shows that most of speech information is accounted for by the first two eigenvectors only.

6.3. Results in the PP/PP condition

This test presents the results obtained with seven frontal parameters (same as the measured parameters used in § 6.1) predicted by our method from different subsets of the first eigenvisemes (Figure 6).

Results are highly similar to those obtained in the previous test, which is not surprising since the same information is only presented differently, up to a linear combination (see § 5).

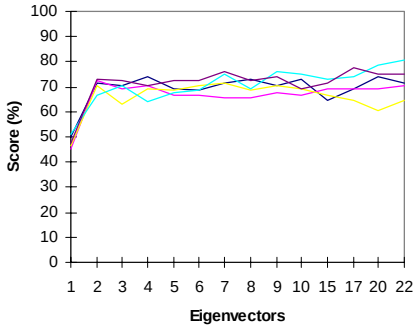

Figure 6. Test and training in PP/PP condition

6.4. Results in the CK/PP condition

This test evaluates the capacity of our system to predict accurately the value of the seven above mentioned parameters from raw images. In this case, the training set consisted of accurately measured parameters whereas the test set consisted of predicted parameters. Recognition scores are presented on Figure 7, depending on the number of eigenvectors used to predict the parameters in the test set.

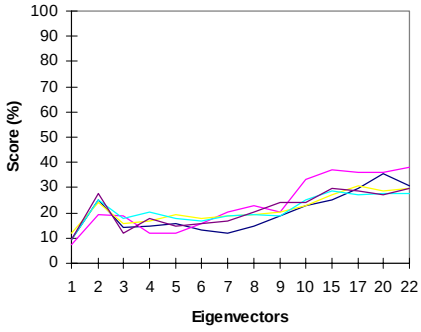

Figure 7. Test and training in CK/PP condition

Most of the errors come from confusions between /a/ and /i/, /b/ and /v/, and /l/ and /r/. This could be explained by the very low resolution that we used for the images (32x24). At this resolution, small variation of opening does not clearly appear.

Although clearly lower than the scores obtained in §6.1, the scores here obtained are well above chance level. A third of the 54 words are correctly recognised when all 22 eigenvectors are taken into account. Except for the noticeable peak observed when the first two eigenvisemes only are considered, performance continuously improves as the number of eigenvectors used increases. This last observation should be interpreted together with the evolution of the PC/PC scores presented in §6.2: The third and following eigenvisemes are probably not necessary to account for the variability associated with the "speech" information contained in our lip images. However, the two techniques used are not strictly comparable, since image processing stores an important information related to the visibility of the teeth and of the tongue which is not processed in the reference "chroma-key" technique.

7. CONCLUSION

We have presented in this paper a complete method to predict articulatory parameters from raw images. The evaluation tests tends to show that speech variation is mostly contained in the first two eigenvectors of a PCA applied to reference images.
Future work will investigate further how to best code visual speech variation. Another major improvement will be to allow the system automatically select an appropriate set of visemes.

8. ACKNOWLEDGMENTS

The author would like to thank Christian Benoît and Ali Adjoudani for their contribution to this work.

REFERENCES

[1] Adjoudani, A., Benoît, C., "On the Integration of Auditory and Visual Parameters in an HMM-based ASR", in *Speechreading by Humans and Machines*, D. Stork and M. Hennecke (eds.), Springer-Verlag, Berlin, 1996, pp. 461-471.

[2] Benoît, C., Lallouache, M.T., Abry, C., "A set of French visemes for visual speech synthesis", in *Talking Machines : Theories, Models and Designs*, G. Bailly and C. Benoit, Eds., Elsevier Science Publishers, 1992, pp. 485-504.

[3] Bregler, C., Konig, Y., "Eigenlips for Robust Speech Recognition", in Proc. *ICASSP'94*, Adelaide, 1994, pp. 669-672.

[4] Brooke , N.M. , Scott , S.D., "PCA Image Coding Schemes and Visual Speech Intelligibility", in *Proc. of the Institute of Acoustics*, Autumn Meeting, Windermere, UK, 16(5), pp. 123-129, 1994.

[5] Kaucic, R. , Dalton, B. , Blake, "Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications", in Proc. *ECCV*, pp. 376-387, Cambridge, UK, 1996.

[6] Lallouache, M.T., "Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres", PhD. dissertation, INPG, Grenoble, France, 1991.

[7] Luettin, J., Thacker, N. A., Beet, S. W., "Speechreading using Shape and Intensity Information", in Proc. *4th ICSLP Conference*, Philadelphia, PA, USA, 1996.

[8] Mase, K. , Pentland, A., "Automatic Lipreading by Optical-Flow Analysis", *in Systems and Computers in Japan*, vol. 22, no. 66, pp. 67-75, 1991.

[9] Petajan, E. , Graf, H. P., "Robust Face Feature Analysis for Automatic Speachreading and Character Animation", in *Speechreading by Man and Machine*, D. Stork and M. Hennecke Eds., Springer-Verlag, Berlin, pp. 425-436, 1996.

[10] Turk, M., Pentland, A., "Eigenfaces for Recognition", *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[11] Yuille, A.L., Hallinan, P.W., Cohen, D.S., "Feature Extraction from Faces using Deformable Templates", *Int. J Computer Vision*, vol. 8, no. 2, pp. 99-111, 1992.