Estimation of Speech Lip Features from Discrete Cosinus Transform

Zuheng Ming, Denis Beautemps, Gang Feng, Sébastien Schmerber

Estimation of Speech Lip Features from Discrete Cosinus Transform

Zuheng Ming[2], *Denis Beautemps*[1], *Gang Feng*[1], *Sébastien Schmerber*[2]

[1]Department Parole & Cognition of GIPSA-lab, CNRS UMR 5216, 961 rue de la Houille Blanche, Domaine Universitaire - BP 46, 38402 Saint Martin d'Hères Cedex, France
[2]Laboratoire Langage et Audition, Service ORL, CHU Michallon – BP 217, 38043 Grenoble Cedex 09, France

{ZMing,SSchmerber}@chu-grenoble.fr;{Denis.Beautemps,Gang.Feng}@gipsa-lab.grenoble-inp.fr

Abstract

This study is a contribution to the field of visual speech processing. It focuses on the automatic extraction of Speech lip features from natural lips. The method is based on the direct prediction of these features from predictors derived from an adequate transformation of the pixels of the lip region of interest. The transformation is made of a 2-D Discrete Cosine Transform combined with a Principal Component Analysis applied to a subset of the DCT coefficients corresponding to about 1% of the total DCTs. The results show the possibility to estimate the geometric lip features with a good accuracy (a root mean square of 1 to 1.4 mm for the lip aperture and the lip width) using a reduce set of predictors derived from the PCA.

Index Terms: Visual Speech, lip-reading, lip geometric feature, estimation, Discrete Cosine Transform.

1. Introduction

The benefit of visual information for speech perception (called "lip-reading") is widely spread. Indeed, from the precursory works of Sumby and Pollack [1] (back to 1954), then those of Summerfield (Summerfield, 1979 [2]; Summerfield et al., 1989 [3]) to those of Benoit et al. (1992) [4] as far as the French language is concerned, it is a well established fact that the visual information from the speaker's face is used to enhance speech perception under noisy environment. But, even in the context of clear auditory speech, vision remains important: shadowing experiments have shown, for instance, that reaction times were lowered by an average factor of 7.5% in case of audiovisual stimuli in comparison with the simple audio presentation (Reisberg et al., 1987 [5]). The well known "McGurk effect" demonstrates the ability to integrate auditory and visual information even if the two modalities are not congruent (McGurk and MacDonald, 1976 [6]; MacDonald and McGurk, 1978 [7]). As we have exposed here, normal-hearing people have competences in lip-reading (Cotton, 1935 [8]; Dodd, 1977 [9]). However, it has been shown that the initial performances – i.e. without specific training - vary greatly from one individual to another. Bernstein et al. (2000) [10] have compared the performances of 96 normal-hearing people with 72 profoundly deaf people. The authors have observed very variable performances between the individuals of both groups, but have clearly showed that the best lip readers were deaf people.

The access to communication technologies has become essential for the handicapped people. The TELMA project (Phone for deaf people, Beautemps et al., 2007 [11]) aims at developing an automatic translation system of visual speech to speech sound and inversely. This project would make possible the communication between deaf users and normal-hearing people through the help of the autonomous terminal TELMA. In the chain of translation of visual speech to speech sound, the automatic processing of the lips requires the extraction of visual features that carry the phonetic information. For this two approaches are generally considered, as suggested by Potamianos (2001 [12], 2003 [13]): shape based features and appearance based features. In the first one, the inner and outer lip contours are extracted from the image view of the face. A lip contour model can be obtained statistically (Luettin et al., 1996 [14]; Dupont and Luettin, 2000 [15]) or parametrically (Hennecke et al. 1996 [16]; Chiou and Hwang, 1997 [17]). Then, the set of model parameters contains the visual information. In the second approach, appropriate transformations, such as discrete cosine transform (DCT) or principal component analysis (PCA), are applied to the pixels of the image corresponding to the speaker's mouth region of interest (ROI) (Gray et al. 1997 [18]; Matthews et al. 1996 [19]). Both approaches are often combined as in Matthews (1998) [20], where active appearance models are built both on shape and appearance features. In this paper, we present a very innovative method that consists to estimate directly speech lip geometric features of the shape approach by a concentrated information of the lips using an appearance based approach. In the following, we first present the experimental set-up and the lip material then the image processing and the modelling before its optimization and evaluation. A general discussion will end this contribution.

2. Experimental set-up and lip material

The data have been derived from a video recording of a speaker pronouncing in French a set of 50 isolated words. The words were made of 31 digits, 12 months and 7 days. Each word was presented once on a monitor placed in front of the speaker, in a random order. The speaker is a female native speaker of French. The recording has been made in a sound-proof booth and the image video recording rate has been set on 50 frames/second. The speaker was seated in front of a microphone and a camera connected to a betacam recorder. A blue mark was placed between eyebrows as a reference point for further lip region location. In addition, a square paper was recorded for further pixel-to-centimeter conversion.

The video recording has been done with the PAL format, thus saved as numerical Bitmap RGB images made of the interlaced half-frames of the video (respectively even and odd lines). Each image was de-interlaced and the missing lines of the half-frames were filled by linear interpolation, as to obtain two de-interlaced full frames corresponding to two recordings separated with 20 ms.

These frames constitute the set of images at the rate of 50 Hz that we will refer to in the following. For its part, the audio of the recording was digitalized at 22,050 kHz. It has been used to segment temporally the words. For each word, the

coordinates of the inner and outer contours of the lips have been manually selected on the corresponding images and converted into centimeters with the use of the pixel-to-centimeter conversion formula. Finally, the following geometric lip features were derived (following Lallouache, 1991 [21]): the lip width (A, Aext), the lip aperture (B, Bext) and the lip area (S, Sext) respectively for the inner and outer contours. The whole process led to a database made of a set of 1570 (A, B, S, Aext, Bext, Sext) sextuplets and the corresponding set of images containing the face of the speaker, at the rate of 50 Hz.

3. Image Processing and Modeling

This section presents the different steps needed to build the model that allows the prediction of the six geometric lip features by an appropriate transformation of the image in the lip region. The transformation is based on a 2-D Discrete Cosine Transformation of the lip region pixels.

3.1. Detection of the Lip Region of Interest

The lip ROI has been defined as a 100 by 100 pixels square in reference to the blue mark placed between the speaker eyebrows. With a color segmentation method the blue mark is easily detected and its center is used to refer the lip ROI (see Figure 1).

Figure 1: *Image of the speaker (left) and the associated lip ROI (right). The landmarks on the hand are used for further Cued Speech processing but not considered in this modelling.*

3.2. The 2D Discrete Cosine Transform (DCT)

The lip ROI obtained from the lip detection step was converted into a 100 × 100 grayscale intensity image matrix. The DCT was chosen as the image transform instead of other transform methods because of excellent energy compaction for highly correlated images and widely used in the image compression techniques. Also this method allows staying in the Real Numbers Domain. The 2-D DCT was used in this work.

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)p}{2M} \cos\frac{\pi(2n+1)q}{2N},$$
$$0 \le p \le M-1, 0 \le q \le N-1, \quad (1)$$

Where

$$\alpha_p = \begin{cases} 1/\sqrt{M}, p = 0 \\ \sqrt{2/M}, 1 \le p \le M-1 \end{cases} \quad \alpha_q = \begin{cases} 1/\sqrt{N}, q = 0 \\ \sqrt{2/N}, 1 \le q \le N-1 \end{cases}$$

The matrix Amn (M=100,N=100) is the set of pixels in grayscale contained in the lip ROI and Bpq is the resulting transformed coefficients matrix. In this work, no specific truncation window is used.

3.3. Using a Mask

At this stage, we need to recall that the aim of the work is to predict the high level geometric features by a limited set of coefficients that concentrate the main lip information. The DCT for its part packs energy in the low frequency regions, so some of the high frequency content can be discarded without significant quality degradation. Therefore we used a mask to select the most significant coefficients located in the top left of the DCT matrix in order to reduce the dimensionality. We tested a triangle and a square mask both to compare the results. Preliminary trials showed that the best results were obtained with masks containing about 100 DCT coefficients, i.e. only 1% of the total number of DCT coefficients (100x100). The optimal size and the shape of the mask will be discussed in the last section, with respect to the precision of the prediction.

3.4. The multi-linear modeling

This subsection establishes the linear relation between the set of DCT coefficients included inside a mask and the six geometric lip features. For this, we used a subset of 811 elements of the database.

Given the matrix $D = [D_1, D_2, \ldots, D_M]$, in which M is the size of the mask. Each line contains the M DCT coefficients of the mask region for a frame instant. The number of line is 811.

Considering that the matrix D contains large image information and the generality of the linear fitting, we could use the linear combinations of vectors D_i to estimate the geometric lip features, by example for B:

$$\hat{B} = f(D_1, D_2, ..., D_p) \qquad 1 \le p \le M$$

But in order to have a set of predictors with an order largely less than M, it is important to analyze the order of importance of each predictor. For this aim, we applied a PCA on the DCTs that finally decorellates the initial DCTs. The first principal component accounts for the maximum variance of the DCT data, and each succeeding component accounts for the maximum of the residual variance. The set of F factors of the PCA are the projections of the DCTs on the principal axes. A subset of these factors have been used to predict B efficiently:

$$\hat{B} = f(F_1, F_2, ..., F_p) \qquad 1 \le p \le M \quad (2)$$

In addition, the geometric lip features can be estimated step by step since the orthogonality property of the F factors.

Thus for B, we obtained (Figure 2):

$$\hat{B}_1 = k_1 F_1 + \bar{B} \quad (3)$$

where k_1 is the linear fitting coefficient in the least square sense and $\bar{B}$ is the mean value of B. We defined $r_1 = B - \hat{B}_1$ as the residual error of the first estimation, and we used F_2 to estimate r_1, (Figure 3):

$$\hat{r}_1 = k_2 F_2 \quad (4)$$

The estimation continues following the same procedure until the order p of prediction is attained. Finally, the estimated B values can be expressed as following:

$$\hat{B} = k_1 F_1 + k_2 F_2 + ... + k_p F_p + \bar{B} \quad (5)$$

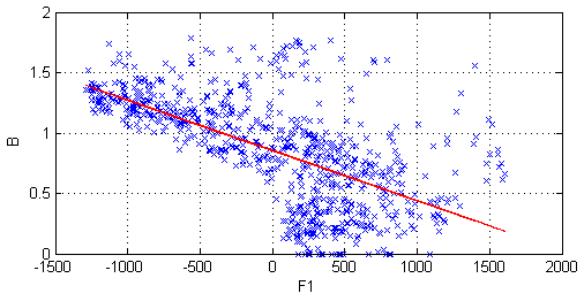

Figure 2: *B values (crosses) and its estimation (straight line) in function of* F_1 *(with the use of a right-angled triangle of side 15 as the mask).*

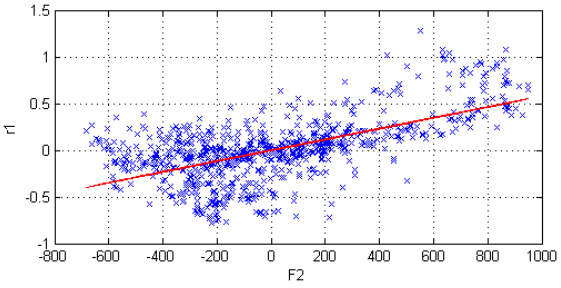

Figure 3*:* r_1 *values (crosses) and its estimation (straight line) in function of* F_2 .

4. Optimizing the multilinear model and Test

The multilinear modeling has been established in the previous section. But the choice of the system parameters, such as the shape and size of the mask, affects the performance of the modeling. Before optimizing the modeling, we set criterions to judge the performance of the modeling. Measures of explained variance (e.g., proportion of variance accounted for) are often considered to indicate the importance of a statistical modeling. The efficiency of the predictor can be measured by the explained variance as follow:

$$VE = 1 - Var_i / Var_0 \qquad (6)$$

where Var_i is the residual variance using F_1 to F_i predictors and Var_0 is the variance of the feature to be predicted (see Fig 4 for B).

The optimization consists to minimize Var_i / Var_0 , the shape and the size of the mask and the prediction order.

If we fix the explained variance to 95%, in other words the Var_i / Var_0 *to 5%*, we find the optimal mask for the modeling.

According to the results of the experiment (Table 1), we found that the triangle mask of side 8 is too small to contain enough information to explain the variance, the same remark as for the square mask of side 7. And we observe that the triangle mask of side 15 and the square mask of side 11 have almost the same good performance.

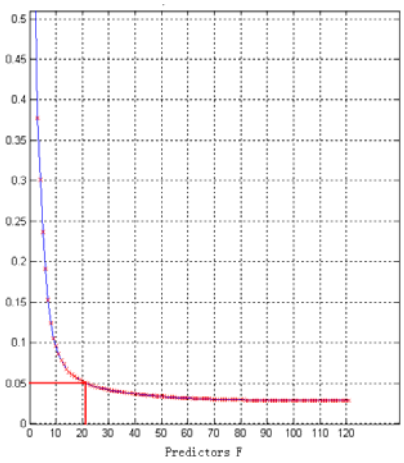

Figure 4: *Evolution of the residual of the B explained variance in function of the prediction order (case of a triangle mask). Var0= 0.1930* cm^2 .

Table 1: *Performance of different masks*

The line corresponding to each mask means the number of the predictors that explain 95% of the variance, otherwise the percentage is the asymptotic value of the explained variance.

RMS = Root Mean Square error (cm), T = Triangle Mask, S= Square Mask

	A	B	S	Aext	Bext	Sext
T of side 15	0.945	21	10	0.93	0.945	40,
RMS	0.30	0.1	0.36	0.116	0.09	0.41
T of side 8	0.90	0.94	13	0.91	0.91	0.93
RMS	0.41	0.17	0.31	0.13	0.12	0.49
S of side 7	0.92	21	12	0.91	0.92	0.94
RMS	0.38	0.1	0.36	0.128	0.11	0.45
S of side 11	63	19	11	0.93	0.945	34
RMS	0.29	0.1	0.35	0.12	0.09	0.41

We can observe from the table that for B , only 19 or 21 predictors are needed to explain 95% of the variance for both masks. It means that less than 20% of the predictors could efficiently estimate the feature.

In order to validate our model, we used it to predict the test data made of the remaining 759 elements of the database that have not been included in the modeling process.

We show in Figure 5 the evolution of the residual variance in function of the prediction order for the test data for B feature. We can see a rapid decreasing of the residual variance to attain a minimum value of 7 % similar to that obtained for the modeling data (see Figure 4). We note that after the order of 20, no significant improvement can be obtained.

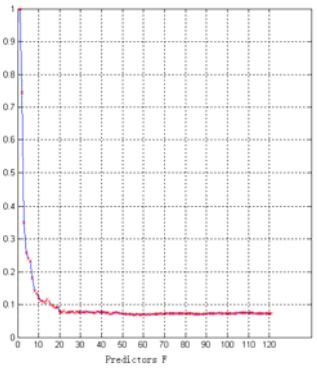

Figure 5: *Evolution of the residual of the B explained variance (test data) in function of the prediction order (case of a triangle mask). Var0= 0.1925 cm*^{2}.

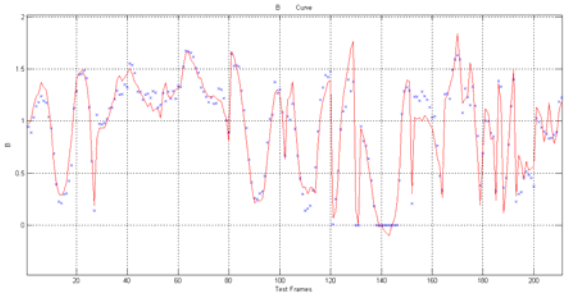

Figure 6: *Estimation Curve for parameter B with 21 predictors, the blue points are the test data, the red line are the estimated values.*

5. Discussion and Conclusion

This paper presents a very innovative modeling for the prediction of speech geometric lip features by concentrated information of the lips obtained with 2-D Discrete Cosine Transform and PCA. This method has the property to concentrate the lip information into a reduce set of predictors. Indeed, the order of prediction could be reduced to 19 or 21 for the explanation of 95% and 93% respectively for the modeling data and test data by comparison to 120 predictors that are needed to gain only 2% in the model data modeling and almost nothing in the test, as it has been observed for the case of the *B* lip aperture feature (see Figure 4 and 5). But the modeling also has some limitation in the sense that it weakly estimates the data close to zero as shown in Figure 6. Finally in the perspective of further works, we extended the test to the case of images containing fingers near the mouth, as it is the case in Cued Speech [22]. The result of this preliminary test shows that the effect of the fingers in the lip ROI is not significant (see also Figure 6 the last 48 frames), the root mean square error being similar for *B* (0.1406 cm with fingers *vs.* 0.1223 cm without).

6. Acknowledgements

The authors would like to thank Myriam Diboui, the speaker, for having accepted the recording constraints. This work is supported by the French ANR/RNTS TELMA project.

7. References

[1] Sumby, W.H., Pollack, I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America 26 (2), 212-215, 1954.

[2] Summerfield, Q. Use of visual information of phonetic perception. Phonetica 36, 314-331, 1979.

[3] Summerfield, Q., MacLeod, A., McGrath, M., Brooke, M. Lips, teeth, and the benefits of lipreading. In: Young, A.W., Ellis, H.D. (Eds.), Handbook of Research on Face Processing. Elsevier Science Publishers, Amesterdam, The Netherlands, pp. 223-233, 1989.

[4] Benoit, C., Lallouache, T., Mohamadi, T., Abry, C. A set of French visemes for visual speech synthesis. In: Bailly, G., Benoit, C. (Eds.), Talking Machines: Theories, Models and Designs. Elsevier Science Publishers, Amsterdam, pp. 485-504, 1992.

[5] Reisberg, D., Mclean, J., Goldfield, A. Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli. In: Dodd, R., Campbell, R. (Eds.), Hearing by Eye : The Psychology of Lipreading. Lawrence Erlbaum Associates Ltd, Hillside, NJ, pp. 97-113, 1987.

[6] McGurk and John MacDonald. "Hearing lips and seeing voices", Nature 264, 746-748, 1976.

[7] MacDonald, J., McGurk, H.. Visual influences on speech perception processes. Perception and Psychophysics 24, 253-257, 1978.

[8] Cotton, J. Normal 'visual-hearing'. Science 82, 582—593, 1935.

[9] Dodd, B. The role of vision in the perception of speech. Perception 6, 31-40, 1977.

[10] Bernstein, L.E., Demorest, M.E., Tucker, P.E. Speech perception without hearing. Perception & Psychophysics 62(2), 233-252, 2000.

[11] Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M.A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Le, V.B., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J.F., Tribout, M., Vidal, S., 2007. TELMA: Telephony for the Hearing-Impaired People, From Models to User Tests. In: proc. ASSISTH'2007, pp. 201–208, 2007.

[12] Potamianos, G., Neti, C., Iyengar, G., Senior, A., Verma, A. A cascade visual front end for speaker independent automatic speechreading. International Journal of Speech Technology, Special Issue on Multimedia, 4, 193-208, 2001.

[13] Potamianos, G., Neti, C, Gravier, G., Garg, A. and Senior, A.W. Recent advances in the automatic recognition of audiovisual speech," in Proceedings of the IEEE, vol. 91, Issue 9, pp. 1306–1326, 2003.

[14] Luettin, J., Thacker, N. A. and Beet., S.W. Visual speech recognition using active shape models and hidden markov models. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 817-820, 1996.

[15] Dupont, S., Luettin, J. Audio-visual speech modelling for continuous speech recognition. IEEE Transactions on Multimedia 2(3), 141-151, 2000.

[16] Hennecke, M., Stork, D., Prasad, K. Visionary speech: Looking ahead to practical speechreading systems. In: Stork, D.G., Hennecke, M.E. (Eds.), Speechreading by Humans and Machines. Springer, Berlin, pp. 331- 349, 1996.

[17] Chiou, G., Hwang. Lipreading from color video. IEEE Transactions on Image Processing 6(8), 1192 – 1195, 1997.

[18] Gray, M., Movellan, J., Sejnowski, T. Dynamic features for visual speech-reading: A systematic comparison. In: Mozer, M.C., Jordan, M.I., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA, pp. 751-757, 1997.

[19] Matthews, I., Bangham, J., Cox, S. Audio-visual speech recognition using multiscale nonlinear image decomposition. In: Proc. International Conference on Spoken Language Processing (ICSLP)'96, Philadelphia, PA, pp. 38-41, 1996.

[20] Matthews, I. Features for Audio-Visual Speech Recognition. Ph.D. thesis, School of Information Systems, University of East Anglia, Norwich, 1998.

[21] Lallouache, M., 1991. Un poste Visage-Parole couleur. Acquisition et traitement automatique des contours des lèvres. Ph.D. Thesis, Institut National Polytechnique de Grenoble, Grenoble.

[22] Cornett, R. Cued Speech. American Annals of the Deaf 112, 3-13, 1967.