Is color information really useful for lip-reading ? (or what is lost when color is not used)

Philippe Daubias

IRISA
Campus Universitaire de Beaulieu,
35042 Rennes CEDEX, France
Philippe.Daubias@irisa.fr

10.21437/Interspeech.2005-366

Abstract

In this paper, we report on experiments aiming at evaluating quantitatively the amount of information carried by color and luminance (gray level images) for automatic lip-reading. More precisely, we focus on the lip location problem: we trained Artificial Neural Networks (ANN) classifiers which were proven to be effective for the lip, skin and inner-mouth classification task, with both color and gray scaled image blocks extracted from the same images. Experiments were conducted with 6 subjects (1 female, 5 males, some with little facial hair) taken from the freely available LIUM-AVS database. A few different ANN architectures were tested, and in all cases, the use of color information enabled an important classification error reduction. Considering all the images blocks from the lip region that were available, the classification error was reduced from 30% for gray-level to 5% using color.

1. Introduction

During the last decade, numerous works [1, 2, 3, 4, 5] have proven the benefit of using visual speech information to improve over acoustic-only automatic speech recognition (ASR): visual information is complementary to acoustic information and enables an error reduction in ASR in all cases although audio-visual (AV) ASR is mainly useful for acoustical noise-prone environments.

Part of these early experiments were either speaker dependent or focussed on hardly constraint environments [6]. Other were carried out using corpora recorded with specific image acquisition devices [7] not available in classical Human-Computer Interaction (HCI) set-ups. Since 2000, with the progresses in calculation speed and data storage capacity, research in the field is more focussed on AV ASR for *open* or *natural* conditions, *i-e* real environments. But accurately extracting the visual speech information in real environments is still a hard task especially if this environment is visually challenging [5]. Whatever the approach, image-based [2, 5] or model-based [6, 8], a reliable estimation of the mouth position is crucial. In practice, the first step for visual speech extraction is to use a face detector [9]. Some perform well and give a precise location of the face, but as the shape of the face differs across subjects, a precise and homogeneous location of the region of interest (ROI) is still necessary.

For the work presented here, we will discuss mainly the ROI location problem, but as the models we use not only detect lip vs non-lip as in [10], but classify each pixel from the region surrounding the mouth as belonging to the lips, skin or mouth cavity classes, they may be used to extract parameters for lip-reading or more directly to produce parameters by themselves, but this is still ongoing research. This paper is organized as follows: in section 2, we propose a quick overview of related works, then, in section 3, we describe in more details the experiments we carried. Results are given in section 4, and we end with our conclusions in section 5.

2. Previous and related work

Some systems are able to produce directly measurements of visual speech without image processing by means of markers such as reflective pastilles and are out of the scope of this paper, where we are only interested in extracting visual speech information from unprepared subjects, if possible recorded with a standard camera in a frontal or near-frontal view.

Since the precursor work in 1984 by Petajan [11] which was using binary black or white images of the mouth cavity, technological progress has given the opportunity to process graylevel images and even color images. Nevertheless, some recent studies [12, 13] still don't use color information, mainly for computation-saving reasons. In the following subsections, we give a quick overview of studies where gray and color images are used.

2.1. Graylevel processing

Humans can extract visual speech information from images of the mouth area even if images are grayscaled. Obviously, this is also the case for machine speechreading as it was proven by [1, 2, 4, 13]. For the oldest studies, the use of gray images was enforced by the limited data acquisition, storage and computation capabilities of that time. For the more recent studies devoted to AV ASR on mobile devices [12, 13], color is available but is not used to keep the processing requirements as low as possible. Even if images are captured in color, they are gray scaled before processing and algorithms which don't use color are applied.

2.2. Use of color information

The first laboratory to use color for automatic lip-reading was probably the ICP in 1990 (work by Lallouache). At this time, blue ink was used to paint the speaker's lips and lighting was cautiously controlled to allow easy lip segmentation. As in real AV ASR applications, the use of makeup is not feasible, some research has been carried since to detect unadorned lips in color images. One of the first studies towards this aim was probably the one by Coianiz [14] who was using hue information centered around red in 1996. Since then, other *a priori* approaches have been proposed [7, 15], mostly relying on hue information. *A posteriori* approaches have also been proposed [10, 16], both

using ANN for the modelization.

For the experiments described next, we used a slightly modified version of the *a posteriori* appearance model proposed in [16] because it achieves very good performance for lip region pixels classification and has the tremendous advantage over *a priori* models being adaptable to different contexts and in that manner to serve as an evaluation tool.

3. Experiments

In this section, we first describe the models which were used, then we focus on the database from which the experimental data was collected and finally we describe in more details the two subsets of data which were used and why they were chosen.

3.1. Model description

For all the experiments presented here, we used artificial neural networks (ANN) also known as multi-layer perceptrons (MLP). These ANN have a very simple topology with one input layer containing i units, one hidden layer of h units and one output layer of o units, and are fully connected from a layer to the next one. The number of input units corresponds to the number of values of the vector to be classified, in our case a 5×5 (color) image block. The output units give, for an image block entered in input, the probability of belonging to the lips, skin and cavity classes, and each hidden unit will be on a different path from input to output thus allowing different modes to go through the network. In the following, networks will be denoted $\mathcal{N}_{i,h,o}$. This type of model was chosen because it reaches as high or higher performance than GMM [16] and through its hidden layer, it doesn't require the user to specify how many mixtures should be used for each class, only the global number h of hidden units has to be specified.

In [16], we used a $\mathcal{N}_{75,15,3}$ network for 3 subjects. Here, a $\mathcal{N}_{75,30,3}$ network will be used for 6 subjects. Hopefully, the number of hidden units doesn't have to be doubled each time the number of subjects is doubled, but we wanted to have high performance for the reference color model and decided to use a greater number of hidden units for that reason. As the number of units in the input layer corresponds to the number of values in the 5×5 image block, and as in gray scaled images, there is only one value per pixel instead of three for color, a $\mathcal{N}_{25,30,3}$ network was used for gray level learning.

All networks were randomly initialized and then trained with the back-propagation algorithm using 20,000 randomly chosen image blocks for each class and for each subject of the database (a total of 360,000 image blocks were used). For the $\mathcal{N}_{25,30,3}$ network, the same blocks were used after gray scaling: For each pixel i, the luminance L_i was used instead of R, G, B color:

$$L_i = \frac{R_i + G_i + B_i}{3} \tag{1}$$

3.2. Database

All the images used for this study were taken from the LIUM-AVS Database [17] which is available freely at <http://www-lium.univ-lemans.fr/avs-database/>. This database was recorded to show the feasibility of an automatic lip-labelling procedure in *natural* conditions [16]. For that reason, it was recorded in lowly constraint conditions, without artificial lighting and is supposed to be close to what would be recorded by a camera in a real HCI environment using AV ASR.

(BJ) (GG) (NR)

(SG) (SJ) (TL)

Figure 1: *A sample image of each subject from the database.*

More precisely, experiments were conducted with 6 subjects (1 female, 5 males) without make up (see Fig. 1) taken from the PBS part of the database. The 5 males have different level of facial hair, subjects BJ and GG having the most. We also carried experiments with 5 subjects (1 female, 4 males) with blue make up taken from the LET part of the database. In total, there were 9 caucasians subjects.

For the blue part of the corpus, fully automated region labelling was used whereas for the *natural* part, a computer-aided hand-labelling procedure had to be done. Next, we'll describe more precisely how the images were processed and explain with which objective these two types of images were used.

3.3. *Natural* images

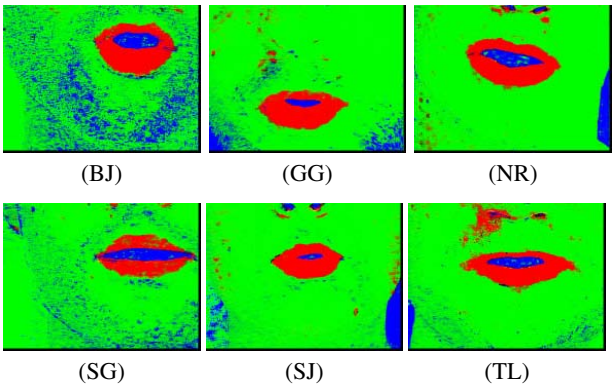
For the *natural* images, a computer-aided hand-labelling procedure was used. The number of hand-labelled images was variable for each subject (from 20 to 70). And from this hand-labelled set, we extracted automatically all 5×5 image blocks from the mouth area and assigned them automatically to the lips, skin or cavity classes according to their position. Heterogenous blocks, *i-e* blocks with parts belonging to more than one class, were discarded for the experiments reported here. In total there were 8 millions image blocks (5.8 million skin blocks, 0.3 million cavity blocks and 1.9 million lips blocks).

3.4. Blue marked lips images

In the LIUM-AVS Database, there are both *natural* and *blue* images and we have also used *blue* images for our experiments. Availability wasn't the only reason to use such *blue* images. They were used for two main reasons:

- 3 out of 5 speakers in the *blue* set are not present in the *natural* set and this should show the generalization ability of the obtained models or in other words, their ability to adapt to other (caucasians) subjects.
- The blue make up does severely modify the lip appearance on color images, but does not affect the skin color, nor much the oral cavity appearance. Nevertheless, blue reflections on the teeth were noticed on some images as they caused errors in previous experiments. When these *blue* images are grayscaled, the blue make up only appears a little darker than unadorned lips. These *blue* images may thus serve to test the models' ability to adapt to a change of lip appearance due to make up.

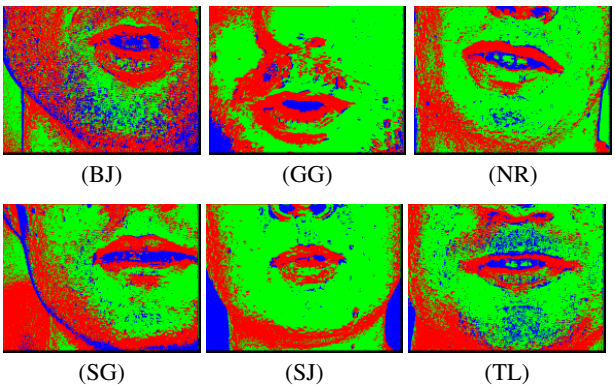
Finally, image blocks acquisition is very fast as it doesn't require any hand-labelling and in the LIUM-AVS Database, all



(BJ) (GG) (NR)

(SG) (SJ) (TL)

Figure 2: *Classification results of the $\mathcal{N}_{75,30,3}$ "color" network for the images shown on Fig. 1.*



(BJ) (GG) (NR)

(SG) (SJ) (TL)

Figure 3: *Classification results of the $\mathcal{N}_{25,30,3}$ "gray" network for the images shown on Fig. 1.*

the *blue* images are already automatically labelled by an accompanying contour file. In total 80 images per subject were used, which lead to 11.7 million homogenous blocks (7.5 million for skin, 1.2 million for cavity and 3.0 million for lips).

4. Results and discussion

In this section, we report in details the results we've obtained for the *natural* images, and then more briefly for the *blue* images. We also discuss the obtained results.

4.1. *Natural* images

Tables 1 and 2 give an overview of the classification results obtained for each subject for the 3 classes of interest using grayscaled or color images. The "global" column presents the results for all the images blocks, but as classes are not equally represented, the more informative "mean" of the three classes is given in the last column.

Results do vary between subjects (see also Fig. 2 and 3), but in all cases and as expected, color blocks are easier to classify than grayscaled blocks. In more details, tables 3 and 4 show the confusion matrixes. One can see that the information carried by gray levels enables to disambiguate the cavity from the skin and lips, but there is more confusion between the skin and lips classes. When using color, there is a lot less confusion between all classes and skin and lips are well discriminated.

One could argue that a $\mathcal{N}_{25,30,3}$ network has less weights to learn than a $\mathcal{N}_{75,30,3}$ network and that the differences between

Table 1: *Classification error rates with grayscaled blocks.*

Subject	Error rates (%)				
	skin	cavity	lips	global	mean
BJ	54.02	24.77	23.64	42.06	34.14
GG	34.24	15.24	25.58	31.34	25.02
NR	21.22	25.49	36.22	25.07	27.64
SG	37.45	23.01	31.06	35.28	30.51
SJ	17.09	28.36	30.72	20.43	25.39
TL	33.80	21.09	31.09	32.94	28.66
Global	30.48	23.67	30.35	30.19	28.17

Table 2: *Classification error rates with color blocks.*

Subject	Error rates (%)				
	skin	cavity	lips	global	mean
BJ	11.45	5.32	3.68	8.51	6.82
GG	2.04	4.41	1.81	2.05	2.75
NR	2.70	9.90	1.37	2.57	4.66
SG	8.90	5.82	12.42	9.66	9.05
SJ	2.39	7.12	0.19	2.01	3.23
TL	6.42	10.18	0.23	5.47	5.61
Global	5.12	7.28	2.48	4.57	4.96

the classification scores are linked to this. Indeed a $\mathcal{N}_{25,30,3}$ network has $25 \times 30 + 30 \times 3 = 840$ weights to learn whereas a $\mathcal{N}_{75,30,3}$ network has $75 \times 30 + 30 \times 3 = 2340$ weights which is about 3 times more, but we also tested gray level image blocks learning with other numbers of hidden units, including a $\mathcal{N}_{25,90,3}$ network which has 2520 weights and for which the results are a little better than for $\mathcal{N}_{25,30,3}$ (about 1% less errors), but are far from the $\mathcal{N}_{75,30,3}$ network, showing that the number of weights isn't the real cause for the higher error rate.

Table 3: *Confusion matrix with grayscaled blocks.*

Class	skin	cavity	lips
skin	69.52	7.35	23.13
cavity	11.65	76.33	12.02
lips	22.36	8.00	69.65

Table 4: *Confusion matrix with color blocks.*

Class	skin	cavity	lips
skin	94.88	2.46	2.66
cavity	5.51	92.72	1.77
lips	1.83	0.65	97.52

4.2. *Blue* images

Contrary to the *natural* images on which the results were quite predictable, the results obtained on the *blue* images were more unexpected. Table 5 shows the results for both color and grayscaled approaches. As expected, the errors in lips classification rise from 2.5 % to 99.5 % for color image blocks, but also rise from 30.3 % to 70.2 % for grayscaled blocks. This degradation of performance is greater than what was expected for the $\mathcal{N}_{25,30,3}$ "gray" network: the blue lips although appearing for us somewhat similar to unadorned lips when grayscaled, are modified in luminance —darkened— by the make up and are misclassified as oral cavity (see table 6).

Looking more precisely at the results through the confusion matrixes which are shown in tables 6 and 7, one can notice that

Table 5: *Global classification error rates on* blue *images.*

Type of blocks	Error rates (%)			
	skin	cavity	lips	mean
color	17.13	29.15	99.53	48.60
grayscaled	58.31	33.84	70.19	54.11

the error rates rise for all classes, not only for lips. For the cavity which is mostly misclassified as lips, an explanation can be given: in the *blue* images, tongue is often visible and image blocks of the tongue may easily be recognized as lips, whereas for the skin blocks which are wrongly classified as cavity, explanations are still to be found.

Finally, the very poor results obtained for blue lips (0.47 % of correct classification), show that our *a posteriori* model has essentially learnt color for lip segmentation.

5. Conclusions and future work

In this paper, we have presented experiments aiming at evaluating the amount of information useful for lipreading carried by luminance (gray levels) and color.

For the six subjects studied, we have obtained similar results, where luminance alone carries about two thirds of the information and the last third may only be recovered through color. Other experiments reported elsewhere [16] have also shown that *a priori* models based only on hue reach about 80 % correct classification compared to more than 95 % with complete color for the same problem. Hue probably carries more information for the lip segmentation task than luminance, but the best results are obtained with full color or in other words with a combination of luminance, hue and saturation.

Although luminance carries some information usable for lip segmentation, an important part of the lip segmentation information is only available in color. When discarding color, this important information which enables the discrimination between lips and skin is lost.

In some of the most advanced research on AV ASR [5], color is used for ROI location but is discarded when calculating the visual speech parameters. We believe that the color information carries information useful for AV ASR and that it should be integrated as a visual speech parameter and will work towards this aim in the future.

6. Acknowledgements

The author would like to thank the LIUM laboratory for providing the LIUM-AVS Database used for this study.

Table 6: *Confusion matrix on* blue *images with gray blocks.*

Class	skin	cavity	lips
skin	41.69	23.13	35.18
cavity	12.35	66.16	21.49
lips	5.20	64.99	29.81

Table 7: *Confusion matrix on the* blue *images with color blocks.*

Class	skin	cavity	lips
skin	82.87	12.92	4.21
cavity	9.31	70.85	19.84
lips	93.58	5.95	0.47

7. References

[1] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *Proc. ICASSP*, vol. II, 1994, pp. 669–672.

[2] I. Matthews, J. Bangham, and S. Cox, "Audiovisual speech recognition using multiscale nonlinear image decomposition." in *Proc. ICSLP*, 1996, pp. 38–41.

[3] M. T. Chan, "HMM-based audio-visual speech recognition integrating geometric- and appearance-based visual features," in *Proc. MMSP*, 2001, pp. 9–14.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[5] J. Jiang, G. Potamianos, H. Nock, G. Iyengar, and C. Neti, "Improved face and feature finding for audio-visual speech recognition in visually challenging environments," in *Proc. ICASSP*, vol. V, 2004, pp. 873–876.

[6] M. Heckmann, T. Wild, F. Berthommier, and K. Kroschel, "Comparing audio- and a-posteriori-probability-based stream confidence measures for AV speech recognition," in *Proc. Eurospeech*, 2001, pp. 1023–1026.

[7] M. Liévin and F. Luthon, "Lip features automatic extraction," in *Proc. ICIP*, vol. 3, 1998, pp. 168–172.

[8] Z. Wu and P. S. Aleksic, "Inner lip feature extraction for mpeg-4 facial animation," in *Proc. ICASSP*, vol. III, 2004, pp. 633–636.

[9] A. W. Senior, "Face and feature finding for a face recognition system," in *Proc. AVBPA*, 1999, pp. 154–159.

[10] J. C. Wojdel and L. J. M. Rothkrantz, "Using aerial and geometric features in automatic lip-reading," in *Proc. Eurospeech*, 2001, pp. 2463–2466.

[11] E. D. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. Thesis, University of Illinois, 1984.

[12] J. F. Guitarte Pérez, K. Lukas, and A. F. Frangi, "Low resource lip finding and tracking algorithm for embedded devices," in *Proc. Eurospeech*, 2003, pp. 2253–2256.

[13] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-visual speech recognition using lip movement extracted from side-face images," in *Proc. AVSP*, 2003, pp. 117–120.

[14] T. Coianiz, L. Torresani, and B. Caprile, "2D deformable models for visual speech analysis," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. Berlin: NATO ASI Series, Springer, 1996, vol. 150, pp. 391–398.

[15] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP JASP, special issue on Joint Audio-Visual Speech Processing*, no. 11, pp. 1228–1247, 2002.

[16] P. Daubias and P. Deléglise, "Statistical lip-appearance models trained automatically using audio information," *EURASIP JASP, special issue on Joint Audio-Visual Speech Processing*, no. 11, pp. 1202–1212, 2002.

[17] ——, "The LIUM-AVS database : a corpus to test lip segmentation and speechreading systems in *natural* conditions," in *Proc. Eurospeech*, 2003, pp. 1569–1572.