Visual Lip Activity Detection and Speaker Detection Using Mouth Region Intensities

Spyridon Siatras,† Nikos Nikolaidis,† Michail Krinidis† and Ioannis Pitas†

Abstract*— *In this paper, we introduce a novel approach for lip activity detection and speaker detection, using solely visual information. The main idea in this work is to apply signal detection algorithms to a simple and easily extracted feature from the mouth region. We argue that the increased average value and standard deviation of the number of pixels with low intensities that the mouth region of a speaking person demonstrates can be used as visual cues for detecting visual speech. We then proceed in deriving a statistical algorithm that utilizes this fact for the efficient characterization of visual speech and silence in video sequences. Furthermore, we employ the lip activity detection method in order to determine the active speaker(s) in a multi–person environment.

[1]

I. Introduction

Speech analysis has been an area of extensive research in recent years, demonstrating impressive growth and significant applications. At first, only the audio information was being exploited, however visual cues have been also incorporated, providing supplementary information in the analysis process.

In this paper, we present a statistical approach for lip activity detection and speaker detection in videos, based solely on visual information. Lip activity detection can be thought of as the visual counterpart of voice activity detection. Naturally, lip activity can be directly related to visual speech, although not all lip movements correspond to visual speech, yawning and chewing being two such examples.

The proposed method aims to distinguish frames that depict visual speech from those that depict visual silence. Furthermore, speaker detection is achieved by applying the lip activity detection algorithm to every detected face of each video frame. The main idea in our work is to apply signal detection algorithms to a simple feature that is easily extracted from the mouth region intensities, leading to a fast and robust method. The proposed system exhibits a wide range of potential applications in areas such as human-computer interaction, video indexing and multimedia retrieval. This paper is an enriched version of the work presented in [4]. In this paper a much more detailed and analytical description of the statistical framework employed by the proposed algorithm is provided. Moreover, the efficiency of the proposed algorithm is verified by conducting more extensive experiments and by comparing it to a state of the art approach. Furthermore, an extension of the proposed algorithm for the efficient detection of the active speaker(s) in a multi-person environment is introduced and tested in this paper.

The main research topic in the area of speech analysis using visual information is automatic visual or audio-visual speech recognition [5], [6], [7]. However, only a few works [9], [10] address the same problem as in our work, i.e. characterizing the frames of a video sequence as containing speaking persons or not using only visual information.

In [8], a system that detects the user's intent to speak considering both the audio and visual information is proposed. First, a frontal face is detected, and then the audio energy and the shape of the speaker's mouth are used to yield an indication of speech activity. The system achieved a maximum correct classification (of silence or speech visemes) equal to 70.27%.

A method for visual detection of silence sections is proposed in [9]. The visual information consists in the time trajectory of basic lip contour geometric parameters, namely the interolabial width and height. These parameters are analyzed in order to characterize the possible differences on visual patterns between silence and non-silence sections for a given speaker, leading to a dynamic model. The proposed method achieved a correct silence detection of 80%, at a false silence detection of 5%.

In [10], Principal Component Analysis (PCA) is applied on the intensities and the first order intensity differences of the pixels in the mouth region, that is derived in each frame using PCA and template matching. The outcome of PCA is used as the feature vector of the algorithm. A single Gaussian is used to model feature vectors of the non–voice frames whereas a mixture of two Gaussians is used to model the feature vectors of voice frames. The parameters of the Gaussians are estimated from feature vectors derived from training data. Once training is completed, the speech/non–speech decision is taken by evaluating the likelihood of the feature vector of each frame for both distributions (voice/non–voice). In the experiments reported in this paper, the method exhibits a frame error rate (incorrect classification of the video frames) equal to 3.37%.

Concerning speaker detection and localization, several methods try to address the problem using only audio information, for instance, using microphone array processing to estimate the direction of arrival from one or more sources [12]. A number of methods proceed into integrating visual information with audio information for the same task [13], [14], [15]. In the vast majority of the speaker detection works, the authors consider only video data where strictly one person is speaking at a time; the algorithms have not been tested

†Aristotle University of Thessaloniki, Department of Informatics, Box 451, 54124 Thessaloniki, Greece, *email:* {*siatras,nikolaid, mkrinidi, pitas*}*@aiia.csd.auth.gr* Fax/Tel ++ 30 231 099 63 04, http://www.aiia.csd.auth.gr/.

This work was supported by the project PYTHAGORAS II "Efficient techniques for information organization, browsing, and retrieval in multimedia" funded by the Greek Ministry of Education and the European Union.

on simultaneous speech cases. The method proposed in this paper has been successfully tested in such simultaneous speech cases.

II. Motivation

Our method for lip activity detection is based on the significant variability of the intensity values of the mouth region in the case of a speaking person. The opening of the mouth produces a radical increase in the number of mouth pixels with low intensity values. This is due to the exposure of a part of the interior of the mouth, which is usually in shade. Since the pronunciation of the majority of phonemes involves an open mouth, it is obvious that during visual speech, the number of low intensity pixels is, in general, high. However, when a person is speaking, the percentage of the oral cavity revealed is related to the pronounced phoneme, and the pronunciation of certain phonemes even involves closed lips. Hence, there is a fluctuation in the number of low intensity pixels. On the other hand, when there is no lip activity (i.e. no speech) the lips are most probably closed and, therefore, there is no increase in the low intensities of the mouth region and no fluctuation of these intensities. We, thus, argue that the increase and the fluctuation of the number of mouth region pixels exhibiting low intensity values can indicate lip activity. This fact is used in this work for the visual detection of speech.

We denote by $x[n]$ the number of pixels of the mouth region at the n-th video frame whose grayscale value is below an intensity threshold T. More specifically, if $H_n(i),\ i = 0, \ldots, 255$ is the grayscale histogram of the n-th frame then:

$$x[n] = \sum_{i=0}^{T} H_n(i) \tag{1}$$

Thus, for a video sequence that consists of N frames, we create a discrete sequence $x[n],\ n \in [0, N-1]$. In order to normalize the value of $x[n]$ for different sizes of the bounding box of the mouth region, we divide it with the area (in pixels) of the bounding box that encloses the mouth. The methodology for calculating the intensity threshold is described in section III.

In Figure 1 we depict $x[n]$ for a video sequence displaying a person who is silent at first, and then is speaking for a number of frames. It is obvious that $x[n]$ obtains much higher values when the person is speaking. Moreover, $x[n]$ exhibits a large deviation of its values in the speaking interval, due to the movement of the lips that affect the visible area of the mouth cavity. In the silent frames, the values of $x[n]$ are much lower (in average) and exhibit a small deviation from their mean value. The proposed algorithm makes use of the increased values and the large deviation of $x[n]$, in the visual speech intervals, in contrast to the low values and the small deviation of the samples of $x[n]$ in the silent intervals.

III. Lip Activity Detection Algorithm

Before proceeding with the detailed description of the proposed algorithm, a short overview of the method will be provided.

Prior to applying the proposed algorithm, we detect the face in the video sequence and then assign at each frame a bounding box encompassing the mouth region. Naturally, the lip activity detection system is prone to suffer from errors of the face and mouth region detectors.

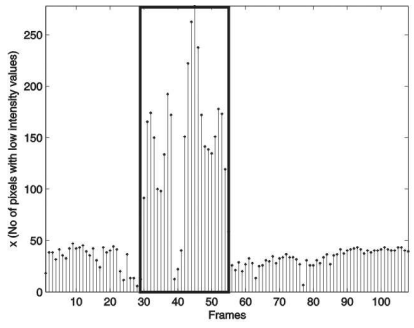


Fig. 1. Plot of the number of low grayscale intensity pixels $x[n]$ for certain frames of a video sequence. The rectangle encompasses the frames where the person is speaking.

The face is detected using the method described in [16], which is one of the most powerful and widely used face detectors proposed in the literature. For the detection of the mouth region within the detected face, the algorithm proposed in [19] was used.

The lip activity detection system is based on statistical algorithms, used in signal detection applications. The speaking and non-speaking intervals are determined by applying on $x[n]$ an *energy detector* and an *averager* [17], using a sliding window which moves frame-by-frame, spanning the whole video sequence. The outcomes of the detectors are compared to their respective thresholds in order to determine the presence of lip activity in each window. The thresholds are computed according to the Neyman-Pearson theorem for each video sequence and depend on the variance of $x[n]$ in the silent frames. A detailed description of the algorithm is provided below.

The first step of the algorithm is to compute the threshold T in (1). Since video excerpts from different movies, TV programs, or personal cameras are acquired in diverse lighting conditions, we do not apply a global intensity threshold for all videos, but a video-specific threshold. The intensity threshold is calculated as half of the average intensity μ_1 of the mouth region in the first frame of the face video sequence. If, for this value, the number $x[0]$ of pixels whose intensity is below T is zero, T is increased in small steps until $x[0]$ obtains a non-zero value. Once T is determined, $x[n]$ is computed for every frame $n = 0, ..., N-1$ using (1) and normalized by the number of pixels of the mouth area.

A. Statistical framework

The aim of the proposed method is to decide between two possible hypotheses: lip activity present versus no lip activity. This hypothesis testing problem can be translated into a problem of signal detection within noise that involves the sequence $x[n]$. We consider as noise $w[n]$ the fraction of $x[n]$ that corresponds to the area of the lips when the mouth is closed, and as signal $s[n]$ the contribution to the value of $x[n]$ of the area of the mouth interior that is revealed when a person is speaking. Hence, in both hypotheses noise is present,

whereas when the person is speaking there is signal present as well. Consequently, our hypotheses can be stated as follows:

$$H_0 \;\; (no\ lip\ activity) : x[n] = w[n],\;\; n = 0, ..., N-1$$
$$H_1 \;\; (lip\ activity) : x[n] = s[n] + w[n],\;\; n = 0, ..., N-1$$

Both our signal $s[n]$ and noise $w[n]$ samples are obtained as the sum of the number of pixels whose intensity is below T (see (1)). If $H_n(i)$ are assumed to be independent random variables, we can assume, according to the central limit theorem, that both $s[n]$ and $w[n]$ follow Gaussian distributions. Furthermore, we assume that both $s[n]$ and $w[n]$ are independent for the various values of n and that the noise $w[n]$ is zero mean. Thus: $w[n] \sim N(0, \sigma^2)$, $s[n] \sim N(\mu_s, \sigma_s^2)$. Therefore, in order to discern between visual speech and silence, we can apply detection theory principles for detecting a Gaussian random signal in Gaussian noise. We have to note that in reality the distribution of $w[n]$ is not zero mean. However, we can convert it to zero mean by estimating the mean value of the noise samples, as will be presented in the following subsection, and subtract it from $w[n]$. The above assumptions are necessary in order to proceed in developing our statistical framework, which is experimentally verified as efficient and reliable.

We define the $N \times 1$ random vector $\mathbf{x} = [x[0], x[1], ..., x[N-1]]^T$. The *Neyman-Pearson theorem* states that in order to maximize the probability of signal detection P_D for a given probability of false alarm $P_{FA} = \lambda$, we decide for H_1, if the likelihood ratio $L(\mathbf{x})$ is larger than a threshold γ:

$$L(\mathbf{x}) \;=\; \frac{p(\mathbf{x}; H_1)}{p(\mathbf{x}; H_0)} > \gamma \tag{2}$$

where the threshold γ is found from:

$$P_{FA} = \int_{\mathbf{x}: L(\mathbf{x}) > \gamma} p(\mathbf{x}; H_0) d\mathbf{x} = \lambda \tag{3}$$

and $p(\mathbf{x}; H_0)$, $p(\mathbf{x}; H_1)$ are the multivariate probability density functions of $\mathbf{x}$ under the respective hypotheses. From our modelling assumptions, $\mathbf{x} \sim N(\mathbf{0}, \sigma^2 \cdot \mathbf{I})$ under H_0 and $\mathbf{x} \sim N(\mu_s \cdot \mathbf{1}, (\sigma_s^2 + \sigma^2) \cdot \mathbf{I})$ under H_1, where $\mathbf{0}$ and $\mathbf{1}$ denote the all-zero and all-one vectors respectively and $\mathbf{I}$ denotes the identity matrix. By substituting these density functions in (2), the likelihood ratio becomes:

$$L(\mathbf{x}) \;=\; \frac{\frac{1}{[2\pi(\sigma_s^2+\sigma^2)]^{\frac{N}{2}}} \exp\left[-\frac{1}{2(\sigma_s^2+\sigma^2)} \sum_{n=0}^{N-1} (x[n] - \mu_s)^2\right]}{\frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]\right]} > \gamma \tag{4}$$

We then compute the log-likelihood ratio by taking the logarithm of (4), and we incorporate the non-data terms (i.e., the terms of the sum that are not related to $x[n]$) in the threshold. Thus, the following expression results:

$$l(\mathbf{x}) \;=\; -\frac{1}{2(\sigma_s^2 + \sigma^2)} \sum_{n=0}^{N-1} (x[n] - \mu_s)^2 + \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] > \gamma' \tag{5}$$

By further processing of the above equation and incorporating its non-data terms in the threshold, the test statistic $T(\mathbf{x})$ is found:

$$T(\mathbf{x}) \;=\; N\mu_s \cdot \frac{1}{N} \sum_{n=0}^{N-1} x[n] + \frac{\sigma_s^2}{2\sigma^2} \cdot \sum_{n=0}^{N-1} x[n]^2 > \gamma'' \tag{6}$$

This test statistic is a linear combination of an averager:

$$T_1(\mathbf{x}) = (1/N) \sum_{n=0}^{N-1} x[n] \tag{7}$$

which attempts to discriminate between the two hypotheses on the basis of the sample mean, and an energy detector:

$$T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n] \tag{8}$$

which attempts to discriminate on the basis of the variance, i.e.:

$$T(\mathbf{x}) = aT_1(\mathbf{x}) + bT_2(\mathbf{x}) \tag{9}$$

In order to use (6) or, equivalently, (9) one needs to estimate the variances σ, σ_s of the noise and the signal as well as the mean value of the signal μ_s. Since such an estimation is difficult to achieve, we have chosen to employ the two detectors separately. In that way, as it will be shown below, one needs to estimate only σ, since the terms $a = N\mu_s$ and $b = \frac{\sigma_s^2}{2\sigma^2}$ in (9) can be incorporated in the respective thresholds.

By applying these two detectors, we can detect lip activity by exploiting the attributes that a speaking person demonstrates. In order to determine the presence of lip activity, both criteria – increased values and large variance of $x[n]$ – have to be satisfied. The two detectors are applied to a sliding window, consisting of N frames, which moves frame-by-frame spanning the whole video sequence. At each window, both detectors are compared to their respective thresholds, γ_1 and γ_2, which are computed according to the analysis that follows. When the outcomes of both detectors exceed their respective thresholds, i.e. when:

$$T_1(\mathbf{x}) > \gamma_1 \;\; AND \;\; T_2(\mathbf{x}) > \gamma_2 \tag{10}$$

lip activity is detected.

The averager (7) is used to detect a DC level in the presence of zero mean Gaussian noise. The detector compares the sample mean to a threshold. The value of the threshold is found by constraining P_{FA}. Being the sum of Gaussian random variables $x[n]$, $T_1(\mathbf{x})$ is also Gaussian. Furthermore, it can be easily proven that, under hypothesis H_0, $T_1(\mathbf{x})$ has mean equal to zero, and variance equal to σ^2/N. Hence, the probability of false alarm of the averager is given by:

$$P_{FA} = Pr\{T_1(\mathbf{x}) > \gamma_1; H_0\} = Q(\frac{\gamma_1}{\sqrt{\sigma^2/N}}) \tag{11}$$

where Q is the right tail probability of a Gaussian random variable. Hence, the threshold can be found from:

$$\gamma_1 = \sqrt{\frac{\sigma^2}{N}} Q^{-1}(P_{FA}) \tag{12}$$

where Q^{-1} is the inverse right-tail probability.

The energy detector (8) is used to detect a random Gaussian signal in zero mean Gaussian noise. The detector computes the energy of the data samples and compares it to a threshold. If the signal is present, the data energy is large. Again, the value of the threshold is found by constraining P_{FA}. The probability of false alarm can be found by noting that under hypothesis H_0, $T_2(\mathbf{x})/\sigma^2$ is distributed according to a *chi-squared* distribution. The right-tail probability function of a chi-squared random variable is expressed as $Q_{\chi^2_N}(x)$ [18]. Therefore, the probability of false alarm is

$$\begin{aligned} P_{FA} &= Pr\{T_2(\mathbf{x}) > \gamma_2; H_0\} \\ &= Pr\{\frac{T_2(\mathbf{x})}{\sigma^2} > \frac{\gamma_2}{\sigma^2}; H_0\} = Q_{\chi^2_N}(\frac{\gamma_2}{\sigma^2}) \end{aligned} \quad (13)$$

Thus, the threshold is given by

$$\gamma_2 = \sigma^2 Q^{-1}_{\chi^2_N}(P_{FA}) \quad (14)$$

However, we have not completely resolved the problem yet, since in our case the noise standard deviation σ, which is involved in threshold determination (12), (14), and the noise mean, required to convert the noise into a zero mean process, are not known a priori.

B. Estimation of noise statistics

In the preceding analysis we have assumed a zero mean Gaussian noise $w[n]$. However, as already mentioned, $w[n]$ is not zero mean. Thus, we have to estimate the actual mean value μ of the noise and subtract it from the noise samples. Furthermore, we have concluded that the noise standard deviation σ is a prerequisite for the computation of the thresholds γ_1 and γ_2. In order to find the actual values of the noise statistics, we apply an estimation algorithm based on the detection theory principles we have presented.

The noise statistics estimation algorithm focuses on distinguishing efficiently the *signal and noise* samples (hypothesis H_1) from the *noise only* samples (hypothesis H_0), and then calculating the actual noise mean μ and standard deviation σ. This is achieved iteratively, by applying the averager and the energy detector to our data sequence, each time with refined estimates of the noise statistics, until they converge to their final values. This approach, referred to as an *estimate and plug* detector [17].

The algorithm first computes initial estimates of μ and σ, in order to apply the detectors. The initial estimates are obtained by evaluating the sample mean and sample variance of the smaller 10% of $x[n]$, assuming that these values belong to the noise samples. Thereafter, we apply the detectors to our data set, employing the computed noise statistics. The detectors distinguish the noise only samples from the signal and noise samples and new noise statistics are calculated from the detected noise samples. This process is repeated until the difference between two consecutive estimates of σ is smaller than 10^{-2}.

C. Speaker Detection

As already mentioned, using the lip activity detection method we have presented, we can detect the active speakers in multi–speaker videos, solely from visual information. In order to do so, we apply the face detector so as to identify the persons that are present at each frame. Subsequently, the mouth detector is applied on each detected face. The outcome of the lip activity detection algorithm, for each mouth region, determines the active speakers. The proposed speaker detection, has been successfully tested in cases where more than one speakers are talking simultaneously. Such cases have not been considered in the experimental evaluation of other approaches presented in the literature.

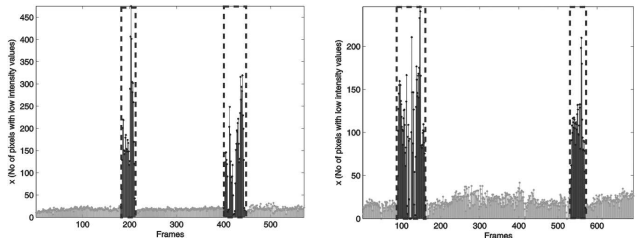
IV. Experimental Results

In order to evaluate the performance of the lip activity detection system, we have tested it in 49 short video sequences consisting of a total of 20429 frames, displaying individuals that exhibit speaking and silent intervals. In particular, our test data consist of 8915 speaking and 11514 silent frames, from 13 individuals. The video sequences are recorded from news programs and talk shows. In general, the selection of video sequences for the creation of our test data was performed so as to ensure that the lips and their deformations are visible. The faces displayed in these videos were chosen to be predominantly frontal, with dimensions ranging from 100×145 to 195×315 pixels. In order to determine the ground truth, each frame has been marked after visual inspection as corresponding to visual speech or silence.

In our experiments, we have constrained P_{FA} to 1% and we have applied the detectors to a data window consisting of 5 frames using the thresholds derived from (12) and (14). The window was moving frame-by-frame, spanning the whole data sequence. The decision obtained for each window position characterized its central frame.

The probabilities of detection P_D (the ratio of the correctly detected visual speech frames to the total number of visual speech frames) and false alarm P_{FA} (the ratio of the silence frames that were mis-detected as speech ones, to the total number of silence frames) were used as performance indicators for our system. The proposed system achieved very good results, namely P_D equal to 98.93%, and P_{FA} equal to 2.16%. Most of the false alarms were produced by the speaker's mouth opening, either to breathe or to establish his intent to speak. The fact that the achieved P_{FA} is very close to the theoretically imposed value (1%), as well as the overall very good performance of the method, justify that the assumptions that have been adopted for the theoretical derivations are, in general, valid. In Figure 2, we depict the lip activity detection algorithm outcomes for two video sequences.

In another set of experiments, a state-of-the-art voice activity detection algorithm proposed in [10] was compared against the proposed approach. A brief description of the algorithm, that will be mentioned from now on as PCA-GMM, is provided in Section I. The algorithm was tested on the same data as the proposed algorithm. The outcome of the face and mouth detectors which were used in the proposed method (scaled by a factor of 1.5), was also used to define the ROI where PCA-GMM is operating. The parameter values that were proposed in [10] as being the ones that provide



(a) Video sequence 1 (b) Video sequence 2

Fig. 2. Lip activity detection. Dark values: Visual speech frames, bright values: silence frames. The dashed rectangle encloses the visual speeech frames as defined in the ground truth.

the best performance, were adopted in our experiments. More specifically, a 12-dimensional feature vector, resulting from the application of PCA to intensity and differences of intensity data was used, whereas a mixture of two Gaussians was used to model the voice data and a single Gaussian to model the non–voice data. In the training phase, 20% of the total number of video frames was used in order to train the voice and non–voice Gaussian models. The rest of the frames were used in order to test the performance of the PCA-GMM. The achieved probabilities of detection and false alarms over the test data, were $P_D = 80.75\%$ and $P_{FA} = 15.32\%$ respectively, proving that the proposed algorithm performs significantly better than the PCA-GMM approach.

In a final set of experiments, the performance of the proposed algorithm in the task of speaker detection was evaluated on video clips depicting discussions, from news programs and TV talk shows, where two or three persons are present, and where (sometimes) more than one persons are speaking simultaneously. More specifically, we have tested our method in 14 short video sequences, for a total of 3452 frames. The total probability of error, defined as the ratio of the mis-detected frames (namely falsely determining the active speaker(s); detecting a speaker when no one is speaking; not detecting a speaker when someone is speaking) to the total number of frames, was used as performance indicator. Our system produced a very low probability of error equal to 1.16%, thus showing that the proposed method achieves very good active speaker detection accuracy.

V. Conclusions

A novel method for detecting lip activity and determining the active speakers in video sequences has been presented in this paper. The method uses the luminance information of the mouth region. The fact that the algorithm is based on a simple feature, that is easy to extract, is, in our opinion, a positive aspect of the algorithm that contributes to its robustness, since it does not require a complex feature extraction procedure (e.g. tracking of lip contours as in [9]) that might be prone to errors. Moreover, the adopted statistical signal detection approach and the fact that the statistics involved in this approach are evaluated from the data in a robust and adaptive way (without fixed parameters or thresholds) further increase the robustness of the algorithm. The proposed system has been tested in a number of video sequences with very good performance. Furthermore, we have applied our lip activity detection method in order to identify the active speakers in sequences having more than one speaker.

Naturally, the proposed system, despite being sufficiently robust to lighting variations, face size and orientation variations, cannot operate correctly under all conditions. Extremely poor lighting conditions and faces that deviate significantly from a frontal pose or are too far apart from the camera to provide sufficient information for the mouth area, would cause the algorithm to perform poorly.

References

[1] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Proc. NIPS 1994*, Denver, Colorado, USA, Nov. 28 - Dec. 3. 1994, pp. 851-858.

[2] A. MacLeod and A. Q. Summerfield, "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use," *British Journal of Audiology*, vol. 24, pp. 29-43, 1990.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.

[4] S. Siatras, N. Nikolaidis and I. Pitas "Visual speech detection using mouth region intensities," in *Proc. of European Signal Processing Conf. (EUSIPCO 2006)*, Florence, Italy, 4-8 September, 2006.

[5] P.S. Aleksic, J.J. Williams, Z. Wu, and A.K. Katsaggelos, "Audio-Visual Speech Recognition Using MPEG-4 Compliant Visual Features," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1213-1227, November 2002.

[6] M. Gordan, C. Kotropoulos and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," *Journal of Applied Signal Processing*, vol. 2002, no. 11, pp. 1248-1259, November 2002.

[7] M. I. Faraj and J. Bigun, "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1169-1175, September 2007.

[8] C. Neti, P. de Cuetos, and A. Senior, "Audio-visual intent-to-speak detection for human-computer interaction," in *Proc. ICASSP 2000*, Istanbul, Turkey, June 5-9. 2000, pp. 1325-1328.

[9] D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," in *Proc. ICASSP 2006*, Toulouse, France, May 14-19. 2006, vol. 1, pp. 601-604.

[10] P. Liu and Z. Wang , "Voice activity detection using visual information," in *Proc. ICASSP 2004*, Montreal, Canada, May 17-21. 2004, vol. 1, pp. 609-612.

[11] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces, " in *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103-108, January, 1990.

[12] D.H. Johnson and D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Prentice-Hall, Englewood Cliffs, N. J., 1993.

[13] M. Gurban and J. Thiran, "Multimodal speaker localization in a probabilistic framework," in *Proc. EUSIPCO 2006*, Florence, Italy, September 4-8. 2006.

[14] C. Zhang, P. Yin, Y. Rui, R. Cutler and P. Viola, "Boosting-Based Multimodal Speaker Detection for Distributed Meetings," in *IEEE 8th Workshop on Multimedia Signal Processing*, October 3-6. Victoria, BC, Canada, 2006.

[15] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran and M. Kunt, "Extraction of Audio Features Specific to Speech Production for Multimodal Speaker Detection," *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 63-73, January 2008.

[16] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137-154, May 2004.

[17] S. M. Kay, *Fundamentals of statistical signal processing, vol. II: detection theory*. Prentice-Hall, Englewood Cliffs, N. J., 1998.

[18] A. Papoulis, *Probability, random variables and stochastic processes*, McGraw-Hill, 2001.

[19] S. Asteriadis, N. Nikolaids, I. Pitas and M. Pardas, "Detection of facial characteristics based on edge information, " in *International Conference on Computer Vision Theory and Applications (VISAPP 2007)*, Barcelona, Spain, 2007.