VISUAL SPEECH DETECTION USING MOUTH REGION INTENSITIES

Spyridon Siatras, Nikos Nikolaidis, and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, 54124, Thessaloniki, Greece
phone: + (30) 2310-996361, fax: + (30) 2310-998453, email: siatras, nikolaid, pitas@aiia.csd.auth.gr
web: www.aiia.csd.auth.gr

ABSTRACT

In recent research efforts, the integration of visual cues into speech analysis systems has been proposed with favorable response. This paper introduces a novel approach for lip activity and visual speech detection. We argue that the large deviation and increased values 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 speech. We describe a statistical algorithm, based on detection theory, for the efficient characterization of speaking and silent intervals in video sequences. The proposed system has been tested into a number of video sequences with encouraging experimental results. Potential applications include speech intent detection, speaker determination and semantic video annotation.

1. INTRODUCTION

Speech analysis systems have attracted increased attention in recent research efforts. At first, the focus was solely on the audio information, however visual cues are currently being incorporated, providing supplementary information in the analysis process. In [6], the authors argue that a major improvement can be obtained by using joint audio-visual processing, compared to the sole processing of the audio information.

Indeed, seeing the face of a speaking person facilitates the intelligibility of the speech, particulary in noisy environments. Laboratory studies have shown that visual information allows a tolerance of an extra 4-dB of noise in the acoustic signal [7]. This is a significant improvement considering that each dB of signal-to-noise ratio is reflected into a 10-15% error reduction in the intelligibility of complete sentences [8].

In human-to-human interaction, lip-reading performance depends on a number of factors [6]. Viewing conditions affect the quality of the visual information. For instance, poor lighting causes difficulties in determining the mouth's shape. Furthermore, as the speaker and the listener move further apart, it becomes more difficult to observe important visual cues. Finally, the viewing angle has a major effect on the recognition process. Inevitably, these limitations are inherited into automatic visual speech analysis systems.

The main research topic in this area is automatic visual or audio-visual speech recognition (ASR) [9]. Methods for speech intent detection for human-computer interaction [10] and multi-modal determination of speaker location and focus [11] have been also proposed.

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.

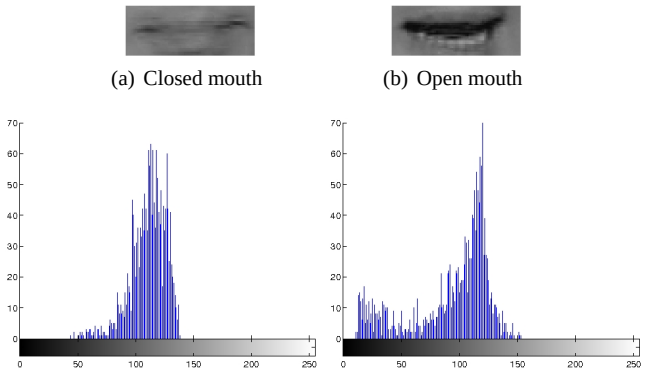


Figure 1: Increase in the number of low intensity pixels in the mouth region when mouth is open

In our work we present a statistical approach for visual speech detection, using mouth region intensities. Our method employs face and mouth region detectors, applying signal detection algorithms to determine lip activity. The proposed system can be used for speech intent detection and speaker determination in human-computer interaction applications, as well as in video telephony and video conferencing systems. It can also be used as a component in a dialogue detection system for movies and TV programs. Such a system can be useful in multimedia data management or semantic video annotation applications.

2. MOTIVATION

Our method is based on the significant variation of the intensity values of the mouth region that a speaking person demonstrates. Specifically, as it can be seen in Figure 1, the opening of the mouth produces a radical increase in the number of pixels with low intensity values. This is due to the exposure of a part of the oral cavity, which is revealed when a person is speaking. The intensity values which the oral cavity pixels possess, belong to the lower grayscale intensity range, since the oral cavity is usually in the shade. Therefore, we argue that a large number of mouth region pixels exhibiting low intensity values can indicate lip activity. This fact can be used for the visual detection of speech.

We denote by x the number of the low intensity pixels of the mouth region at a single video frame. In particular, x is the number of pixels of the mouth region whose grayscale value is below an intensity threshold t. Since video excerpts from different movies, TV programs, or personal cameras are acquired in diverse lighting conditions, we do not apply a

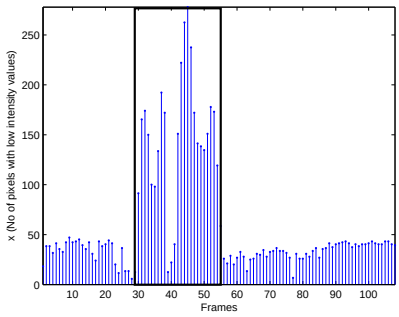


Figure 2: Distribution of the number of low grayscale intensity pixels of a video sequence. The rectangle encompasses the frames where the person is speaking

global threshold for all videos, but a video specific threshold, computed prior to the analysis of each video sequence. In order to normalize the value of x for different sizes of the bounding box of the mouth region, we divide its value with the area of the bounding box. Thus, for a video sequence that consists of M frames, we create a discrete sequence $x[n]$, $n \in [0, M-1]$.

In Figure 2 we depict $x[n]$ for a video sequence displaying a person that is silent at first, speaking for a number of frames – the frames included in the rectangle drawn in Figure 2 – and then silent again. It is obvious that $x[n]$ obtains much higher values when the person is speaking. Moreover, $x[n]$ exhibits a larger deviation of its values in speaking intervals, due to the moving lips that affect the visible area of the mouth cavity. For instance, at frame 39 $x[n]$ takes a very small value, even smaller than the values of some of the 'silent frames'. This is because at this particular instance the speaking person has his lips joined together to produce the letter "m". In the silent frames, the values are much lower (in average) and exhibit a small deviation from their mean value. The proposed algorithm exploits the attributes that a video sequence of a speaking person exhibits. In particular we make use of

- the increased values of $x[n]$
- the large deviation of $x[n]$

which are present at the sequence intervals where a person is speaking.

3. SYSTEM OVERVIEW

Our system consists of three parts:

- Face detection
- Mouth region detection
- Visual speech detection

The main focus of this paper is in describing our approach for visual speech detection. However, before applying our detection algorithm, we first have to detect the face in the video sequence under examination, and then assign at each frame a bounding box encompassing the mouth region of the detected face. The face detector we employ is based on the techniques presented in [2, 3, 4].

For the detection of the mouth region we use the technique described in [5] for eye detection, modified to detect mouth regions in facial images. In [5], each pixel is assigned the slope and the magnitude of the vector from the pixel to the closest edge point. Thus, a slope and a magnitude map are formed for each candidate region. Eye detection is performed by comparing these maps against the corresponding maps of an eye model, in a suitable space derived through PCA. In our case, a similar approach, employing a mouth model is applied for mouth region detection.

The visual speech detection system is based on statistical algorithms, used in signal detection applications. At first the intensity threshold t is determined, as half the average intensity of the mouth region in the first frame, and the number of pixels below it is computed. The intensity threshold is increased iteratively when it can not provide sufficient information about the intensity values of interest, i.e. when the threshold is low and the number of the selected pixels is inadequate. The speaking and non-speaking intervals are determined by applying an *energy detector* and an *averager* to 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 visual speech in each window. The thresholds are computed according to the Neyman-Pearson theorem for each video sequence and are depended on the distribution of the silent frames.

4. VISUAL SPEECH DETECTION ALGORITHM

The proposed method for the efficient determination of speaking and non-speaking intervals is based on statistical signal processing principles, incorporating detection theory algorithms. Our aim is to decide between two possible hypotheses; visual speech present versus no visual speech. We can translate our hypotheses into a problem of signal detection within noise. We consider as noise the value of x when the mouth is closed, i.e. the value corresponding to the area of the lips, and as signal the contribution of the area of the oral cavity that is revealed when a person is speaking to the value of x. Hence, in both hypotheses there is noise present (the pixels of the lip area) whereas when the person is speaking there is signal present as well. Consequently, our hypotheses can be stated as follows

H_0: Noise only (visual silence)
H_1: Signal and noise (visual speech)

Both our signal and noise samples are obtained as the sum of a number of pixels whose intensity is below t. Thus, according to the central limit theorem, we can consider that the data samples, $x[n]$, follow Gaussian distributions under both hypotheses. Therefore, in order to discern between visual speech and silence, we can apply the detection theory principles for detecting a Gaussian random signal in white Gaussian noise. We assume that the signal $s[n]$ is a Gaussian process with variance σ_s^2 and mean μ_s and the noise $w[n]$ is zero mean white Gaussian, with variance σ^2. We have to note that actually 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 presented in the following subsection. Consequently, our detection problem can be described as

$$H_0 : x[n] = w[n],\ n = 0, 1, ..., N-1$$
$$H_1 : x[n] = s[n] + w[n],\ n = 0, 1, ..., N-1$$

where $w[n] \sim N(0, \sigma^2)$, $s[n] \sim N(\mu_s, \sigma_s^2)$, and $s[n]$, $w[n]$ are assumed to be independent and identically distributed and

also independent from each other. Hence, the signal can be discriminated from the noise, based on its mean and variance differences.

We define the $N \times 1$ random vector $\mathbf{x}$, consisting of the random variables $(x[0], x[1], ..., x[N-1])$. 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}, decide H_1 if the likelihood ratio $L(\mathbf{x})$ is larger than a threshold γ:

$$L(\mathbf{x}) = \frac{p(\mathbf{x};H_1)}{p(\mathbf{x};H_0)} > \gamma \quad (1)$$

where $p(\mathbf{x};H_0)$, $p(\mathbf{x};H_1)$ are the multivariate probability density functions for the respective hypotheses.

From our modelling assumptions, $\mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ under H_0 and $\mathbf{x} \sim N(\mu_s\mathbf{1}, (\sigma_s^2 + \sigma^2)\mathbf{I})$ under H_1, where $\mathbf{0}$ and $\mathbf{1}$ denote the all-zero and all-one vectors respectively. Thus, by substituting the density functions in (1), manipulating the likelihood ratio and incorporating the non-data terms in the threshold, we have the test statistic $T(\mathbf{x})$:

$$T(\mathbf{x}) = N\mu_s \cdot \frac{1}{N}\sum_{n=0}^{N-1} x[n] + \frac{\sigma_s^2}{2\sigma^2} \cdot \sum_{n=0}^{N-1} x[n]^2 > \gamma \quad (2)$$

which is a function of an *averager* $T_1(\mathbf{x}) = (1/N)\sum_{n=0}^{N-1} x[n]$, which attempts to discriminate between the two hypotheses on the basis of the sample mean, and an *energy detector* $T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$, which attempts to discriminate on the basis of the variance. With γ we denote the detection threshold.

Consequently, by applying these two detectors, we can detect visual speech by exploiting the attributes that a speaking person demonstrates. In order to determine the presence of visual speech 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. The non-data terms of (2) are incorporated in the thresholds.

The averager 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}. The probability of false alarm of the averager is given by

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

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

$$\gamma_1 = \sqrt{\frac{\sigma^2}{N}}Q^{-1}(P_{FA}) \quad (3)$$

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

The energy detector 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 energy of the data 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 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 $Q_{\chi^2_N}(x)$. Therefore, the probability of false alarm is

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

Thus, the threshold is given by

$$\gamma_2 = \sigma^2 Q^{-1}_{\chi^2_N}(P_{FA}) \quad (4)$$

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

4.1 Noise Estimation

In the preceding analysis we have assumed zero mean Gaussian noise and we have concluded that the noise standard deviation is a prerequisite for the computation of our threshold. 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 philosophy of the estimation algorithm focuses on distinguishing efficiently the *signal and noise* samples from the *noise only* samples, and then calculating the noise's μ and 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 [1], suffers from the possibility that the estimation will be biased if a signal is present in the initial estimation.

The algorithm first computes initial estimates of μ and σ, in order to apply the detectors. The initial estimates are computed from the smaller 10% of the data set values, assuming that these values belong to the noise samples. Thereafter, we apply the detectors to our data set, employing the noise characteristics we have computed. The detectors distinguish the noise only samples from the signal and noise samples and new noise characteristics emerge. This process is repeated until the difference between two consecutive estimations of σ is smaller than 10^{-2}.

The stages of the noise estimation algorithm for a video sequence are displayed in Figure 3. It is obvious that the initial values of the noise statistics result in a modest estimation of the noise, as depicted in Figure 3b, and only a portion of the noise samples is identified. These noise samples, however, are used to obtain a better estimation of the noise characteristics. After two more iterations of the algorithm, shown in Figures 3c and 3d, where every time more noise samples are identified and better estimations of the noise characteristics are obtained, the noise only samples are efficiently distinguished. Hence, in the final step, an accurate estimation of the noise statistics is available.

It should be noted here that the visual speech detection procedure outlined in this section involves certain assump-

Frames	Speaking	Silent	P_D	P_{FA}
10281	3849	6432	97.14%	3.56%

Table 1: Experimental results

tions as well as small deviations from the statistical detection theory formulae. However, the experimental results presented in the next section verify that the proposed methodology is valid and efficient.

5. EXPERIMENTAL RESULTS

In order to evaluate the performance of our system, we have tested it in 28 short video sequences consisting of a total of 10281 frames, displaying individuals that exhibit both speaking and silent intervals. In particular, our test data consist of 3849 speaking and 6432 silent frames, from 7 individuals. The video sequences are recorded from news programs and talk shows, hence they correspond to real-life conditions.

As we have already mentioned, the performance of a visual speech system is influenced by a variety of factors, such as the viewing angle, the lighting and the distance of the speaker. In our experiments, the displayed faces are predominantly frontal with dimensions ranging from 100×145 to 195×315 pixels. The video sequences selection for our test data was performed so as to ensure visibility of the lips and the deformation of the mouth cavity at the speaking intervals. The frames of the video sequences have been manually marked as speaking or silent, in order to determine the ground truth. The ground truth was acquired from the visual perspective; no audio cues were used. The beginning of a speaking interval is considered as the first frame where the lips are slightly detached from one another.

In our experiments, we have (theoretically) constrained P_{FA} to 1% and we have applied the detectors to a data window consisting of 5 frames. The window was moving frame-by-frame, spanning the whole data sequence. The decision obtained for each window position characterized its central frame, i.e. when it was determined that signal was present in a certain window, the central frame of the window was marked as "speaking".

The probabilities of detection and false alarm were used as performance indicators for the visual speech detection problem. Probability of detection (P_D) can be defined as the ratio of the correctly detected speaking frames to the total number of speaking frames, whereas probability of false alarm (P_{FA}) can be defined as the ratio of the silent frames mis-detected as speaking, to the total number of silent frames. The experimental results are displayed in Table 1. Most of the false alarms are produced by the opening of the speakers mouth, either to breathe or to establish his intent to speak. Furthermore, the visual speech detection system is prone to suffer from detection errors of the face and mouth region detectors.

In Figure 4, we present the detection algorithm outcomes for four video sequences.

6. CONCLUSION

We have presented a novel method for determining lip activity and detecting visual speech. Our method uses the intensity information of the mouth region. We argue that the increase of the number of pixels with low intensity values that is produced by the opening of a speaker's mouth can be used as a cue for the visual detection of speech.

We have implemented a system that employs face and mouth region detectors and applies an averager and an energy detector to efficiently distinguish the speaking from the silent frames of a video sequence. The proposed system has been tested in a number of video sequences consisting of both speaking and silent intervals with encouraging performance. In the future, we plan to test our system to an even larger data set, and explore the influence of the viewing conditions into our method's performance. Additionally, we intend to incorporate into our system further visual cues, such as edges, as well as to examine the integration of audio cues.

REFERENCES

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

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

[3] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE CVPR*, Kauai, HI, USA, December 9-14. 2001, vol. 1, pp. 511-518.

[4] R. Lienhart, and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *IEEE ICIP*, Rochester, NY, USA, September 22-25. 2002, vol. 1, pp. 900-903.

[5] S. Asteriadis, N. Nikolaidis, A. Hajdu, and I. Pitas, "A novel eye-detection algorithm utilizing edge-related geometrical information," to appear in *Proc. EUSIPCO 2006*, Florence, Italy, September 4-8. 2006.

[6] T. Chen, and R.R. Rao, "Audio-visual integration in multimodal communication,"*Proc. IEEE*, vol. 86, no. 5, pp. 837-852, May 1998.

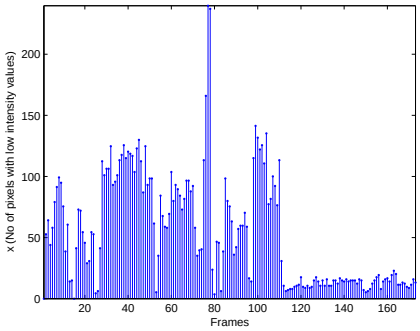
[7] J. R. Movellan, "Visual speech recognition with stochastic networks," in *Proc. NIPS 1994*, Denver, Colorado, USA, Nov. 28 - Dec. 3. 1994, pp. 851-858.

[8] 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.

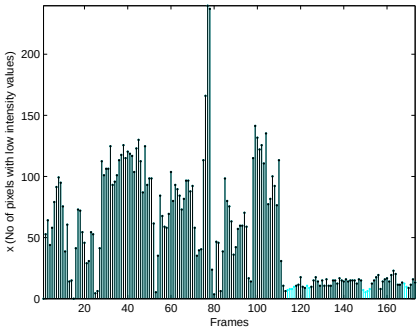
[9] J. Luettin, N.A. Thacker, and S.W. Beet, "Speechreading using shape and intensity information," in *Proc. ICSLP 1996*, Philadelphia, PA, USA, October 3-6. 1996, vol. 1, pp. 58-61.

[10] 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.

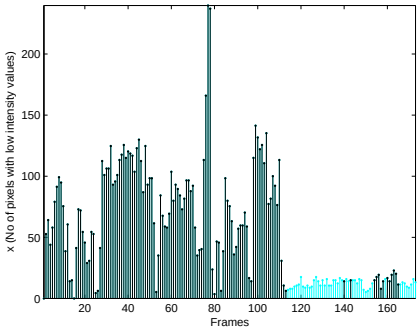
[11] M. Siracusa, L.P. Morency, K. Wilson, J.W. Fisher, and T. Darrell, "A multi-modal approach for determining speaker location and focus," in *Proc. ICMI 2003*, Vancouver, Canada, November 5-7, 2003, pp. 77-80.



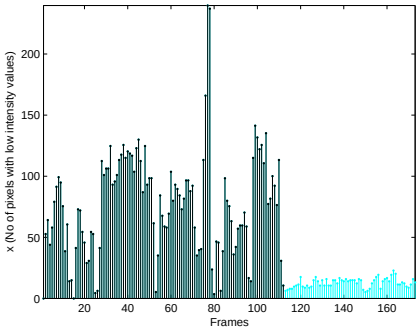
(a) Data set



(b) Noise estimation: 1st iteration

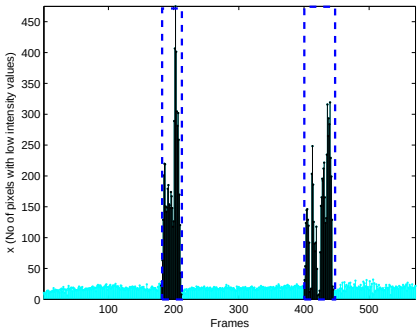


(c) Noise estimation: 2nd iteration

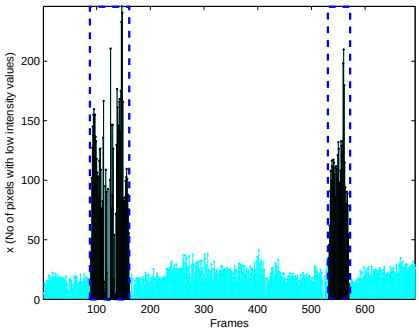


(d) Final noise estimation-Signal detection

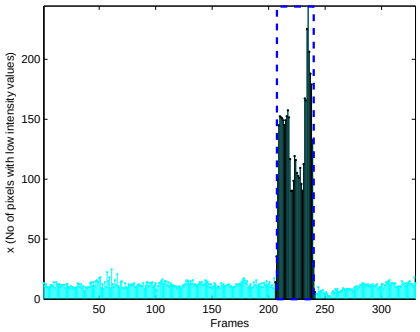
Figure 3: Noise estimation steps. Dark values: signal and noise present, bright values: only noise present



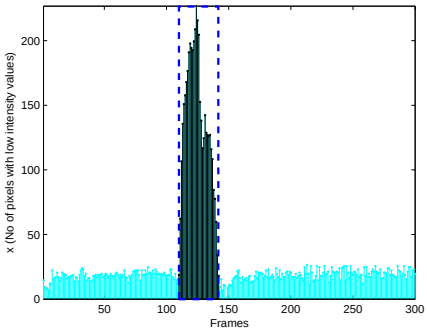
(a) Video sequence 1



(b) Video sequence 2



(c) Video sequence 3



(d) Video sequence 4

Figure 4: Visual speech detection. Dark values: Speaking frames, bright values: silent frames. The dashed rectangle encloses the speaking frames as defined in the ground truth