Adaptive Weighting Parameter in Audio-Visual Voice Activity Detection

Matar Buchbinder, Yaakov Buchris and Israel Cohen
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200003, Israel
matarb22@gmail.com, sbucris@gmail.com, icohen@ee.technion.ac.il

Abstract—**Audio-visual voice activity detectors are traditionally based on fixed algorithms and do not consider the quality of the signals in each modality. This could significantly decrease the detector's performance in cases when one of the signals is relatively of poor quality. We proposed an improved solution, which evaluates the signal's quality in each modality and weights them accordingly. In this paper, we present a method for estimating the video quality, particularly in the presence of noisy motion vectors or global motion of the camera. The fussy motion vectors are intended to simulate blurred, unfocused video or low resolution sensor. An adaptive setting of the weighting parameter between the audio and the video signals ensures an optimal bi-modal detector. The proposed method was incorporated with an audio-visual voice activity detector, and was tested with a real data set. Simulation results have shown an improved performance compared to the existing fixed method.**

I. Introduction

Voice activity detection algorithms often rely on the audio signal. Usually those algorithms assume that the noise is slowly time-varying with respect to the speech, and use that assumption to separate them efficiently. However, this assumption does not hold in the presence of transient noise, like keyboard typing or door knocking. In that case, a significant decrease in the audio detector's performance is expected. Mousazadeh and Cohen, [1], recently introduced a voice activity detector (VAD) for non-stationary noises. The proposed detector employs spectral clustering, and tries to deal with the transient's problem by averaging noise statistics over short windows in the time domain. The results of this detector in the presence of transients were limited.

For that reason, VADs which use the video signal are advantageous. The video signal is immune to all kinds of noises, including transients. The first step in the detection process based on video is to focus on the relevant block in each frame that contains the lips. The extraction of the lips can be based on several algorithms. One approach for lips detection is to use the features of the lips. In [2] and [3] the extraction of the lips features is based on the contours of the lips. Another approach is to exploit the unique shape and color of the lips ([4], [5]). One of the main approaches for visual voice activity detection is to estimate the movement of the lips during speech intervals. This approach has shown mainly good performance, although it was found sensitive to movement of the lips in non-speech frames. In [6] and [7] motion estimation approach was utilized, which exploits the motion vectors for voice activity detection. The energy in the mouth region is determined using optical flow, and the classification is based on a Hidden Markov Model (HMM).

While VADs which are based on visual detection may overcome the problem of transient noise, their performances are usually inferior to the audio-based detectors in a quiet enviroment or in the presence of stationary noise. This leads to a growing interest in the combination of both detectors, thus, exploiting the strengths of each modality [8].

Recently, Dov, Talmon and Cohen [9] introduced a new audio visual voice activity detector. This detector is based on a supervised learning procedure, and a labeled training data set is taking into consideration. Diffusion map is applied separately and similarly to both the audio and video signal's features in order to build a low dimensional representation. The visual features are based on the motion vectors of the video. The calculation is based on the Lucas-Kanade method, as described in [10] and [11]. The measures of the two modalities are equally merged into one bi-modal detector. The experimental results in [9], show that the suggested detector outperforms other state-of-the-art AV-VADs.

However, the combined bi-modal detector presented in [9] does not consider the signals' absolute quality or their relative quality (for example when one signal is significantly better than the other), and makes a blind decision, using a fixed formula that weights the two signals equally.

in this paper we take into consideration the quality of the video signal, to wisely determine the weighing of the two input signals. The purpose of this work is to efficiently estimate the quality of the video signal, and adaptively build the weighting function between the audio and the video modalities. The video quality is based on characteristics like image sharpness, texture, global motion and rotation of the camera. Experimental results demonstrate the improved performance of the adaptive weighting over the fixed weighting modal. The algorithm calculates the variance of the motion vector in a frame-by-frame manner and sets a quality factor for the entire video.

The paper is organized as follows: in Section II, the problem

This Research was supported by Qualcomm Research Fund and MAFAAT Israel Ministry of Defense.

978-1-5090-2152-9/16/$31.00 ©2016 IEEE

is described and formulated. The description of a video modulation, in particular noise and global motion modulation is presented in Section III. Experimental results demonstrating the improved performance of the proposed algorithm are presented in Section IV.

II. PROBLEM FORMULATION

Let $P(a_i)$ and $P(v_i)$ denote measures of voice activity from the audio and the video signals, respectively, which were calculated in [9]. The bi-modal measure of voice activity is given by

$$P^B(a_i, v_i) = (1 - \alpha)P(a_i) + \alpha P(v_i), \tag{1}$$

where α is a weighting parameter in the range of $[0, 1]$. For example, in case when the audio signal is relatively clean and the video signal has poor quality, α should be close to 0.

In this work, we focus on the quantification of the video signal quality, while assuming the audio signal is in a decent condition, meaning that α is in the range of $[0, 0.5]$, depending solely on the video quality. The goal in this paper is to adjust α over time. Specifically, α should be set adaptively according to the video quality parameter, i.e.,

$$\alpha = f(Q(v_i)) \tag{2}$$

where $Q(v_i)$ is a general quality factor of the video signal v_i, and $f(Q(v_i))$ is a function of $Q(v_i)$. The quality factor addresses challenging real scenarios in the video signals, from blurred and unfocused video, to global motion of the camera. Detailed definition of the quality factor $Q(v_i)$ is described later in the paper.

In future work the quality of the audio signal should also be taken under consideration. In that case, the quality factor of each signal should be normalized to the range of $[0, 1]$, and α would be expressed as

$$\alpha = \frac{Q(v_i)}{Q(a_i) + Q(v_i)}, \tag{3}$$

where $Q(a_i)$ and $Q(v_i)$ are the general quality factors of the audio and video signals, respectively.

III. VIDEO MODULATION

The proposed algorithm evaluates the quality of the video signal. In order to do that, a set of objective parameters of the video signal has to be defined. Let $\{v_i\}_{i=0}^{N}$ be the video data set comprising of N consecutive video frames $v_i \in \mathbb{R}^{WxH}$ where W and H are the number of pixels in the raw and column of each frame, respectively. In each frame a cropping the bounding box of the mouth is performed as a preprocessing stage. The cropping method extends the scope of this work. Two scenarios that are taken under consideration in this paper are noisy frames and global motion of the camera.

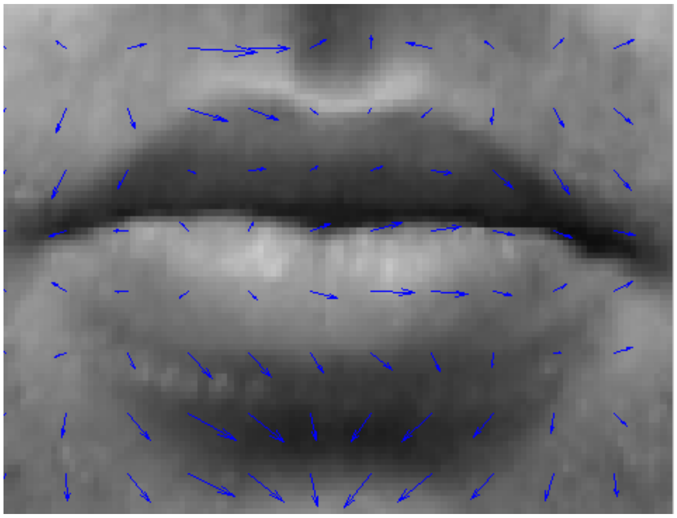

Fig. 1: Motion vectors presented on their matching frame.

A. Noisy Frame Modulation

Noisy image may results from a variety of reasons, like poor quality of the sensors, blurred, unfocused frame, or low textured and uniformed areas within the frame. Each frame in the video was devided into 100 macroblocks, where a macroblock consist from a group of pixels and is used as a processing unit in the image. We use the motion vectors of each frame to determine its quality, hence the motion vectors associated to each motion-informative macroblock depending upon the quality of the frame. Motion vectors are used to represent a macroblock in a frame, based on the position of this macroblock (or a similar one) in the previous frame. The optical flow method tries to calculate the motion between two consecutive image frames, i.e. it calculates the difference between these two frames. Under this assumption it's easy to see that the motion vectors are extremely informative to determine the frame quality. While the motion vectors of a clean frame are gradually varying, those of a noisy frame are random, and have almost no correlation to one another. Figure 1 presents a clean frame from the detection process, with its corresponding motion vectors.

Let γ_i be characteristic angle of the macroblock m_i, which states for the angle of its motion vector in degrees. The characteristic differential angle of m_i is given by

$$\Delta_i = \frac{1}{4}\sum_{j=1}^{4} \gamma_i - \gamma_j, \tag{4}$$

where m_i, m_j are consecutive macroblocks so that $\{m_1, m_2, m_3, m_4\}$ are the up, down, right and left neighbor of the macroblock m_i. The parameters $\{\Delta_i\}_{i=1}^{M}$ are random variables, where M is the number of macroblocks in the frame. Δ_i represents the similarity between a certain macroblock to its surrounding, so as homogeneous the region is, Δ_i get closer to 0. The variance of Δ_i is given by

$$V_f = E[\Delta_i^2] - E[\Delta_i]^2, \tag{5}$$

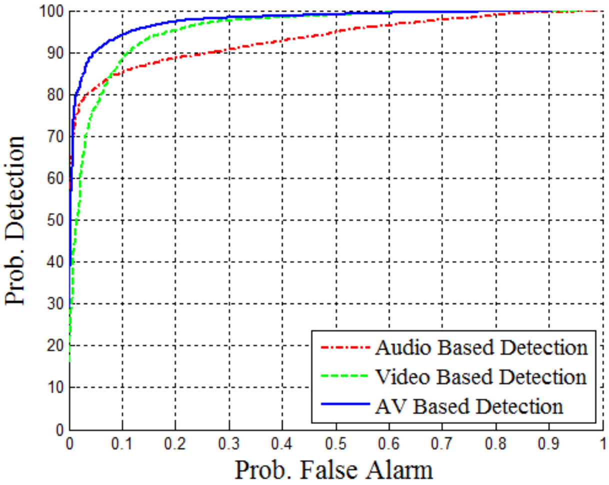

Fig. 2: ROC curves of the original detectors presented at [9]. α set to be a constant at value 0.5

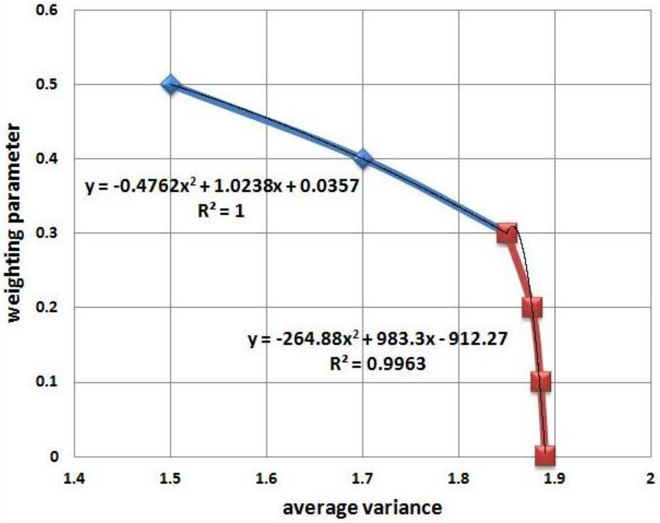

Fig. 3: the empirical dependence between the noise parameter V and the weighting parameter α.

where V_f is defined as the characteristic parameter of the entire frame. Intuitively, for highly noisy frames, V_f should get high values, and vice versa. This method of V_f calculation is applied in a frame-by-frame manner to determine the noise characteristic parameter for the entire video

$$V = \frac{1}{N}\sum_{f=1}^{N} V_f, \tag{6}$$

where the given video signal comprises N frames.

B. Global Motion Modulation

When global motion is present in the frame, the motion of the lips is usually negligible, what makes it harder to detect. Therefore, a significant global motion of the camera could have an impact on the video-based detection process. Multiple works practice and explore the issue of global motion estimation. For example, Basu and Pentland [12] investigate how raw, noisy motion vectors can be used to estimate global camera motion. For our use, an estimation of the absolute value of the global motion is required. Recall that in [9], the proposed visual features for voice activity detection are based on the motion vectors of the video signal. The motion vectors field does not properly represent fast global motion in the video. However, absolute values of the motion vectors are always significantly higher in global motion's frames than in regular ones. Let m_i be the i-*th* macroblock of the frame v_f, and let MV_ix and MV_iy denote the horizontal and the vertical components of the motion vector. We define a characteristic value for each frame v_f given by

$$|MV_f|_{avg} = \frac{1}{M}\sum_{i=1}^{M} \sqrt{[MV_ix]^2 + [MV_iy]^2}, \tag{7}$$

where M is the number of macroblocks in the frame v_f. This calculation of $|MV_f|_{avg}$ is applied in a frame by frame manner to determine the global motion characteristic parameter for the entire video

$$|MV|_{avg} = \frac{1}{N}\sum_{f=1}^{N} |MV_f|_{avg}, \tag{8}$$

where the given video signal comprising N frames. In cases that $|MV|_{avg}$ is greater than a certain threshold value, global motion of the camera is assumed.

The quality factor $Q(v_i)$ is calculated independently for each of the two parameters presented above, i.e., the noise parameter (6) and the global motion parameter (8). For each parameter, a different value for $Q(v_i)$ is set, where the overall $Q(v_i)$ is the lowest one. The explicit calculation for each parameter is described in the next section.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

The experimental setup is based entirely on the simulation described in [9]. The data set is obtained from 7 speakers loudly reading an article. The training data set contains 30 sec of 6 speakers, while the test data is using 60 sec of each of the 7 speakers. The video is recorded using a frontal camera of the smartphone (25 [fps], 640x480 resolution). A bounding box of the mouth (110x90 pixels) is cropped out of the video. The number of macroblocks of the motion vector calculation is chosen to be M=100, which means that each macroblock consist of 11x9 pixels. Figure 1 shows an example of high quality, non-noisy frame, and its corresponding motion vectors.

B. Receiver Operating Characteristic

In order to evaluate the detector's performance of the proposed algorithm, an objective, reliable measure is needed to be defined. First, it should be mention that the ground truth for

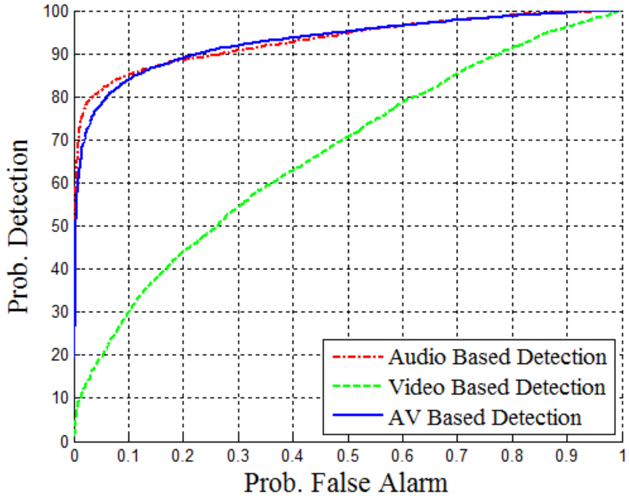

Fig. 4: The ROC curves of the detectors while adding white gaussian noise with variance 1 and $\alpha = 0.5$.

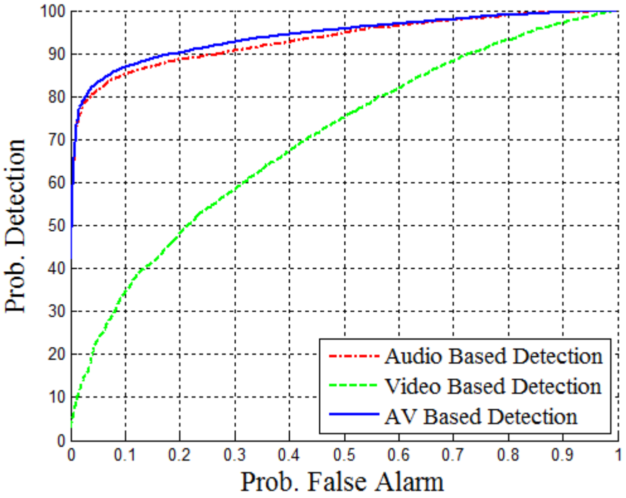

Fig. 5: The ROC curves of the detectors while adding white gaussian noise with variance 1 and α adaptively set.

TABLE I: Experimental Results of the Adaptive Weighting Parameter in Several Variances Values of the Noise

Var=0.1	Var=0.2	Var=0.5	Var=0.8	Var=1	Var=1.3	Var=1.5
0.5	0.447	0.329	0.303	0.253	0.182	0.133
0.5	0.446	0.332	0.288	0.263	0.150	0.088
0.5	0.431	0.319	0.292	0.222	0.153	0.067
0.5	0.439	0.335	0.291	0.233	0.113	0
0.5	0.445	0.332	0.289	0.265	0.038	0.040
0.5	0.441	0.330	0.289	0.255	0.195	0.023
0.5	0.446	0.337	0.292	0.251	0.127	0.039

both the audio and the video modalities is given. This ground truth is labeled '1' in the presence of speech, and '0' otherwise, for each modality. The uniform audio-visual ground truth is the "or" function between the two ground truth functions of the separate modalities.

We evaluate the detector performance using the receiver operating characteristic (ROC) curve as a quality measure. ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (detection) against the false positive rate (false alarm) at various threshold settings.

In our experiments, the proposed AV-VAD, with the combine ground truth, is compared to each of the single modality versions, upon a ROC curve. The original ROC curve of the detector [9] is presented in Figure 2. The video data has good quality and the weighting parameter α is set to be 0.5.

C. Weighting Parameter Evaluation

1) Noisy frame: For noisy frames simulation, white gaussian noise was added to the horizontal and vertical components of the motion vectors, MV_ix and MV_iy, respectively. Sereval simulations were performed, in each one the noise had different variance, and therefore different impact on the detection process. The link between V, the noise characteristic parameter presented in (6), and the quality factor $Q(v_i)$, or the weighting parameter α, was empirically found. This connection is described at the graph presented in Figure 3, which indicates different dependence for different values of V. Also, a threshold value was set, so if the variance is smaller than τ, the frame is declare as 'not noisy', and α is set to be 0.5 accordingly.

Several experiments has been made, and for each value of V, the optimal α in terms of the ROC was found and set. Those discrete results were generalized for the continuous case, as presented in Figure 3. The adaptive algorithm calculate $Q(v_i)$ and α from the range of $[0, 0.5]$ according to the formulas mentions in Figure 3. the two colors in the graph indicate different link between V and α. An example for the ROC curves of the detector containing the noisy video, with noise variance of 1, and α set to be 0.5, can be shown in Figure 4. In Figure 5, the ROC curves of the same detector is presented, when the value of α is adaptively set. Comparing the two figures, it can be seen that the proposed algorithm improve the resulted bi-modal detector. In particular, the audio version had outperformed the bi-modal version in low values of false alarm, and after adjusting α, the bi-modal detector outperform the audio version, for all possible values of false alarm rates. The values of α from the adaptive algorithm, for numerous variances of the noise, are summarized in table I. Note that α is a vector in the length of 7 since the test data contains 7 speakers.

2) Global motion: The simulation of the global motion divided into several tests: global motion in the horizontal axis, in the vertical axis, in both horizontal and vertical and rotation of the camera. Each test aimed at finding the influence of each movement of the camera on the video data quality in terms of motion vectors extraction and the voice detection performance. For the global motion simulation, a constant value was added to MV_ix and MV_iy for horizontal or vertical motion, respectively, when the value of this constant represents the movement extent.

It was empirically found that all the simulations described above have no influence whatsoever on the detector performance, apart from the global motion in the horizontal axis, which show slightly inferior results. The uninfluenced results consistent with the logic, because the definition of the motion vectors. Since the motion simulation is applied exactly the same way for each macroblock in the frame and for each frame in the video, the difference stays approximately the same. The slightly inferior results in the horizontal direction can be explained by the fact that most of the mouth movements (i.e. the signal) in speech frames are in the vertical direction, as mentioned in [13]. This work provides a three-dimensional model of human lips motion trained from video. Since most of the relevant data is in the vertical direction, the signal might overcame a small values of noise added in this direction (high SNR) so the vertical global motion is hardly affect the detection. However, the horizontal noise added to the signal is easily overcome the signal in this direction, which is low to begin with.

It should be noticed that for each and every one of the global motions simulated, including movements in every direction and extent, the quality of the video signal remains the same as if there was no motion whatsoever. namely, the speech detection based on the video modality does not influenced from global motion, hence α should set to be 0.5. Nevertheless, one extreme and yet realistic scenario should take under consideration in further research - large global motion of the camera, where the speaker head gets out of the frame, thereby making the video signal irrelevant. According to the size of the frame, the average size of the head in it and the average absolute values of the motion vectors, a threshold value τ should be set, so if $|MV_f|_{avg}$ of a certain frame is bigger then τ, α set to be 0, and the detection is based merely on the audio signal.

V. Conclusions

We have presented an adaptive algorithm for calculating the weighting parameter between the audio and video signals for audio-visual voice activity detection. The algorithm evaluates the video signal quality based on modeling blurred unclear frames, low resolution sensor and global motion of the camera. Experimental results show that the proposed adaptive weighting parameter improves the performance of a bi-modal detector compared to a constant weighting parameter. The proposed algorithm can be incorporated in any AV-VAD which employs a weighting parameter for the two modalities.

Here, extreme scenarios, like getting out of the frame of the speaker, were not addressed, and need to be investigated in future work. Furthermore, the audio signal quality estimation is also not in the scope of this work, and need to be considered in future research.

References

[1] S. Mousazadeh and I. Cohen, “Voice activity detection in presence of transient noise using spectral clustering,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1261–1271, June. 2013.

[2] D. Sodoyer, B. Rivat, L. Girin, J. Schwartz, and C. Jutten, “An analysis of visual speech information applied to voice activity detection,” *Proc. 31st IEEE Int. Conf. Acoust., Speech, Signal Process (ICASSP).*, vol. 1, 2006.

[3] B. Rivat, L. Girin, and C. Jutten, “Visual voice activity detection as a help for speech source separation from convolutive mixtures,” *Speech Commun.*, vol. 49, no. 7, pp. 667–677, 2007.

[4] D. Scott, C. Jung, J. Bins, A. Said, and A. Kalker, “Video based vad using adaptive color information,” *Proc. 11th IEEE Int. Symp. Multimedia (ISM).*, pp. 80–87, 2009.

[5] C. Lopes, J. Goncalves, A.L.and Scharcanski, and C. Jung, “Color-based lips extraction applied to voice activity detection,” *Proc. 8th IEEE Int. Conf. Image Process (ICIP).*, pp. 1057–1060, 2011.

[6] A. Aubrey and Y. C. J. Hicks, “Visual voice activity detection with optical flow,” *IET Image Process.*, vol. 4, no. 6, pp. 463–472, 2010.

[7] P. Tiawongsombat, M. Jeong, J. Yun, B. You, and S. Oh, “Robust visual speakingness detection using bi-level hmm,” *Pattern Recogn.*, vol. 45, no. 2, pp. 783–793, 2012.

[8] L. Peng, W. Zuo-ying, J. Yun, B. You, and S. Oh, “Audio-visual voice activity detection,” *Translated from Journal of Tsinghua University (Science and Technology)*, vol. 45, no. 7, pp. 896–899, 2005.

[9] D. Dov, R. Talmon, and I. Cohen, “Audio-visual voice activity detection in using diffusion maps,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, April. 2015.

[10] J. Barron, D. Fleet, and S. Beauchemin, “Performance of optical flow techniques,” *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.

[11] A. Bruhn, J. Weickert, and C. Schnörr, “Lucas/kanade meets horn/schunck: Combining local and global optic flow methods,” *Int.J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, 2005.

[12] M. Pilu, “On using raw mpeg motion vectors to determine global camera motion,” *Digital Media Department, HP Laboratories Bristol.*, pp. 97–102, August. 1997.

[13] S. Basu and A. Pentland, “A three-dimensional model of human lip motions trained from video,” *M.I.T Media Laboratory Perceptual Computing Section Technical Rep ort .*, no. 441, June. 1997.