Development and comparison of two approaches for visual speech analysis with application to voice activity detection

Bertrand Rivet[1,2], *Andrew Aubrey*[3], *Laurent Girin*[1], *Yulia Hicks*[3], *Christian Jutten*[2], Jonathon Chambers[3]

[1,2]Grenoble Image Parole Signal Automatique (GIPSA - [1]ICP/[2]LIS)
CNRS UMR 5216 , Grenoble Institute of Technology (INPG), Grenoble, France
emails: {rivet,girin}@icp.inpg.fr, {rivet,serviere,jutten}@lis.inpg.fr
[3] Centre of Digital Signal Processing
Cardiff School of Engineering, Cardiff University, U.K
emails: {aubreyaj,hicksya,chambersj}@Cardiff.ac.uk

Abstract

In this paper[1] we present two novel methods for visual voice activity detection (V-VAD) which exploit the bimodality of speech (*i.e.* the coherence between speaker's lips and the resulting speech). The first method uses appearance parameters of a speaker's lips, obtained from an active appearance model (AAM). An HMM then dynamically models the change in appearance over time. The second method uses a retinal filter on the region of the lips to extract the required parameter. A corpus of a single speaker is applied to each method in turn, where each method is used to classify voice activity as speech or non speech. The efficiency of each method is evaluated individually using receiver operating characteristics and their respective performances are then compared and discussed. Both methods achieve a high correct silence detection rate for a small false detection rate.

Index Terms: visual voice activity detection, audiovisual speech

1. Introduction

Voice activity detectors (VADs) are used to detect the presence or absence of speech in an acoustic environment. As VAD methods traditionally rely on acoustic information, their accuracy is highly dependent on the acoustic environment (e.g. the presence of competitive sources or highly non stationary noise). However, speech is a bi-modal signal with both audio and visual aspects. The most visible aspect of speech production is the movements of lips; in the past it has been shown that there is a coherence between the speaker's lips and the resulting acoustic signal [1]. This characteristic has already been used to improve speech recognition [2] and speech enhancement [3]; and more recently in blind speech separation [4, 5]. Recently, VAD based on visual data as opposed to audio data has been developed [6, 7, 8]. Visual voice activity detection (V-VAD) has an advantage over audio based VAD in that it is not susceptible to the problems associated with the acoustic environment (e.g. noise, simultaneous speakers, reverberations, etc.).

Previously, Iyengar and Neti [6] developed a V-VAD which was used for deciding a person's intent to speak. The V-VAD uses a head pose and lip motion detector to switch a microphone on and off in a speech recognition system. The drawback of this method is that it does not distinguish between speech and non-speech movement of the lips. Liu and Wang [7] proposed a V-VAD that used statistical models of speech and non-speech activities. Visual information relating to non-speech was modelled using a single Gaussian distribution and visual speech information modelled using a mixture of two Gaussian distributions. New data were classified on the basis of a likelihood calculation. However, their method does not model the dynamics of lip motion. More recently, Sodoyer *et al.* [8] proposed a method for V-VAD that uses temporal smoothing of dynamical lip motion. Unfortunately, their method is a high computational cost chroma-key system, which is impractical for a natural environment.

In this paper, we propose two new methods that use visual information for VAD. The first method, presented in Section 2, is a dynamical model that classifies appearance parameters of a speaker's lips using an HMM. The second method, introduced in Section 3, exploits a retinal filter on the region of the lips, is a simple and low computational cost V-VAD based on dynamical lip motion. Section 4 describes numerical evaluation before conclusions and perspectives are given in Section 5.

2. Visual Detection using a Statistical Model of Appearance Parameters

Cootes and Taylor [9] introduced active appearance models (AAMs) as a way of modelling selected visual features. An AAM is a joint statistical model of shape and colour values (texture), where a single appearance parameter defines a corresponding texture and shape vector. The model is built in several stages. Firstly, the lip shape is tracked through the video by placing landmarks (manually or automatically) on the outer edge of the lips (Fig. 1). Each landmark is represented with its Cartesian coordinates (x_i, y_i). For a single image, the vector $\mathbf{z}$ describing the shape of the lips is defined as:

$$\mathbf{z} = [x_1, \cdots, x_N, y_1, \cdots, y_N]^T. \quad (1)$$

A collection of vectors $\{\mathbf{z}(j)\}_{1 \leq j \leq Tmax, j \in \mathbb{N}}$ describes the variation of the lip shape over a set of images. Next, a statistical model of the shape variation is generated from $\{\mathbf{z}(j)\}$. To build the texture model all images within the boundaries described by the set $\{\mathbf{z}(j)\}$ are first warped to the mean shape to create a 'shape-free' patch. We then obtain the texture (colour values) within this patch for each image to form a texture vector for each image: $\mathbf{g} = [g_1, g_2, \cdots, g_N]^T$. Next we perform PCA on the shape and texture separately. The sum of the outer products

[1]This paper is based on work already submitted to EUSIPCO 2007.

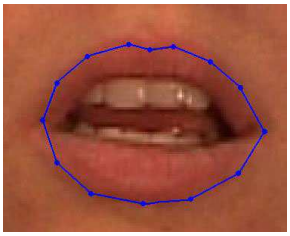


Figure 1: Example landmarks extracted from the lip edges, shown as the dots on the connected outline.

of the set of shape vectors and the sum of the outer products of the texture vectors form two matrices. For each matrix we compute the eigenvectors and eigenvalues, where the selection of the significant eigenvectors is performed by examining how many of the significant eigenvalues must be retained to keep a percentage of the energy. The shape and texture of any image in the set may be represented using the following models:

$$\mathbf{z}(j) = \bar{\mathbf{z}} + \mathbf{P}_s \mathbf{b}_s(j), \tag{2}$$

$$\mathbf{g}(j) = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g(j), \tag{3}$$

where $\bar{\mathbf{z}}$ and $\bar{\mathbf{g}}$ are the mean shape and texture vectors. $\mathbf{P}_s$ and $\mathbf{P}_g$ are matrices formed from eigenvectors (obtained from the PCA operation). By varying the shape and texture parameter vectors $\mathbf{b}_s(j)$ and $\mathbf{b}_g(j)$ we are able to approximate the shape and texture of any of the existing images. Let $\mathbf{b}(j) = [\mathbf{W_s}\mathbf{b}_s^T(j), \mathbf{b}_g^T(j)]^T$ denote the concatenated vector of shape and texture parameters, where $\mathbf{W_s}$ is a diagonal matrix of weights to account for the difference between the shape and texture values. The required appearance parameters $\mathbf{c}(j)$ are then obtained by performing PCA on the set of vectors $\{\mathbf{b}(j)\}$ and forming a matrix $\mathbf{P}_c$ from a certain number of the resulting eigenvectors and applying this to $\mathbf{b}(j)$ as follows.

$$\mathbf{c}(j) = \mathbf{P}_c^T \, \mathbf{b}(j) \tag{4}$$

Thus $\mathbf{c}(j)$ is a vector of appearance parameters describing shape and texture of a speaker's mouth region at time j. Given a set of appearance parameters $\mathbf{c}(j)_{1 \leq j \leq T, j \in \mathbb{N}}$ sampled over time, we can model their dynamical changes over time using an HMM. HMMs have been used extensively in the past to model the dynamics of speech (e.g. [10]) and more recently to model joint audio-visual features [5]. For training an HMM, we use the standard Baum-Welch algorithm [10], which gives us the model $\lambda = (A, B, \pi)$, where π is a vector of the initial state probabilities, A is the state transition matrix and B is the state probability distribution.

The task is to determine if a person is speaking or silent in a given period of time, *i.e.* in a given sequence of appearance parameters. For this we calculate the likelihood that a sequence of appearance parameters is generated by our HMM λ . We calculate the likelihood $P(\mathbf{O}|\lambda)$ for a sequence of consecutive frames $\mathbf{O}$ ($\mathbf{O} = \mathbf{c}(t_k) \dots \mathbf{c}(t_l)$), where the number of frames between k and l is unchanged for all sequences. Each observation $\mathbf{O}$ will generate an associated likelihood value P. The early experiments showed it was necessary to filter the likelihood values to remove minor false detections. For this purpose we use a simple median filter (of length 25) to smooth the output and reduce false detections, where the filtered likelihood is denoted P_f (see Figure 6). We compare each value of P_f to a threshold value β (the value of β is found experimentally).

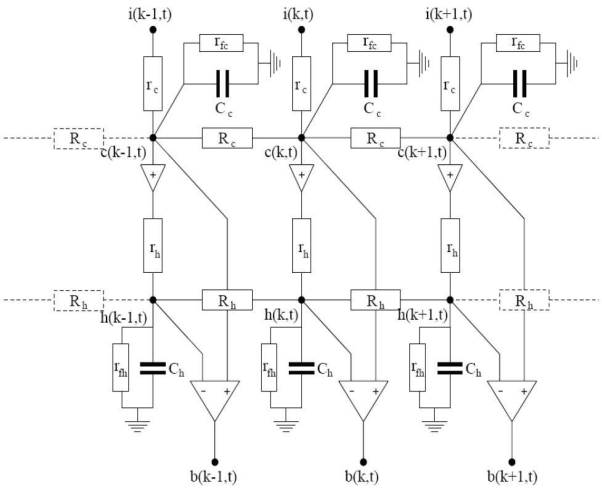


Figure 2: Mono-dimensional electrical scheme to model human retina [11].

If $P_f < \beta$ then the current sequence of frames is classified as speech, if $P_f > \beta$ then the sequence of frames is classified as non-speech.

3. Visual Voice Activity Detection based on a Retinal Filtering

In [8], Sodoyer *et al.* proposed a V-VAD based on the movement of the lips. Indeed, in silence sections, the lip-shape variations are generally quite small. On the contrary, during speech sections, these variations are commonly much stronger. However, this V-VAD requires the lips to be painted blue so that the internal height and width of the lips can be extracted by a chroma-key system. In this section, we propose a different and possibly simpler but equally efficient approach for the detection of the movement of the lips. As explained below, this new method requires no *a priori* information: it applies a retinal filter to each image and calculates the change in energy to classify voice activity.

Considering the lips region (which can be obtained manually or automatically), the first stage of this V-VAD is an enhancement of the contours of the lips based on a spatio-temporal filter which models the human retina behavior [11, 12]. This filter can be modelled by a mono-dimensional electrical scheme (Fig. 2) whose transfer function is given by

$$G(z_s, f_t) = \frac{1}{1 + \beta_c + \alpha_c \left(-z_s^{-1} + 2 - z_s\right) + \imath 2\pi f_t \tau_c} \times \frac{\beta_h + \alpha_h \left(-z_s^{-1} + 2 - z_s\right) + \imath 2\pi f_t \tau_h}{1 + \beta_h + \alpha_h \left(-z_s^{-1} + 2 - z_s\right) + \imath 2\pi f_t \tau_h} \tag{5}$$

where $\alpha_c = r_c/R_c$, $\beta_c = r_c/r_{fc}$, $\tau_c = r_c\, C_c$, $\alpha_h = r_h/R_h$, $\beta_h = r_h/r_{fh}$ and $\tau_h = r_h\, C_h$ (the component values relate to those shown in Figure 2). The resulting enhanced image denoted $B(t)$ is obtained from the original image of the lips region $I(t)$ by separately applying filter (5) on each row and column of $I(t)$. To detect the motion of the lips over time, we compute the following temporal derivation image

$$\Delta R(t) = \left| \left|R(t)\right|^2 - \left|R(t-1)\right|^2 \right|, \tag{6}$$

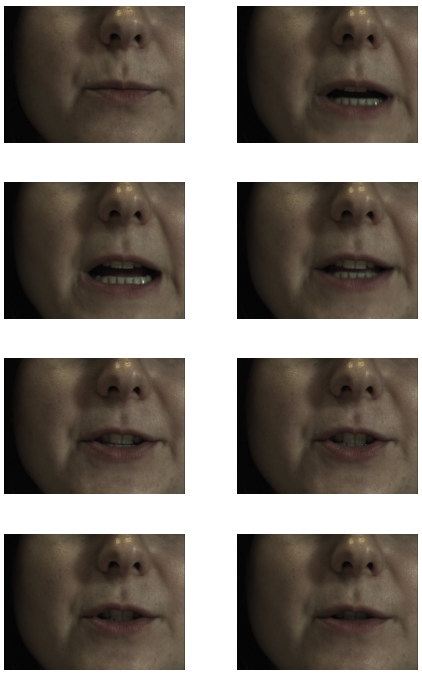
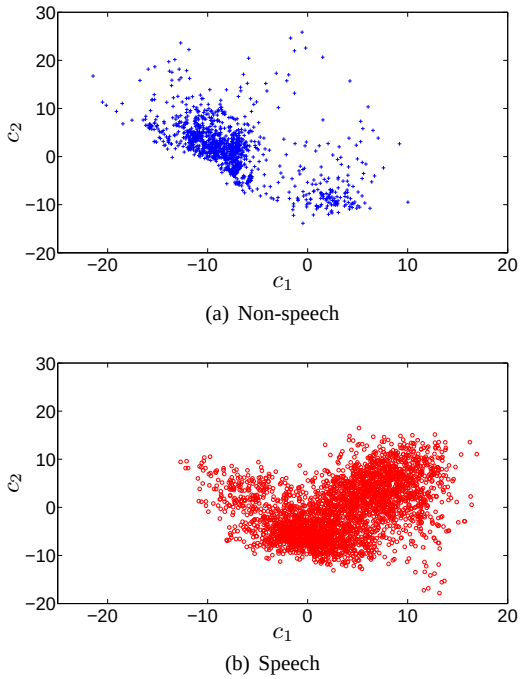


Figure 3: Frames from the dataset of the speaker saying the word 'much'. Frames read, top to bottom, left to right.



(a) Non-speech

(b) Speech

Figure 4: Distribution of the first two dimensions of non-speech (Fig. 4(a)) and speech (Fig. 4(b)) appearance parameters.

where $R(t)$ is the image obtained by the windowed two dimensional Fourier transform of the enhanced contours image $B(t)$. The change in energy is then obtained by the mean value of $\Delta R(t)$:

$$v(t) = \frac{1}{N_r}\frac{1}{N_c}\sum_{i=0}^{N_r-1}\sum_{j=0}^{N_c-1}\Delta R_{ij}(t), \tag{7}$$

where $\Delta R_{ij}(t)$ is the *(i,j)-th* pixel of image $\Delta R(t)$, N_r and N_c are the numbers of rows and columns of $\Delta R(t)$ respectively. The *t-th* input frame is classified as silence if $v(t)$ is lower than a threshold Λ and it is classified as speech otherwise. However, direct thresholding of $v(t)$ does not provide optimal performance: for instance, the speaker's lips may not move during several frames, while he is actually speaking. Thus, we smooth $v(t)$ by combining T_F consecutive frames

$$V(t) = \sum_{l=0}^{T_F-1} h^l\, v(t-l) \tag{8}$$

where h is a number between 0 and 1. Finally, the *t-th* frame is classified as silence if $V(t)$ is lower than a threshold Λ ($V(t) < \Lambda$) and speech otherwise ($V(t) \geq \Lambda$).

4. Numerical experiments

In this section, we first describe the database and the visual features used to conduct the numerical evaluation. The reason for recording our own database is that we are unaware of any existing audio-visual database where there is significant lip motion during silence sections of continuous speech.

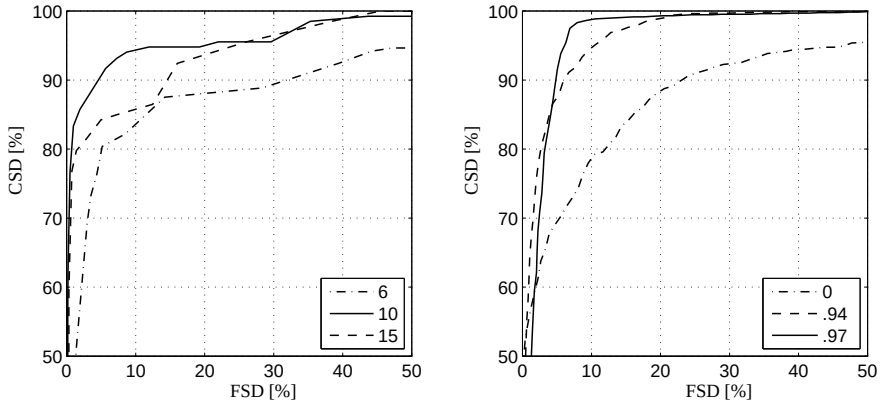
4.1. Audio Visual Corpus

The corpus used in these experiments consists of a single speaker reciting a well known poem in English. It consists of approximately 2.5 minutes of audio and video synchronously recorded. The video was recorded at 30fps giving 4400 useable video frames and the audio was sampled at 44.1KHz. The resolution of each frame of video is 640×480. To test rigorously the capabilities of both V-VAD methods, the speaker's lips during the silence periods were not always stationary. In fact, during the silence periods the speaker purposefully performed complex movements (*e.g.* smiling, biting lips and licking lips). Indeed, in spontaneous speech, people regularly perform natural movements such as those listed above during silence phases. Several example images from the dataset are shown in Fig. 3

4.2. Visual Features

The active appearance model described in Section 2 produces 400 dimensional vectors $\mathbf{c}(t_j)$, which are too large for numerical calculations. To reduce the dimensionality, $\mathbf{P}_c$ is only composed of the parameters associated with the N most important eigenvalues. However, there is a large overlap between the speech and non speech features (Fig. 4). Thus, N is a trade off between the size of the appearance parameter vector and the ability to separate speech and non-speech events. In this experiment, we retained the ten first eigenvectors which contained 75% of the original appearance energy. The HMM was trained solely using non-speech lips movements, where the training data consisted of the appearance parameters for 600 frames of video.

To apply the retinal filtering method described in Section 3, first the lips region $B(t)$ was extracted from $I(t)$ resulting in



(a) AAM (b) Retinal filtering

Figure 5: ROC curves. Fig. 5(a): AAM based method, the legend indicates the number of consecutive frames. Fig. 5(b): retinal filtering based method, the legend indicates the integration parameter h (8), with the classical convention $0^0 = 1$.

an image of 325×200 pixels. To obtain $R(t)$ in (6) the two-dimensional Fourier transform was applied on $B(t)$ using a Hamming window of 325×200 pixels with no zero-padding. Finally, the smoothing filter (8) was applied on $T_F = 20$ consecutive frames.

4.3. Results

Figure 5 shows the Receiver Operating Characteristics (ROC) curves for the two presented methods. They represent the ratio of correct silence detection to false silence detection. The correct silence detection (CSD) is defined as the ratio between the number of actual silence frames correctly detected as silence ($N_{\mathrm{Sil|Sil}}$) and the number of actual silence frames (N_{Sil}):

$$\mathrm{CSD} = \frac{N_{\mathrm{Sil|Sil}}}{N_{\mathrm{Sil}}}. \tag{9}$$

The false silence detection (FSD) is defined as the ratio between the number of actual speech frames detected as silence ($N_{\mathrm{Sil|Spe}}$) and the number of actual speech frames (N_{Spe}):

$$\mathrm{FSD} = \frac{N_{\mathrm{Sil|Spe}}}{N_{\mathrm{Spe}}}. \tag{10}$$

The ROC curve was produced by varying the thresholds β or Λ between the maximum and the minimum values of P_f and $V(t)$ (8) for the respective methods. One can see that the performance of the V-VAD is dependent upon the integration (8) (Fig. 5(b)) or the window size (Fig. 5(a)). A large window size (e.g. 15 frames) provides poor results since it is less likely that the model generates such a large sequence of frames. Similarly, a short window (e.g. 6 frames) is not satisfactory since there is not enough data to accurately classify. On the other hand, choosing a correct window size (e.g. 10 frames) or a correct integration parameter h (e.g. 0.97) provides reasonable performance: the two methods are able to achieve a CSD of 90% for a FSD of 5%.

Finally, Fig. 6 shows the temporal results of both methods (*i.e.* silence probability obtained from the AAM based method (Section 2) and video parameter (7) (Section 3)). One can see that the both proposed methods are able to easily discriminate the silence phases with no movements or short movements (e.g. between 30s and 40s or between 130s ans 150s). Moreover, even if it is (obviously) more difficult to detect silence phases with complex movements, the proposed silence probability obtained from the AAM based method or the proposed video parameter (8) still highlight such silence phases providing quite good silence detection performances as shown in Fig.5.

5. Conclusion

Both of the novel methods presented herein obtain similar performances for V-VAD. They both obtain low false detection rates for high true detection rates. One of the observations made during the experiments was that due to the use of *a-priori* information the AAM approach was more consistent with the detection of the non-speech sections containing complex lip movements. Similarly, the retinal filter was more consistent in the detection of non-speech where the lips show less motion. However it should be noted that the reliance on *prior* information restricts the AAM method to be person specific (a generic method is currently being investigated). Finally, the retinal filtering has a much lower computational cost compared to the AAM method. The results indicate that each method has its strengths and weaknesses in different areas, and this leads us to the conclusion that combining the two methods will result in an increase in performance, which is the subject of future research.

6. References

[1] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.

[2] G. Potamianos, N. Chalapathy, J. Luettin, and I. Matthews, "Chapter 10: Audio-Visual Automatic Speech Recognition: An Overview," *In Audio-Visual Speech Processing, MIT Press*, September, 2005.

[3] L. Girin, J.L. Schwartz, and G. Feng, "Audio-Visual Enhancement of Speech in Noise," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 3007–3020, 2001.

[4] B. Rivet, L. Girin, and C Jutten, "Solving the Indetermi-

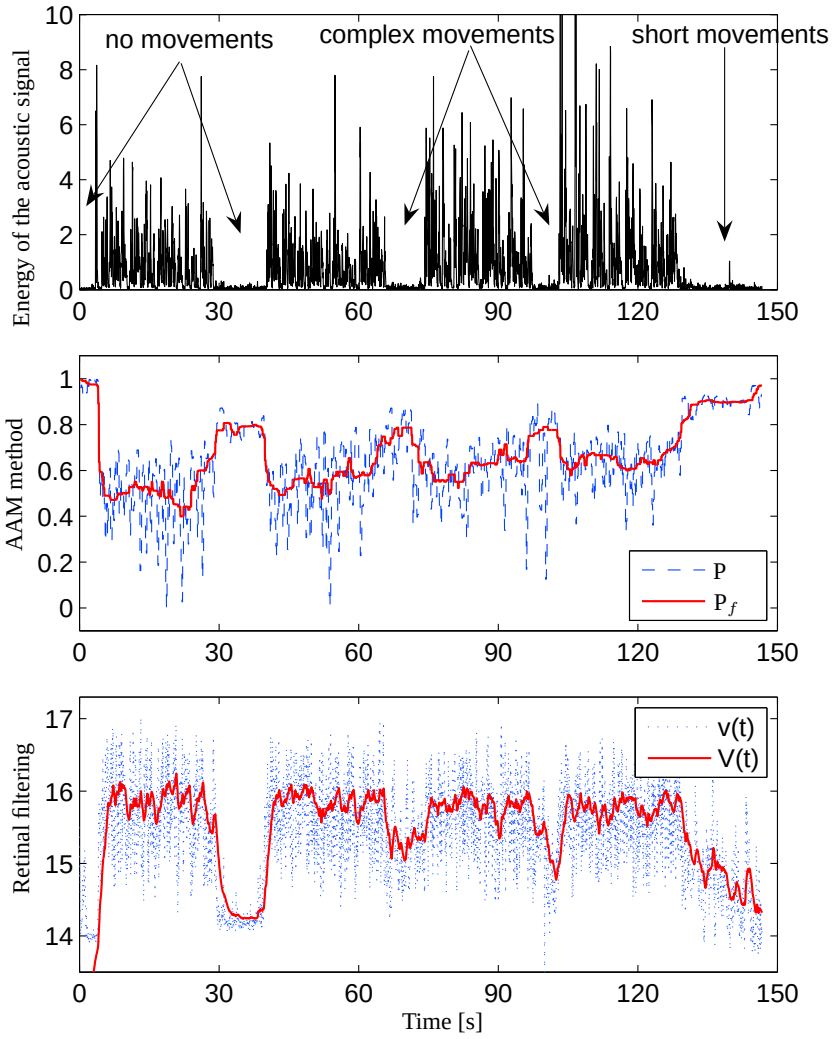


Figure 6: Temporal results. From top to bottom : energy of the acoustic signal, silence probability obtained from the AAM based method, and finally logarithm of video parameters (7) and (8) with $h = .97$.

nations of Blind Source Separation of Convolutive Speech Mixtures," in *ICASSP*, Philadelphia, 2005, pp. 533–536.

[5] A. Aubrey, J. Lees, Y. Hicks, and J. Chambers, "Using the Bi-modality of Speech for Convolutive Frequency Domain Blind Source Separation," in *IMA 7th International Conference on Mathematics in Signal Processing*, December 2006.

[6] G. Iyengar and C. Neti, "A Vision based Microphone Switch for Speech Intent Detection," in *Recognition, Analysis and Tracking of Face and Guestures in real time systems workshop, ICCV*, Vancouver, Canada, 2004.

[7] P. Liu and Z. Wang, "Voice Activity Detection Using Visual Information," in *ICASSP*, Montreal, Canada, 2004.

[8] D. Sodoyer, B. Rivet, L. Girin, J.L. Schwartz, and C. Jutten, "An Analysis of Visual Speech Information Applied to Voice Activity Detection," in *ICASSP*, Toulouse, France, 2006, pp. 601–604.

[9] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[10] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] J. Hérault and W. Deaudot, "Motion Processing in the Retina: about a Velocity Matched Filter," in *European Symposium on Artificial Neural Networks (ESANN)*, Brussels, Belgium, April 1993, pp. 129–136.

[12] A. Benoit and A. Caplier, "Motion Estimator Inspired from Biological Model for Head Motion Interpretation," in *WIAMIS*, Montreux, Switzerland, 2005.