Cascading Appearance-Based Features for Visual Voice Activity Detection

Rajitha Navarathna[1], *David Dean*[1], *Patrick Lucey*[1,2], *Sridha Sridharan*[1], *Clinton Fookes*[1]

[1]Speech, Audio, Image and Video Technology Lab, Queensland University of Technology, Australia
[2]Robotics Institute, Carnegie Mellon University,
Department of Psychology, University of Pittsburgh, USA
{r.navarathna, d.dean, p.lucey, s.sridharan, c.fookes}@qut.edu.au

Abstract

The detection of voice activity is a challenging problem, especially when the level of acoustic noise is high. Most current approaches only utilise the audio signal, making them susceptible to acoustic noise. An obvious approach to overcome this is to use the visual modality. The current state-of-the-art visual feature extraction technique is one that uses a cascade of visual features (i.e. 2D-DCT, feature mean normalisation, inter-step LDA). In this paper, we investigate the effectiveness of this technique for the task of visual voice activity detection (VAD), and analyse each stage of the cascade and quantify the relative improvement in performance gained by each successive stage. The experiments were conducted on the CUAVE database and our results highlight that the dynamics of the visual modality can be used to good effect to improve visual voice activity detection performance.

Index Terms: visual speech, voice activity detection, CUAVE database, static features, dynamic features

1. Introduction

An interesting problem in the speech processing field is the determination of speech and non-speech segments, as it is useful in many applications (e.g. real-time speech transmission on the Internet [1] and mobile communication services [2]). A significant amount of research has been conducted in the field of VAD in the presence of acoustic noise over the past decade [3, 4, 5, 6], however, the robustness and effectiveness depends on the acoustic environment and very poor when the level of background noise increases.

Using the visual modality is a potential method of improving the robustness of VAD in noisy environments. Recently, there have been a few attempts to incorporate the visual movements in VAD. In 2004, Liu et. al [7] used template matching to extract the region-of-interest (ROI) and applied PCA on the ROI to extract features, which were then used for VAD. In 2006, Sodoyer et. al [8] used lip width and height as features to perform VAD. More recently, Libal et al. [9] developed a real-time system to recognise visual speech activity on low cost embedded platforms. This system uses a camera mounted on the rearview mirror to monitor the driver. It detect face boundaries and facial features, and finally use lip motion clues to recognize VAD.

As described above, these above methods only use primitive visual features[1] and do not utilise both the static and dynamic speech information encoded within the region around a speaker's mouth. A technique which incorporates this information is based on a cascade of appearance based features, first devised by Potamianos et. al [10]. This technique has been established as the state-of-the-art for visual feature extraction for audio-visual automatic speech recognition (AVASR) [11, 12, 13]. In literature, there has been no detailed study reporting the effect that the static and dynamics portions of visual speech has on VAD.

Figure 1: *Examples of the CUAVE individual sequences*

In this paper, we analyse the effect that different types of features (both static and dynamic) have on the performance of visual VAD. The experiments are carried out on the CUAVE database and each stage of the cascading appearance based features are analysed. The rest of the paper is organized as follows: Section 2 describes the experimental data (CUAVE database). Section 3 describes the visual front-end, while Section 4 describes the cascading appearance based features. Section 5 outlines the visual VAD system. Experiments are described in Section 6, which is followed up with some concluding remarks.

2. CUAVE database

The CUAVE database [14] is a publicly available audio-visual database which contains speakers talking in frontal and non-frontal poses. The main motivation behind the creation of the CUAVE database was to create a flexible, realistic and easily distributable database that allows for representative and fairly comprehensive testing.

The CUAVE database consists of two sections, the first containing individual speakers, and the second containing groups of two (or more) speakers. Both sections were designed to represent realistic capture conditions for visual speech, including speaker movement and pose variation. The CUAVE database consists of 36 speakers (19 male and 17 female speakers). The database is a speaker independent corpus of over 7000 utterances and all the recorded speech is in English. Some examples of the visual frames from the individual section of the CUAVE

[1]It must be noted that in the case of Libal et al. [9], that this was due to real-time constraints

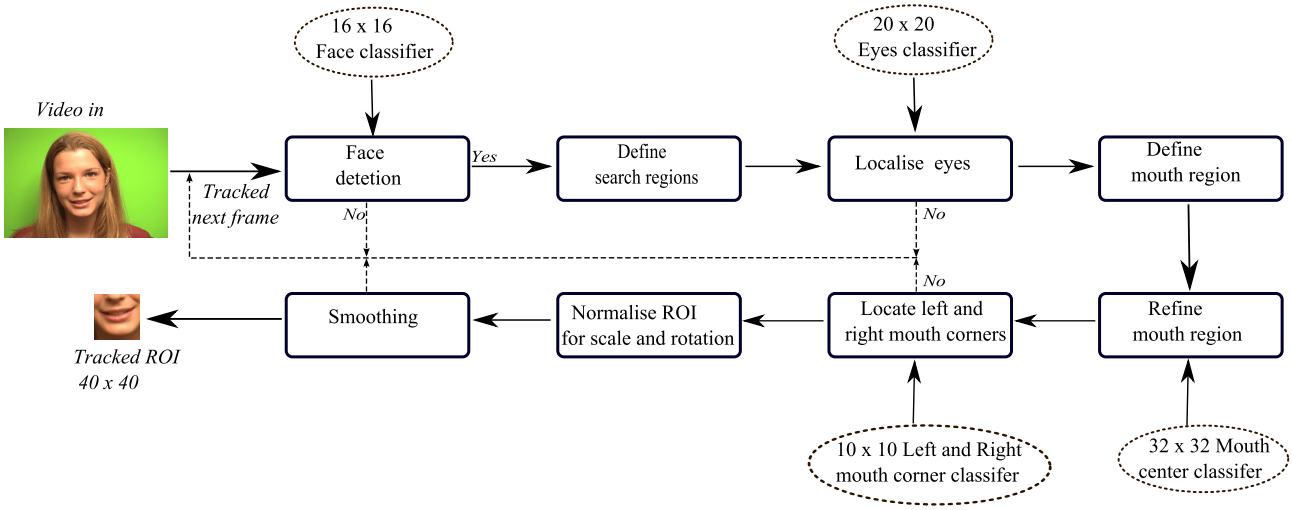


Figure 2: *Block diagram of the visual front end system*

Figure 3: *Examples of tracked 40×40 ROI*

database are shown in Figure 1. For the purposes of this work, only the sections while the speaker remained stationary were used in order to simplify the task of the visual front end.

3. The Visual front-end

Before any visual features can be extracted, a visual front-end has to be developed which is able to track and locate the region of interest (ROI) from the speaker's face. This was done using the Viola-Jones algorithm [15]. An overview of the visual front-end system which was used to extract the mouth ROI for the frontal pose is presented in Figure 2.

Given a video of a speaker, the face is located using a 16×16 face classifier. Once the face was located, the eyes were located in the upper half of the face region. Next, the lower half of the face was used to locate the mouth region. The resulting mouth region was then used as the search region to locate the right and left mouth corners. After these were located, the extracted mouth ROI was then rotated so that these two points were aligned horizontally. Finally, the mouth ROI was downsampled to 40×40 to keep the dimensionality low. The tracked ROI was smoothed using a mean filter. This process was performed on every incoming video frame. All the classifiers were developed using the OpenCV [16] libraries.

Overall, the performance of the visual front-end system was quite good. There was only a few number of poorly or mistracked ROIs, which could be attributed to random head movement. Figure 3 presents some of the examples of mouth ROIs.

4. Cascading appearance-based visual features

Cascading appearance based features, first devised by Potamianos et. al [10]. This technique has been established as the state-of-the-art for visual feature extraction for AVASR [11, 12, 13]. Figure 4 shows a block diagram of the cascading appearance-based visual feature extraction system. Essentially, it is broken into two sections: 1) static feature extraction and 2) dynamic feature extraction. The following subsections describe these in detail.

4.1. Static feature extraction system

Following the ROI extraction from the visual front-end system, a image mean normalization step was performed to remove any irrelevant information, such as illumination or speaker variances. The mean image was calculated from the given entire utterance and subtracted from the every incoming frame in the utterance. Then a two-dimensional separable, discrete cosine transform (DCT) is applied to the mean-removed image as shown in Figure 4. Finally, the top 30 higher energy components were selected to capture the static information.

4.2. Dynamic feature extraction system

Visual speech is best discriminated by the movement of the visual articulators [10]. The best features for representing visual speech are generally considered across a small window of around 5 to 7 frames, rather than within just one frame. One technique that can extract such information is through the use of linear discriminant analysis (LDA) to extract the relevant dynamic speech features from the ROI.

In order to incorporate the dynamic speech information, the static features around the current frame are concatenated into a single feature vector. We used seven of these neighboring static feature vectors over ±3 consecutive frames, and were projected via an inter-frame linear discriminant analysis (LDA) step to yield a 50-dimensional "dynamic" visual feature vector, extracted at the video frame rate of 30 Hz. The classes used for the LDA matrix calculation were hidden Markov model states, based on a forced alignment of known-good, whole-word acoustic models with the aligned acoustic CUAVE speech.

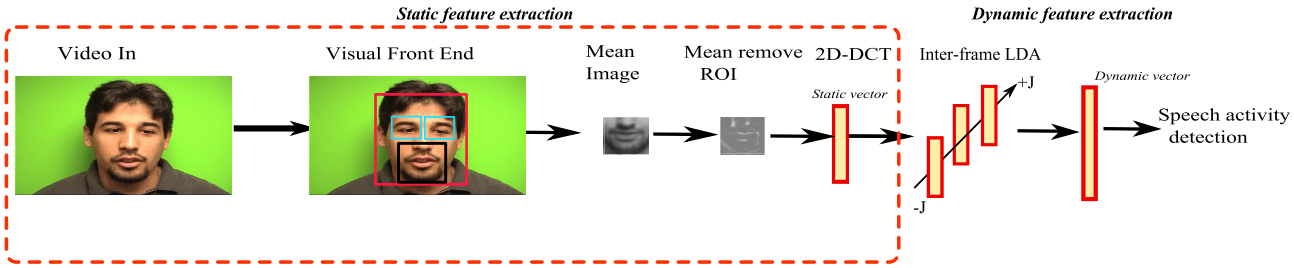


Figure 4: *Block diagram of the feature extraction system: Given the input video the face and facial features are tracked via the visual front-end. From the resulting ROI, visual features are extracted based on a combination of 2D-DCT and LDA techniques. These features are then used to speech activity detection (dashed box represent the static-only feature extraction system)*

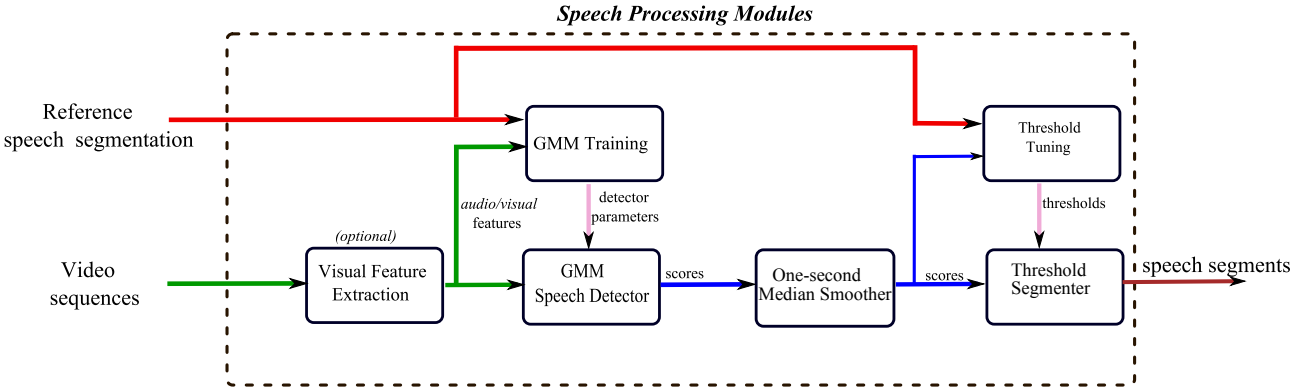


Figure 5: *An overview of the speech detection framework*

5. Gaussian mixture model (GMM) based VAD

5.1. VAD framework

An overview of the operation of the main components of speech detection framework is outline in Figure 5. We describe each of these modules in the following subsections.

5.1.1. GMM speech detectors

An 8-mixture GMM-based speech detection module was used to calculate the speech likelihood of each individual feature-frame obtained from the feature extraction outlined in Section 4. Each speech likelihood score was calculated as the difference between a speech and a non-speech GMM for the particular set of features under observation. A separate GMM training module was used to estimate the parameters of the speech and non-speech GMMs on separate training data.

5.1.2. Score smoother

The speech likelihood scores obtained for each frame were smoothed using a one-second median filter to attenuate the effect of short-term variation. Removing the short-term variation in the GMM-based speech likelihoods was found to provide better performance; hence providing the ability to handle the effect of long vowels.

5.1.3. Speech segmenter

The smoothed speech likelihood scores were then segmented into speech and non-speech decisions according to a simple threshold segmenter. Frames above the threshold were designated speech, and those below, non-speech. Similarly to the GMM parameter estimation, the segmentation threshold was trained on separate training data.

Table 1: *Speaker list*

Group	Speakers
Training	s01, s02, s03, s05, s06, s08, s09, s10 s11, s12, s13, s14, s15, s16, s18, s19 s20, s21, s22, s23, s24, s25, s26
Testing	s27, s28, s29, s30, s31, s32, s34, s36

5.2. Evaluation protocol

The VAD experiments were conducted using the stationary, frontal-view portion of the CUAVE database. Only the isolated digits section, covering 5 repeats of 10 English digits ('zero' - 'nine'). For the purposes of our VAD experiments, the entire 10-digit sequences were considered to be speech, while the silences between the 10-digit sequences were designated non-speech. From the CUAVE database 31 subjects (5 subjects were discarded due to the poor tracking) were selected for the experiments and they were categorised as 23 subjects for training of the GMM models and for the tuning of the segmentation thresholds based on the minimising the half total error rate (HTER) and 8 subjects for testing. The training and testing speakers are listed in Table 1.

In order to evaluate the performance of the VAD system, the

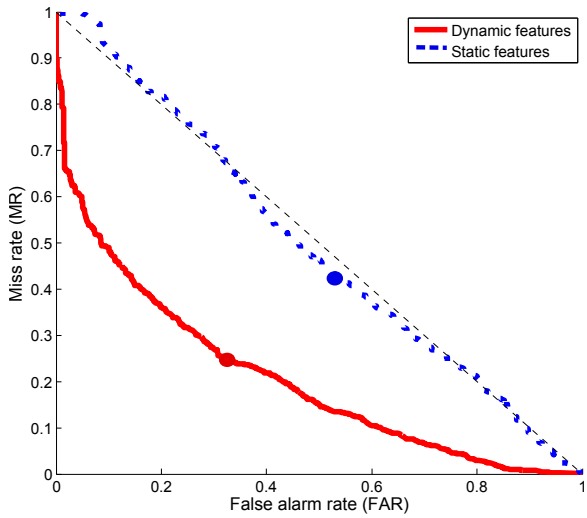


Figure 6: *Variation of receiver operating characteristic (ROC) curve with static and dynamic features*

output speech segmentation was compared with the reference segmentation derived from CUAVE word-level transcriptions. The differences between these two segmentations are reported using the false alarm rate (FAR) (i.e. *How often non-speech frame is detected as a speech frame*), miss rate (MR) (i.e. *How often a real speech frame is miss*) and the HTER. HTER define as the mean of the MR and FAR. The metrics are calculated as follows in this framework. Each of the metrics are calculated over a wide variety of speech sequences.

$$MR = \frac{T_{fa}}{T_{sys}} * 100\% \quad (1)$$

$$FAR = \frac{T_{m}}{T_{ref}} * 100\% \quad (2)$$

In equation (1), T_{fa} represents the duration of speech in false-alarm and T_{sys} represents the duration of speech in the system. In equation (2), T_m define as the duration of speech misses and T_{ref} represents the duration of the reference event transcriptions.

6. Results

6.1. Static feature results

Initially, the VAD experiments were run using only the static features. These experiments are shown as the thick dashed line in Figure 6, showing the trade-off between the MR and the FAR at each possible segmentation threshold. The chosen operating point, based on minimising the HTER, on the curve is also indicated.

Using the static-only features, the FAR was 55.20% and the MR was 41.40% which resulted in a HTER of 48.30%. The overall results were very poor. The main reason for this is that there is little distinction in visual information represented by the static lip features between the speech and non-speech events.

6.2. Dynamic feature results

The next experiments were conducted to obtain the dynamic cascading appearance based visual features. These experiments are shown as the solid line in Figure 6, with the chosen operating point, based on minimising HTER, also indicated. As can be seen the inclusion of dynamic features shows great improvement in performance. Minimum HTER was 28.50% with 31.70% FAR and 25.30% MR. This is 19.80% of improvement in HTER compared with only static features. Table 2 shows the comparison of static and the dynamic feature results.

Table 2: *Comparison of static and dynamic results*

Performance metrics	Static results (%)	Dynamic results (%)
FAR	55.20	31.70
MR	41.40	25.30
HTER	48.30	28.50

As an example, the results of an individual testing speaker obtained from the visual VAD system is shown in Figure 7 for both (a) static and (b) dynamic visual feature extraction. The clear improvement of the dynamic visual features can be seen here, showing the importance of capturing the lip movements for visual VAD.

7. Conclusion and future work

This paper presents a visual VAD system that uses the cascading appearance based features developed by Potamianos et. al [10]. The results show that using the dynamic features clearly improves performance, rather than just utilising static information. The paper also described an efficient visual front-end system to extract the ROI and a speech activity detection framework. This research clearly showed the importance of the lip movements of the visual articulators not only for the speech recognition but also for the visual VAD.

Our current research is focus on VAD using different dataset which has significant amount of "visual silence" in frontal view and profile view as well as visual VAD experiments in a "real-world" environment.

8. Acknowledgment

We would like to thank Clemson University for freely supplying us CUAVE database [14] for our research. This work was supported through the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

9. References

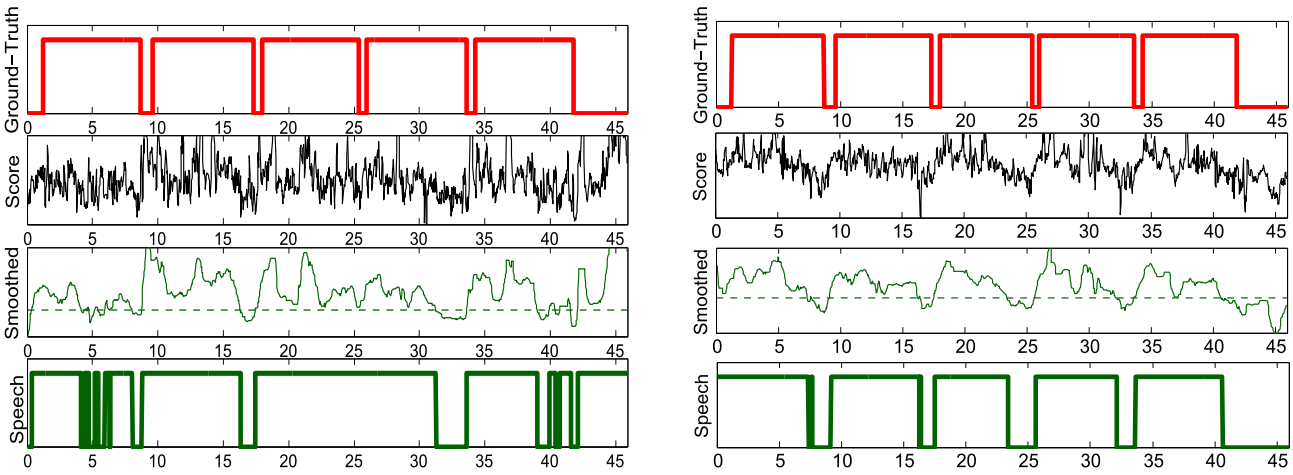
[1] A. Sangwan, M. Chiranth, H. Jamadagni, R. Sah, R. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the internet," *IEEE International Conference on High-Speed Networks and Multimedia Communications*, pp. 46–50, 2002.

[2] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the PAN-european digital cellular mobile telephone service," *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 369–372.

[3] J. Sohn, N. Kim, and W. Sung, "A statistical model based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, pp. 1–3, 1999.

[4] Y. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, pp. 276–278, 2001.

[5] S. Gazor and W. Zhang, "A soft voice activity detector based on



(*a*) With static features (*b*) With dynamic features

Figure 7: *Results of VAD on a sample testing speaker with (a) static and (b) dynamic visual feature extraction*

a Laplacian-Gaussian model," *IEEE Transaction on Speech and Audio Processing*, vol. 11, pp. 498–505, 2003.

[6] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," *Proceedings of the EUROSPEECH 2003*, Geneva, 2003.

[7] P. Liu and Z. Wang, "Voice activity detection using visual information," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 609–612, 2004.

[8] D. Sodoyer, B. Rivet, L. Girin, J. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2006.

[9] V. Libal, J. Connell, G. Potamianos, and E. Marcheret, "An embedded system for invehicle visual speech activity detection," in *Proceedings of the International Workshop on Multimedia and Signal Processing*, Chania, Greece, 2007, pp. 255–258.

[10] G. Potamianos, A. Verma, C. Neti, and S. Iyengar, G. Basu, "A cascade image transform for speaker independent automatic speechreading," *IEEE International Conference on Multimedia and Expo 2000, ICME 2000.*, vol. 2, pp. 1097–1100, 2000.

[11] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *in Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[12] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1293–1296, Geneva, Swizterland, 2003.

[13] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," *in Proc. Int. Works. Multimedia Signal Process. (MMSP)*, pp. 24–28, 2006.

[14] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL,USA ,2002.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition, 2001. CVPR 2001*, vol. 1, pp. 511–518, 2001.

[16] *Open Source Computer Vision Library*, Std. [Online]. Available: http://www.intel.com/research/mrl /research/opencv