AUDIOVISUAL SPEECH ENHANCEMENT EXPERIMENTS FOR MOUTH SEGMENTATION EVALUATION

Pierre Gacon[(1)], Pierre-Yves Coulon[(1)], Gérard Bailly[(2)]

LIS-INPG[(1)], ICP-INPG[(2)]
46, Avenue Félix Viallet
38031, Grenoble, France
{gacon, coulon}@lis.inpg.fr, bailly@icp.inpg.fr

ABSTRACT

Mouth segmentation is an important issue which applies in many multimedia applications as speech reading, face synthesis, recognition or audiovisual communication. Our goal is to have a robust and efficient detection of lips contour in order to restore as faithfully as possible the speech movement. We present a methodology which focused on the detection of the inner and outer mouth contours which is a difficult task due to the non-linear appearance variations. Our method is based on a statistical model of shape with local appearance gaussian descriptors whose theoretical responses were predicted by a non-linear neural network. From our automatic segmentation of the mouth, we can generate a clone of a speaker mouth whose lips movements will be as close as possible of the original ones. In this paper, results obtained by this methodology are evaluated qualitatively by testing the relevance of this clone. We carried out an experience which quantified the effective enhancement in comprehension brought by our analysis-resynthesis scheme in a telephone enquiry task.

1. INTRODUCTION

Lips segmentation can apply to various research areas such as automatic speech recognition (in human-computer interface), speaker recognition, face authentification, or to improve speech intelligibility in noisy situation for audio-video communication. Extracting the shape of lips and modeling it with a few number of parameters can allow low-bandwidth communication or to animate a clone or an avatar of a person.This last application is in the continuity of RNRT project TEMPOVALSE (http://www.telecom.gouv.fr/rnrt/rnrt/projets/res_d17_ap99.htm).

If there are various applications, it still remains a task difficult to achieve. The variability of lips' shape or of skin colour from a speaker to another, inconstant lighting conditions, unusual or extreme lips movements are factors which can decrease the precision and robustness of algorithms.

So a huge diversity of methodologies have been developed and tested to achieve lips segmentation in the last few years. We can broadly discern three big families of methods: 1) methods without lips model, 2) methods with parametrical lips model and lastly 3) methods with statistical lips models.

In the first case, with no lips model of any kind, only information as colour or edge are used.

For example, Zhang [1] proposed to used the hue information and a contour detection to segment the mouth. But without shape constraints the result were quite crude. Delmas [2] used the same informations but with an active contour (or snake) approach by using a gradient criterion. The parameter of the snake could be adjusted in order to have enough flexibility to depict the lips shape with regular contours. This type of method can give convincing results if the conditions are favourable : controlled lighting, good contrast between colour of lips and skin. But in more tricky cases, the segmentation will become far more difficult and robustness will heavily decrease and that can be a problem for application where the precision of the contour should be high. Above all, nothing insure that the result will even be realistic.

In order to obtain a realistic contour in almost all cases, it is very useful to have a model for the shape of the lips. Many authors proposed to use parametric models in which the lips shape will be described by curves controlled by a limited set of parameters. The main difficulty for this type of method consists in finding the good dosage of flexibility. A model too inflexible will always give credible lips contours but will fail to detect faithfully uncommon mouth shape. A model too flexible will adapt itself to much of the situations but will give occasionally unrealistic results.

For example, Hennecke et al. [3] used a deformable template. The template is a model of the lips controlled by a set of parameters which are chosen by minimizing a criterion based on the edges of the lips. For this kind of approach, the lack of flexibility of the template can be a problem. Eveno [4] proposed to use cubic curves to describe the lips with constraints and limits to the values of the derivative at salient key points. These curves are fitted to the processed image contours by using a gradient information based on hue and luminance. These curves are very flexible, but can still generate impossible shapes.

In the last few years the statistical model approach has been used very often. This kind of method can successively segment mouth shape as long as the training data used to build the model is big enough to be representative of the real variability of cases (if a configuration is not present in the training set it will be unlikely well segmented). In the beginning, Cootes et al. [5] introduced active shape models (ASM). The shape of an object is learned from a training set of annotated images. After a principal component analysis (PCA) a limited number of parameters drives the model. The main interest is that the segmentation will always give a realistic result (given that the distribution of the data is effectively Gaussian). Values of the parameters are selected with an appropriate criterion based on gradient information. Later, Cootes et al. [6] introduced active appearance models (AAM) in which appearance (grey-level pixels values) is also modelized by PCA. The model converges by iteratively reducing the difference between the modeled appearance and the real appearance of the processed image. Luettin [7] also developed an active shape model method in which he learned a grey-level profile model around lips contour in the training set. This profile model provides a measure of the goodness of fit between the model and the image.

In our work, we chose to follow the statistical model approach for its ability to always segment realistic shape of lips. We built an

active shape and sampled-appearance model to describe the mouth area and we placed a special emphasis on the segmentation of the inner mouth area by trying to build a model able to adapt itself progressively to a speaker during a video sequence.
We created a sampling mesh to capture pixels values in precise location. This is particularly adapted for the inner lip contour and mouth bottom which present a high variability and non-linearities (mouth close or open, teeth or tongue presence).
The appearance (pixels values) has been divided in two components: a static appearance (which is constant for one speaker) and a dynamic appearance (which corresponds to the variation induced by movement and, particularly, speech).
We find mouth corners with a local appearance model, which estimates the position and the scale of the mouth.
We use a non-linear cost functions to optimize the parameters of our active model. It is computed as the difference between the response of local gaussian derivative descriptors computed on the processed image and the prediction of this response made by a non-linear neural network with the mouth model parameters as entries.
This whole methodology has been presented and detailed in our previous papers ([11], [12], [13]) for mono and multi speaker tasks. In these papers, the relevance of the segmentation was classically evaluated quantitatively by computing the positioning error of control points but it does not really tell if the accuracy is sufficient for lip reading.
So, even if our work does not deal directly with the speech recognition problem, we checked the benefit of our analysis/resynthesis system using classical audiovisual enhancement experiments. It is indeed well-known that visual information correspond to 12dB signal-to-noise enhancement when compared to signal-only condition ([14], [15]). Normaly test stimuli are isolated words or nonsense words. Here perception of phone numbers in noisy condition is used for comparing the impact of the original video versus our system. We show that our processing scheme tracks and synthesize lips movements with enough reliability to offer 75% of the benefit brought by the natural face

2. DATA SET

As color and brightness are mixed in the RGB color space we chose to work with the YCbCr space where chromatic and luminance components are separated.
The data-set consists in 40 videos of a speaker saying 10 digits phone number (approximately 8000 frames). In half of these videos this speaker is whispering while in the other half his elocution is normal.
N=100 images of this data-set were selected to be our training set : key-frames in order to have a wide variety of mouth shape.

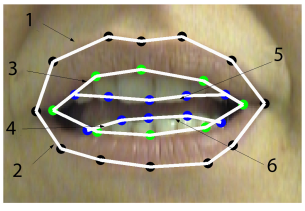

Figure 1 - Example of annotated image, Curves 1 and 2: outer mouth contour, 3 and 4: inner mouth contour, 5 and 6: teeth

Then, this training set was manually annotated to build the model (the remaining images were used to test the algorithm). The general shape is described by 30 control-points: 12 for the outer lip contour, 8 for the inner lip contour and 10 for teeth, as shown in figure 1. If the mouth is closed or if the teeth don't appear, the corresponding points are merged with those of the inner lip contour.
In our work, we consider that the face is found in a pre-processing step and so that the mouth neighbourhood (the low part of the face) is known.

3. MODELS

3.1 Mouth corners local model and detection

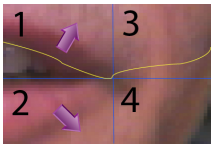

Figure 2 - Mouth corners area with characteristics regions, edge direction and line of luminance minima (yellow)

Mouth corners points are used as key-points to determine the position and scale of the mouth. Mouth corners are considered as the intersection point between 4 regions : regions 1 and 2 are non-homogeneous as they are characterized by an edge between lips and skin while regions 3 and 4 are homogeneous on a chromatic point of view (figure 2). Mouth corners will then be described by a set of 4 Gaussian derivative filters and the statistical distribution of the responses of the filters are described by Gaussian mixture models.
As Eveno [4] proposed, the mouth corners are supposed to be on the line which links luminance minima for each column so we compute the local descriptors for each pixel of the line of interest and the most probable couple of point is selected as mouths corners. More details about this detection can be found in [11].

3.2 Active Shape and Appearance Models

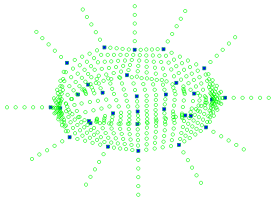

Figure 3 - Mesh used for the sampling of appearance. Plain circles: control-points, empty circles: mesh

After annotating the training in part 2), we have obtained vectors $\boldsymbol{X_i}$ ($1 \le i \le N$) which contain the shape. The next step is to learn the sampled-appearance for every image. The three YCbCr components are extracted at 728 features-points given by a mesh computed from the $\boldsymbol{X_i}$ (as shown in figure 3).
Using this mesh defines precisely if an appearance sample corresponds to skin, lips, teeth or inner mouth. It is particularly adapted for the inner mouth area which has a non-linear behavior: mouth is closed or open, teeth and tongue are apparent or not.
Sampled appearance YCbCr values in 2184 (728x3) values are saved in vectors $\boldsymbol{A_i}$ ($1 \le i \le N$).
We then proceed to dimensional reduction by doing a PCA on the data $\boldsymbol{S_i}$ and $\boldsymbol{A_i}$. The respective mean vectors ($\overline{X}$, $\overline{A}$), the covariance matrices ($\mathbf{C_x}$, $\mathbf{C_a}$), their respective eigenvectors ($\boldsymbol{p_{x,m}}$, $\boldsymbol{p_{a,n}}$) and eigenvalues ($\lambda_{x,j}$, $\lambda_{a,n}$) with $1 \le j \le 60$ and $1 \le n \le 2184$) are then calculated.
The eigenvectors of the covariance matrices correspond to the

various variation modes of the data. As the eigenvectors with large eigenvalues describe the most significant part of the variance, the selection of a few modes can reduce the dimensionality of the problem.

We keep 95% of the variance and the selected eigenvectors are saved in matrix $\mathbf{P}_X$. For example with our data, it corresponds to 7 eigenvectors. So shape will be described by a few parameter : any shape of the training set or new plausible examples can then be generated by simply adjusting vector parameter x with the following equation:

$$X = \overline{X} + \mathbf{P}_x x$$

Figure 4 shows the effects of the first two modes of variation for shape. The weight parameter σ varies from -2*standard deviation to +2*standard deviation.

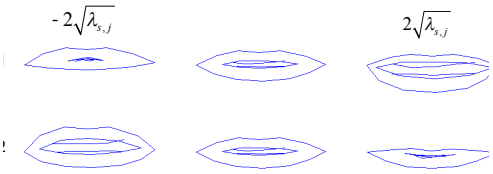

Figure 4 - First 2 modes of variation for the shape model

For appearance, we obtain the following set of equations, with the 9 selected eigenvectors corresponding to 90% of the variance saved in matrices $\mathbf{P_a}$:

$$A = \overline{A} + \mathbf{P}_a a$$

Next we want to make a second level PCA (as in [6]) for shape and appearance parameters in order to have a coherent modelization between the two informations. First, we have to compute for each image the values of the corresponding parameters-vectors X_i (for shape) and A_i (for appearance) with $1 \le i \le N$.

$$x_i = \mathbf{P}_x^T \left(X_i - \overline{X}\right) ; a_i = \mathbf{P}_a^T \left(A_i - \overline{A}\right)$$

We then proceed to the new PCA with the values of these parameters and we obtain a statistical model which links shape variation (represented by the X_i) and appearance variability (represented by the A_i).

Finally we can generate any shape and sampled-appearance of the training set or new plausible examples by simply adjusting c in the following set of equations:

$$(1)\ C = \begin{bmatrix} Wx \\ a \end{bmatrix} = \mathbf{P}_c c \Rightarrow \begin{cases} (2)\ X = \overline{X} + \mathbf{P}_x x \\ (3)\ A = \overline{A} + \mathbf{P}_a a \end{cases}$$

Equation (2) controls the shape model, equation (3) controls appearance and equation (1) controls the combined model for shape and sampled appearance. Segmenting mouth on an image will then consist in finding the best set of 9 parameters contained in combined parameter vector c.

3.3 Non-linear cost function and parameters optimization

The optimization of our active model parameters is achieved by minimizing an adapted cost function that will be called C_f.

The minimization itself, which is a high dimension problems, is performed by the Downhill Simplex Method (DSM), which is a minimization/maximization classical method.

Our cost function is based on the idea that for a tested set of parameters, we are able to predict the response of some adapted local appearance descriptors and to compare it to the response observed on the image. If this prediction is non-linear, then the cost function would be particularly adapted for the inner lip contour and mouth bottom which present a high variability and non-linearities.

We finally chose to use the first gaussian derivative filters [8] as local descriptors. These filters are convolution windows (their sizes is one tenth of the mouth width) and we limit the model to the first Gaussian derivatives so we will compute the convolutions between the 3 filters G and Gx and Gy (mean and horizontal and vertical gradients, ie figure 5) and the image.

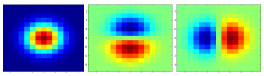

Figure 5 - Gaussian derivative filters G and Gx and Gy

As we want to predict the response of the gaussian filters from the model parameters, this response has to only depend of them. As we work on a single speaker task, there is no inter person variability. Nevertheless lightening change can modify the response of the filter for the same person and the same shape. So we use the retina filter [10] on the luminance to diminish this variability. This filter is a band pass spatial filter which enhance contours and reduce illumination variation.

This prediction will be achieved by a neural network which has to be able to deal with non-linear problems (as the response of filters on the inner contour vary non-linearly when the mouth is opening, or when teeth or tongue appears) so we chose to use feed forward backpropagation to build a two layers neural network. To diminish the its size, we use the active model parameters as entry and we do a PCA on the outputs (descriptors responses) to keep 80% of the variance.

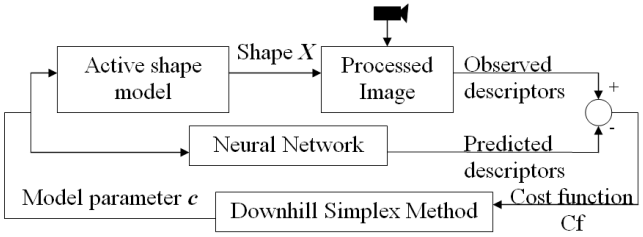

Figure 6 - Principle of segmentation of mouth segmentation with local descriptors

For a tested set of parameters, our cost function C_f is then defined as the mean square error between the response predicted from the parameters values and the actual response of the filters computed on the processed image. Figure 6 summarizes the whole optimization scheme which allows to find the best set of parameters.

The relevance of this cost function, how to adapt it to a multi-speaker task and preliminary steps to improve the speed and accuracy of the segmentation were fully presented and discussed in our previous papers [11], [12], [13].

4. RESULTS AND EVALUATIONS

4.1 Objective Evaluation

outer lip contour	inner lip contour	teeth	all points	number of iterations
1.4/0.8	1.8/0.9	1.8/0.9	1.5/0.8	9.8

Table 1 - Results of a objective, or quantitative, evaluation of the method.

Results are 2-D mean errors and standard deviation for the shape points between detected and annotated points when the method runs on an unknown image. These mean errors are given in percentage: the difference between detected points and annotated point is normalized by the width of the mouth.
These results are quite good and seems to prove the segmentation is accurate and robust, yet they do not really give any information on the realism of the speech movement that can be synthesized from the analysis.

4.2 Subjective Evaluation : listening tests

We saw that when the values of parameters are known, the sampled appearance can be interpolated to create a synthetic clone of the speaker's mouth. If the reconstruction gives a result which seem convincing and quite realistic, it is difficult to really evaluate its quality.

Figure 7 - Examples of mouths rendering

It has been demonstrated that the visual information improved the speech intelligibility in degraded acoustical situations. The extreme example is the pure lip-reading: absence of sound so only the lip contour can be used.
So if our analysis-resynthesis scheme is relevant, it should improve the speech intelligibility in noisy conditions. In order to quantify the effective enhancement in comprehension brought by our method, we carried out an understanding evaluation in a telephone enquiry task.
To do so we used 40 videos of a speaker pronouncing 10 digits phone number. These 40 sequences last 8 seconds in average so in the whole it represents almost 8000 frames. In the half of the sequences, the speaker is whispering while in the other half his elocution is normal.
We segmented the mouth area for all the sequences and we generated 8 categories of stimuli for normal elocution and the same 8 categories for whispered elocution:
1) Audio alone with a reference noise SNR=0 dB
2) Audio alone with a strong noise SNR=-18 dB
3) Natural video and audio with a reference noise SNR=0 dB
4) Natural video and audio with a strong noise SNR=-18 dB
5) Synthetic video and audio with a reference noise SNR=0 dB
6) Synthetic video and audio with a strong noise SNR=-18 dB
7) Natural video alone
8) Synthetic video alone
The protocol of the experience was as follow: each subject watched 4 random stimuli as example (for various level of noise and elocution).
Next, each subject watched 40 stimuli (20 in whispered elocution and 20 normal elocution) and writes down the phone numbers. As the subject were asked to always look the screen, they had to write without looking at their sheet.

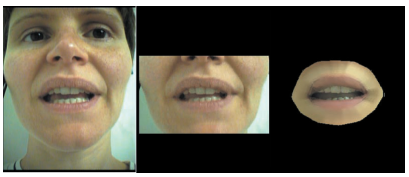

Figure 8 - Original frame, example of a frame of a "natural video" used for listening tests, example of a frame of a "synthetic video"

16 subjects took part in this experience so that each of the 40 sequence was listened twice for each stimuli. Mistakes made on the retranscription of digits of the phone numbers have been counted for each stimuli. There were 10 digits by phone numbers, but as the first digit was always '0', it was not take into count for the statistics. The results, in percentage of good understanding (mean and standard deviation) are shown in Table 2.

Type of Stimuli	Normal elocution	Whispered elocution
1)	97.5 % 8.4 %	92 % 10.3 %
2)	44.2 % 16.5 %	35 % 19.1 %
3)	99.4 % 3.4 %	98.9 % 3.3 %
4)	85 % 15.5 %	85.9 % 17 %
5)	100 % 0 %	98.8 % 3.7 %
6)	76 % 16 %	73.3 % 15.4 %
7)	42.5 % 26 %	56.1 % 28.1 %
8)	27 % 19 %	33.9 % 23.1

Table 2 - Results of our subjective evaluation. Values are the mean percentage of digits successfully understood by a subject, and its standard deviation, for each of the 8 stimuli and for each elocution

If we compare the stimuli "audio and synthetic video" (5-6) with the stimuli "audio alone" (1-2), we see that with a strong noise, our resynthesis scheme improves understanding by 32% for a normal elocution and by 38% for a whispered elocution. With the reference noise, almost no mistakes were made by subjects with the help of synthetic video.
If we compare the stimuli "audio and synthetic video" (5-6) with the stimuli "audio and natural video" (3-4), we see that our resynthesis scheme does not totally meet the improvement in understanding brought by the original video. But the result remains comparable: in average, with a strong noise, subjects made one additional mistake on a digit with our synthetic video.
If we compare the stimuli "synthetic video alone" (8) with the stimuli "natural video alone" (7), the difference of results is more important as in average people made two additional mistakes with our synthetic videos than with the natural ones. This difference may be in part due to the fact that our model only synthesize the

mouth area, while in the natural image the whole low face area was observable by the subjects.

To look further these results, we made a control experience by testing the understanding improvement in cases where the control points had been manually annotated and the appearance directly sampled on the original image (for 5 normal elocution cases). This is exactly as if we had performed a "perfect" segmentation of the mouth on these videos (the appearance being directly extracted on the original images), the results are given by Table 3.

If we compare the comprehension percentage brought respectively by the clones generated from the automatic and manual segmentation, we notice that the results are quite similar. So it seems that the difference of results observed with the natural video does not fully come from an eventual lack of precision on the contour detection (the analysis part) but also from the generation of the clone (the resynthesis part) even if it remains difficult to determine their respective impact on the result

Nevertheless, a more realistic aspect of the clone and a modelization of other useful speech clues (like jaw movements) might improve the results of this listening test.

Type of Stimuli	Automatic segmentation	Manual segmentation	Natural video
Audio and video SNR 0dB	100 % 0 %	99.4 % 3.4 %	99.4 % 3.4 %
Audio and video SNR -18dB	76 % 16 %	74 % 16 %	85 % 15.5 %
Video alone	27 % 19 %	25 % 20 %	42.5 % 26 %

Table 3 - Results of our subjective evaluation for automatic and manual evaluation

These results show that despite the poor intelligibility of synthetic stimuli for lip reading compared to natural lip movements, they enhance audiovisual comprehension. In short they are not able to provide phonetic information but enhance acoustic-to-phonetic decoding. This is quite in accordance with current models of audiovisual integration that put forward early fusion mechanisms where the shallow correlation between facial movements and speech spectrum [16] is used to enhance certain spectral zones of the input signal [17]. Such adaptive filtering techniques have been already applied to audiovisual source separation [18].

5. CONCLUSION

We presented in this paper a novel methodology based on a non-linear local appearance description to segment the lips contours. We also presented a way to evaluate the relevance of the segmentation with some listening tests. Finally, we can say that our analysis/ resynthesis scheme do bring an improvement in understanding even if it does not match the real video case.

It must be noted that in this present work, we only studied a mono speaker task. We have already tested our segmentation method for a multi speaker task in past papers with some objective tests and it proves to be quite accurate and robust. But this kind of qualitative evaluation should be performed in the future in order to fully validate our analysis scheme.

Segmenting other speech clues, like jaw movements for example, might also improve the efficiency of the resynthesis of the speech.

6. ACKNOWLEDGMENTS

This work has been financed by federation ELESA FR8-CNRS-INPG-UJF.

7. REFERENCES

[1] X. Zhang, R. M. Mersereau, M. A. Clements and C. C. Broun, "Visual Speech Feature Extraction for Improved Speech Recognition", In Proc. ICASSP'02, pp. 1993-1996, 2002.

[2] P. Delmas, N.Eveno, and M. Lievin, "Towards Robust Lip Tracking", *International Conference on Pattern Recognition (ICPR'02),*Québec City, Canada, August 2002

[3] M. Hennecke, V. Prasad, and D. Stork. "Using deformable templates to infer visual speech dynamics", *28 h Annual Asimolar Conference on Signals, Systems, and Computer*, *volume 2*, IEEE Computer, Pacific Grove, pp. 576-582, 1994.

[4] N.Eveno, A. Caplier, and P-Y Coulon, "Automatic and Accurate Lip Tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no.5, pp. 706-715, May 2004

[5] T.F. Cootes, C.J. Taylor, and D.H. Cooper, "Active Shape Models - Their Training and Application", Computer Vision and Image Understanding, Vol. 61, No. 1, Janvier, pp. 38-59, 1995

[6] T. F. Cootes, G.J. Edwards, and C.J. Taylor. "Active Appearance Model", *Proc. European Conference on Computer Vision 1998 (H. Burkhardt and B. Neumann Ed.s)*, Vol. 2, pp. 484-498, Springer, 1998.

[7] J. Luettin, N.A. Thacker, S.W. Beet, "Locating and Tracking Facial Speech Features", *Proceedings of the International Conference on Pattern Recognition*, Vienna, Austria, 1996.

[8] T. Lindeberg, "Feature detection with automatic scale detection", *IJVC*, vol. 30, no.2, pp. 77-116, 1998.

[9] T. Poggio, and A. Hulbert, "Synthesizing a Color Algorithm From Examples", *Science*, Vol 239, pp 482-485, 1998.

[10] Beaudot W.H.A., "The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision", PhD Thesis in Computer Science, INPG (France) december 1994

[11] P. Gacon, P.-Y. Coulon, G. Bailly. "Statistical Active Model for Mouth Components Segmentation", *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, Philadelphia, USA, 2005.

[12] P. Gacon, P.-Y. Coulon, G. Bailly. "Non-Linear Active Model for Mouth Inner and Outer Contours Detection", *2005 European Signal Processing Conference (EUSIPCO'05)*, Antalya, Turkey, 2005

[13] P. Gacon, P.-Y. Coulon, G. Bailly. "Modèle Statistique et Description Local d'Apparence pour la Détection du Contour des Lèvres", *20e Colloque sur le traitement du signal et des images (GRETSI'05)*, Louvain-la-Neuve, Belgium, 2005

[14] Q. Summerfield, P. MacLeod, M. McGrath, N.M. Brooke. "Lips, teeth, and the benefits of lipreading". *Handbook of Research on Face Processing*, A.W. Young and H.D. Ellis (eds.) New Holland: Elsevier, pp. 223-233, 1989.

[15] Q. Summerfield, M. McGrath. "Detection and resolution of audio-visual incompatibility in the perception of vowels". *Quarterly Journal of Experimental Psychology*, 36,51-74, 1984.

[16] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson (1998). Quantitative association of vocal-tract and facial behavior. Speech Communication, 26(1): 23-43.

[17] J.-L. Schwartz, F. Berthommier, C. Savariaux. "Seeing to hear better: evidence for early audio-visual interactions in speech identification", *Cognition*, 93, 69-78, 2004.

[18] J.-L. Schwartz, D. Sodoyer, L. Girin, J. Klinkisch, C. Jutten. "Separation of Audio-Visual Speech Sources: A New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli", *EURASIP Journal on Applied Signal Processing*, 11, 1165-1173, 2002.