Visual Voice Activity Detection at different Speeds

Bart Joosten[1], *Eric Postma*[1], *Emiel Krahmer*[1]

[1]Tilburg center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands

bart.joosten@gmail.com, eric.postma@gmail.com, e.j.krahmer@uvt.nl

Abstract

Visual Voice Activity Detection (VVAD) refers to the detection of speech from a video sequence by means of visual cues. VVAD provides a useful addition to auditory voice activity detection, in particular in cases involving multiple speakers or background noise. This paper focusses explicitly on the measurement of facial movements at different speeds to determine which rates of movement contribute to VVAD. Facial movements in video sequences of talking faces are measured using a spatiotemporal Gabor transform. VVAD performances based on these measurements are determined for different speeds and compared to simple frame-differencing. In addition, performances are assessed for the entire frame, the head region, and the mouth region. The results obtained reveal an elevated VVAD performance for large speeds as compared to low speeds. In addition, frame differencing performs at a level comparable to that of the spatiotemporal Gabor method at the optimal speeds.

Index Terms:visual active speech, frame differencing, Gabor transform, spatiotemporal Gabor transform

1. Introduction

Human speech comprises two modalities: the auditory modality and the visual modality. Although auditory cues are dominant, visual cues such as, lip, jaw, head, and eyebrow movements, provide useful additional information to support speech detection.

Automatic Voice Activity Detection (VAD) benefits greatly from visual cues, especially in case of multiple speakers or background noise. It has been found that mouth movements typically precede the vocal signal [1]. Exploiting the visual cue of mouth movement can help VAD to accurately determine the onset of speech. Visual-cue enhanced VAD could, for instance, facilitate the tagging of speakers in automatic conference recording applications.

Visual Voice Activity Detection (VVAD) refers to detection methods that focus solely on visual cues for speech detection. Previously proposed VVAD methods mostly relied on lip tracking [2, 3, 4]. Aubrey *et al.* present and compare two VVAD methods [2], viz., a method based on Active Appearance Model (AAM) parameters pertaining to the lip region, and a method that applies a retinal filter to the lip region. Sodoyer *et al.* [3] propose a VVAD system that focusses on silences. They use a dynamical parameter derived from the relationship between the inner lip width and height to distinguish between silence and non-silence. The width and height of the tracked lips are also used in the work of Liu *et al.* [4] to construct a VVAD. Combined with an appearance feature extracted a the center of the mouth they generate a series of static and dynamic features per frame and use AdaBoost to determine the most informative ones for classification of voice activity.

All these VVADs recognize the importance of motion, since they all incorporate dynamic features. However, in their approaches movements are implicitly captured in the extracted features at the location of the lips. Our objective is to focus explicitly on all facial movements associated with human speech. More specifically, we address the question: How much do different speeds of facial movement contribute to VVAD performance? In addition, since movement during speech in not limited to the mouth region, we will answer this question for three different scales of analysis, i.e., the entire video frame, the head region, and the mouth region.

We assess the VVAD performances as a function of different movement rates by using Spatiotemporal Gabor Filters (SGF) [5] which can be constructed to respond maximally to moving contours at a specific speed and direction of movement.

The performances of our SGF-based VVAD will be compared to straightforward Frame Differencing (FD) [6]. FD is pixel-based and measures movement by means of pixel-wise intensity changes, whereas SGF measures movement by means of biologically-informed filters. Figure 1 shows two original frames, a non-speech and a speech frame and their corresponding SGF and FDM applied outputs.

The outline of the remainder of the paper is as follows. In section 2 we describe our methods for VVAD based on Spatiotemporal Gabor filters and Frame Differencing. Section 3 describes our experimental setup, where we address our dataset and the settings we used to compute our features and to do the classification. We present our results in section 4 and we conclude our paper with a discussion and conclusion of the results in sections 5 and 6 respectively.

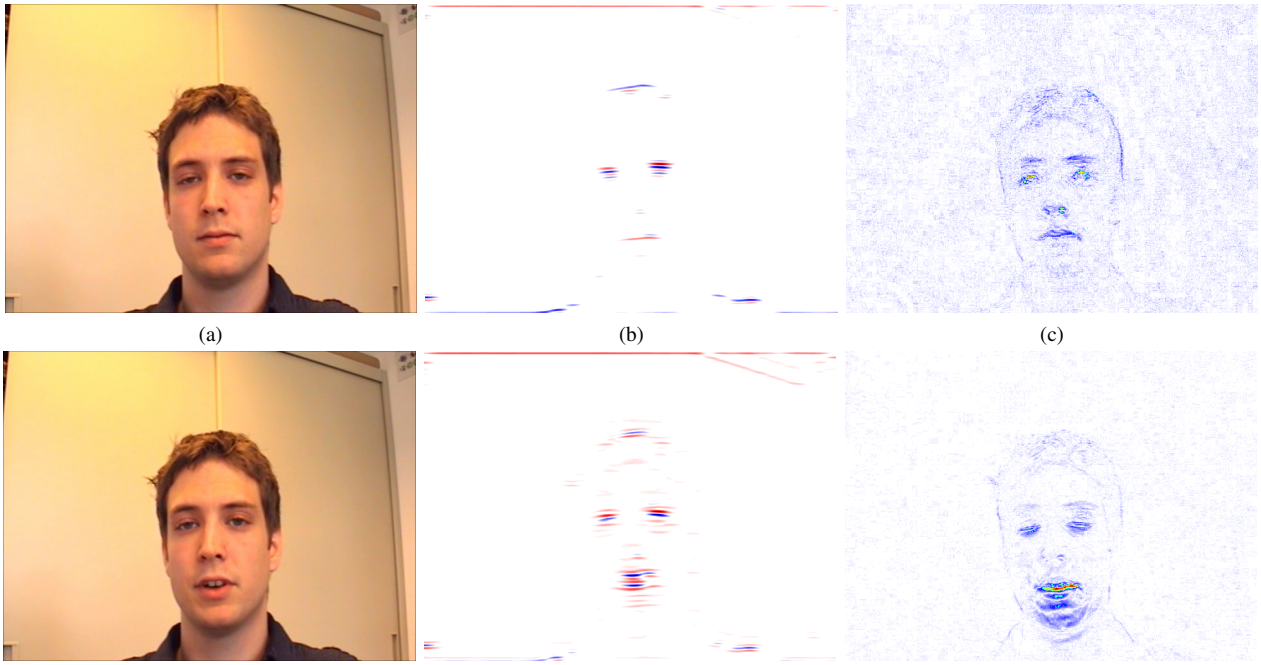

Figure 1: Comparison between a no-speech (a-c) and a speech (d-f) frame. The columns represent, from left to right: (a) and (d) original frame, (b) and (e) FDM output, and (c) and (f) SGF output where the prefered speed and direction are four pixels upwards, respectively. Activation ranges from dark blue (low) to red (high). Colormaps have been scaled to map uniform regions with low activation to white pixels.

2. Visual Voice Activity Detection Method

Spatiotemporal Gabor Filters (SGFs) were developed by Petkov and Subramanian [5] and are biologically informed in that they are based on the functional properties of cells in the primary visual cortex. These cells have a sharp tuning to motion with a certain speed and direction. SGFs extend traditional 2D Gabor filters [7] with the temporal dimension. 2D Gabor filters respond to oriented spatial frequencies. SGFs respond to moving oriented spatial frequencies. The outputs (energies) of the SGFs are aggregated to yield features for our VVAD method. We aggregate all filter responses corresponding to a single speed and orientation by summing them. As a reference to the performance of the SGFs VVAD method we compare them to aggregated output of a straightforward Frame Differencing Method [**?**].

To obtain a VVAD method, the SGFs and FDM are combined with a Support Vector Machine (SVM) that takes the aggregate SGF and FDM features as input. Classification is done on a per frame basis.

3. Experimental method

Our dataset consists of video sequences of participants in a surprise-elicitation experiment. In the experiment, participants were instructed to read aloud a word displayed on a computer screen [8]. Participants pronouncing the Dutch word for liver[1] In the dataset, the "liver" fragments plus any additional utterances are positive instances (i.e., speech) and the rest of the fragments serve as negative instances (no speech). The speech fragments were labeled using a VAD based solely on the audio signal. Each fragment has a length of approximately four seconds (i.e. 120 frames).

For the SGF VVAD method, individual frames are convolved with filters that operate on 23 speeds and on 8 orientations, resulting in 8 feature values per frame for each evaluated speed. A similar approach is taken for the FDM based VVAD method, however, besides summing the pixel values we also store the mean and the standard deviation of the differenced frame's pixel values, resulting in a 3-dimensional feature vector per frame. All features were normalized to z-scores.

To assess the performance of the VVADs at the head and mouth regions, we rely on an automatic face detector to label the frames with the head location. For the mouth region we use the lower half of the rectangle including the face. Since the face detector was unable to sufficiently determine the head locations in five fragments, we excluded those for evaluation of the head and mouth region

[1]For the purpose of the original experiment, the word was pronounced in two conditions: a neutral condition and a surprise-eliciting condition. In the present study we ignore both conditions and treat them equally as speech.

VVADs. Therefore the ratio of speech and non-speech fragments for the entire frame experiment is 633 to 3260 respectively, and 582 to 2987 respectively, for the head and mouth region experiments.

In the spatiotemporal Gabor transform, we tested filters sensitive to 23 different speeds where $v = \{0.1, 0.5, 0, 75, \ldots, 4, 5, \ldots, 10\}$ and 8 directions where $d = 0$ to 315 degrees in steps of 45 degrees.

The VVAD performances of the SGFs at each speed as well as the FDM based features were evaluated using a radial basis Support Vector Machine. Finding the SVM's optimal kernel parameters of C and γ which yield the highest accuracy, was done using a grid search. At each stage in the parameter grid search the resulting SVM was evaluated using a leave-one-subject-out cross validation scheme, i.e., an SVM was trained on all but on subjects and then tested using the left out subject. The methods were evaluated using the F-score, precision and recall. Performance scores were averaged over each fold.

4. Results

Figure 2 shows the performance scores of the SGFs based VVAD at the 23 evaluated speeds and at three areas. The plots show that the performance increases at higher speeds for all three areas. The mouth area yields the overall best performances when compared to the other areas, except at speed 0.75 pixels per frame. There we see an increase in performance for the performance of the total frame and head region, whereas the performance for the mouth region slightly drops. The performance measures We listed the speed that yielded the highest F-score at each area in table 1 and compared those to the F-scores of the FDM based VVAD. Frame Differencing performs slightly better than the Spatiotemporal Gabor filters at their optimal speed, except for the head area.

Table 1: *Optimal speed and associated f-score of SGFs for VVAD obtained with SVM training compared to SVM f-score of FD. Speed is in pixels per frame.*

	SGF		**FD**
	Speed	F-score	F-score
Total frame	4	60.6	64.3
Head	8	66.8	66.2
Mouth	3.25	67.1	71.5

5. Discussion

The FDM based VVAD performance generally outperforms the single speed SGFs based VVAD. SGF tend to smooth the image, which makes them more sensitive to contour changes over time, however this also makes them more sensitive to noise. In this study paper we tried to determine the optimal speed at which visual voice activity could be detected. Speech, however, typically occurs over multiple consecutive frames. The performance of our SGFs based VVAD would presumably increase if it would take this temporal information into account, e.g., by using a sliding window approach to classify individual frames.

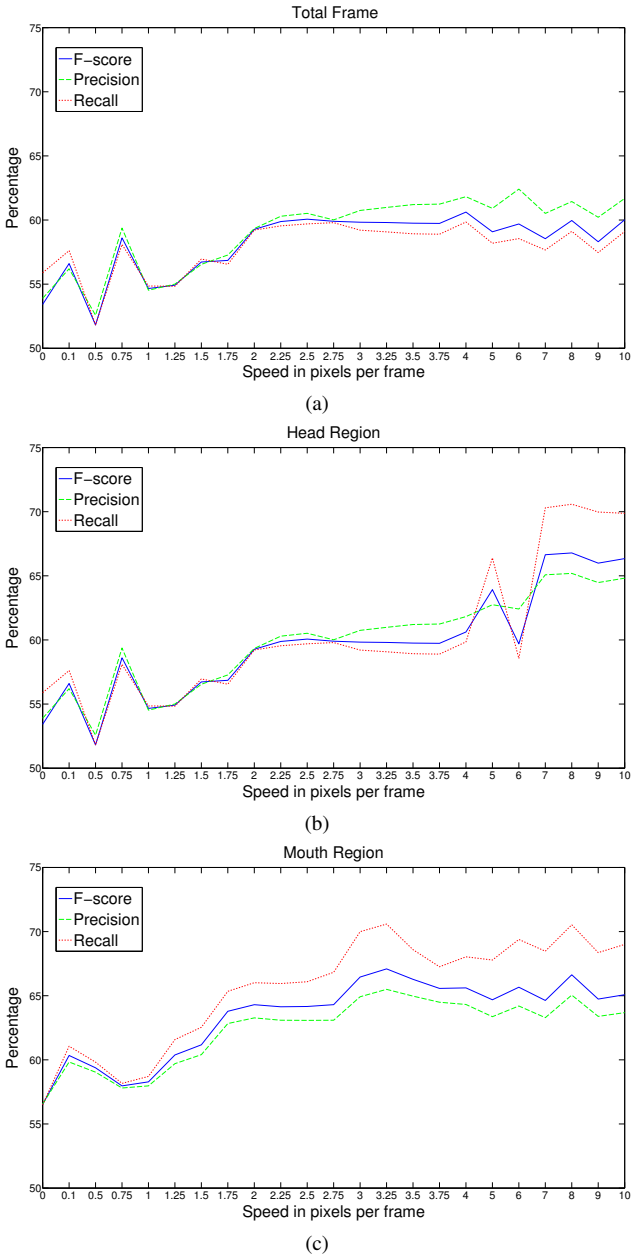

Figure 2: Performance scores of the SGFs based VVAD evaluated at 23 different speeds (x-axis) for (a) the total frame, (b) the head region, and (c) the mouth region.

6. Conclusion

Speeds of over two pixels per frame appear to be relevant for VVAD. As expected, the VVAD methods applied to

the mouth region yields the highest performance results. However, applied to the total frame and head region the performances do not differ greatly. Future research on our SGFs based VVAD must determine if combining features at different speeds and areas results in a higher performance over the FDM based VVAD. To examine this, we will also apply our methods to a dataset that is better suited for VVAD performance comparison.

7. References

[1] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The Natural Statistics of Audiovisual Speech," *PLoS Computational Biology*, vol. 5, no. 7, p. e1000436, Jul. 2009.

[2] A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal filltering," 2007.

[3] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *The Journal of the Acoustical Society of America*, vol. 125, p. 1184, 2009.

[4] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," *Sensor Signal Processing for Defence (SSPD 2011*, pp. 1–5, 2011.

[5] N. Petkov and E. Subramanian, "Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biological Cybernetics*, vol. 97, no. 5-6, pp. 423–439, Oct. 2007.

[6] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*, 3rd ed. Prentice Hall, Aug. 2007.

[7] A. B. Ashraf, S. Lucey, and T. Chen, "Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 7, pp. 1335–1341, 2010.

[8] B. Joosten, E. Postma, E. Krahmer, M. Swerts, and J. Kim, "Automated Measurement of Spontaneous Surprise," in *Proceedings of Measuring Behavior 2012*, A. J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L. P. J. J. Noldus, and P. H. Zimmerman, Eds., Utrecht, The Netherlands, Aug. 2012, pp. 385–389.