Voice Activity Detection for Monaural Speech Enhancement Using Visual Cues

S. Balasubramanian, R. Rajavel, and S. Shoba

Abstract Voice activity detection (VAD) is a vital module in various applications like speech recognition, speech enhancement, and dominant speaker identification. The performance of voice activity detectors using audio cues considerably declines under low SNR conditions. One method to improve the performance of VAD is to use the video signal, which is independent to acoustic background. At the present time, video calls have become a more popular way of communication, and recent products like laptops and smartphones, have cameras and microphones inbuilt. The accessibility of a video signal, along with the audio signal, can be used for voice activity detection particularly in a noisy environment, as the video signal is independent to background acoustic noise. This paper aims to develop a binary mask for voice activity detection to separate the target speech from the noisy speech mixture using the visual cues. The visual cues will be extracted by the mouth detection using Viola–Jones algorithm and the lip movement tracking using Kanade–Lucas–Tomasi (KLT) algorithm. Finally, the extracted mask using visual cues is compared with the mask obtained using audio cues under low SNR conditions. The performance is evaluated for the proposed system using PESQ. The experimental result shows that the proposed system performs well under low SNR conditions and improves the average PESQ score of 0.57 as compared to the other existing systems which use only auditory cues for voice activity detection.

Keywords Voice activity detector · Visual cues · Binary mask · Visual-VAD (V-VAD) · Audio-VAD (A-VAD)

S. Balasubramanian (✉)
Meenakshi Sundararajan Engineering College, Chennai, India

R. Rajavel
SSN Engineering College, Chennai, India
e-mail: rajavelr@ssn.edu.in

S. Shoba
VIT University (Chennai Campus), Chennai, India
e-mail: shoba.s@vit.ac.in

A. Sivasubramanian et al. (eds.), *Futuristic Communication and Network Technologies*,
Lecture Notes in Electrical Engineering 792,
https://doi.org/10.1007/978-981-16-4625-6_25

1 Introduction

In noisy environment, the speech spoken by a person will be affected by the noise. For example, the noise can be other person speaking, noise due to a passing car, and many other sounds. Many applications need a system that separates the speech from the noisy speech and enhancement of the target speaker's speech. In cellular phone communication, the voice will be affected by the surrounding noise at the transmitter. A speech enhancement system can be used at the receiving end to improve the quality of the speech.

In speech separation process, the target speech signal is separated from acoustic mixture. The acoustic mixture may be another speech or environmental noise or both. Speech separation can be used in speech/speaker recognition, voice communication, air–ground communication, hearing aids, etc. Spectral subtraction, subspace analysis, hidden Markov modeling, and sinusoidal modeling are some of the methods propose earlier proposed for monaural speech separation. These approaches usually require a prior knowledge of noise signal. In the last few decades, many researchers have developed monaural speech separation using adaptive energy threshold [1], and image analysis techniques as in [2, 3]. The performance of the speech is improved using genetic algorithm-based fusion scheme in [4], and speech intangibility is improved by fusing voiced and unvoiced speech segments as in [5]. All these methods used audio as the only input to improve the speech quality.

Visual cues refer to the voice activity of the target speaker which is used in this paper to produce an enhanced output of the target speaker's speech. Normal speech enhancement techniques which are used to enhance a single (monaural) person speech do not provide the expected result.

Lip-reading depends on number of factors in human-to-human communication [6]. The quality of visual information plays a vital role. For example, in poor lighting condition it is hard to detect the shape of the mouth. Additionally, it is difficult to detect visual cues as the listener and the speaker are moving apart.

Hence simulating the target person's speech activity by tracking the lip movement of the person using Viola–Jones algorithm [7, 8] and Kanade–Lucas–Tomasi (KLT) [9] and comparing that with the original audio sample to ensure that the presence of voice activity from the lip moment matches with the original audio onsets and offsets.

The presence of speech activity and the voice activity can be detected from both audio and video stream. If the SNR is high, speech is dominant compared to noise; thus, it is more reliable to detect the presence of speech activity from audio stream itself. But when the SNR reduces, and as the noise dominates speech, it is not reliable to detect onsets and offsets from audio stream as it may treat some noisy parts as speech or vice-versa. In such cases, it is advisable to detect onsets and offsets from video stream as it is independent of SNR of the signal. The onset and offset times of video stream can be detected by tracking the mouth of the target speaker. The detected onset and offset time of the video stream is then plotted and compared with the plot of audio stream, and checked for one-to-one correspondence.

The reminder of this paper is organized as follows. Section 2 describes the proposed system about speech onset and offset detection using video stream. The existing system of speech onset and offset detection using audio stream is explained in Sect. 3. The experimental results are given in Sect. 4. Conclusion is outlined in Sect. 5.

2 Speech Onset and Offset Detection Using Visual Stream

In this paper, the main focus is to obtain a voice activity detection mask using visual cues (V-VAD) for target speech detection. The mask here is a binary mask which specifies the absence or presence of voice activity in each frame. First the video stream is splitted in sequence of image frames. Second, the face is detected in the given frame and the mouth region is encompassed with the bounding box in the detected face. The first frame is processed using the Viola–Jones algorithm [7, 8, 10] to get the bounding box for the mouth region. Viola–Jones algorithm uses Haar features to detect the face. The four stages of Viola–Jones algorithm are: (a) choice of Haar feature, (b) creation of integral image, (c) Adaboost training (d) cascading classifiers.

2.1 Binary Mask Detection

The mouth region in the image is identified using Viola–Jones algorithm. It uses a cascade of classifiers to detect the presence of target object in the image frame. Every stage in the cascade rejects the region which has no target object. As the sliding window travels over the image it is possible to produce multiple object detection near the target object. This multiple identification is combined to produce one bounding box for one target object. The bounding box is represented using the top left corner coordinates and the height and width of the box. Figure 1 highlights the bounding box of the mouth region detected.

2.2 KLT Feature Tracking

Kanade–Lucas–Tomasi (KLT) algorithm [9, 11, 12] is used for tracking the two points in the video frames. KLT tracking algorithm works with two simple steps: in initial frame tracking features are identified, and then tracks all the detected features in the remaining frames. Assume the first and the next images were taken at time t and $t + \tau$, respectively. Number of frames per second captured by the video camera determines the time τ. Let an image be represented as a function of two variables x and y. Now the variable t is added to represent the time at which the image was

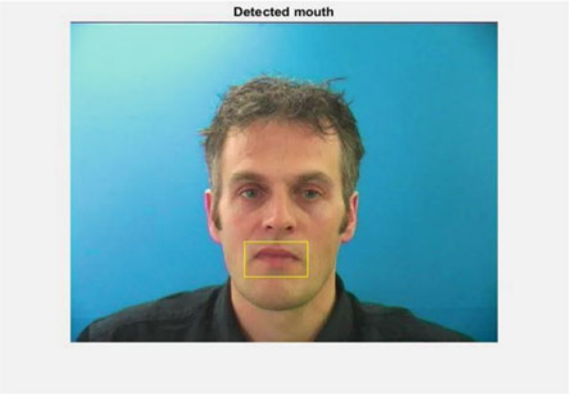

Fig. 1 Mouth detection of first frame

captured by the camera. Now any point in an image is defined by the function $f(x, y, t + \tau)$. The assumption made the KLT tracking algorithm is

$$f(x, y, t + \tau) = f(x - x, y - y, t) \tag{1}$$

From Eq. (1), it is understood that each point in the first frame is shifted by an amount (x, y) to obtain the second frame. This shifting amount is represented by displacement $d = (x, y)$, and the objective of tracking is to compute d. The two feature points taken from the first frame in our system for tracking throughout the frames are given as $[a + (w/2), b]$ and $[a + (w/2), b + h]$ where (a, b) are top left corner coordinates and w is the width and h is the height of the bounding box respectively.

Displacement between the feature points is calculated from one frame to other fame. Lip movement is computed using the displacement. This displacement calculation is used to find the onset and offset of speech using video. This decision will form the binary mask for speech enhancement. The steps involved in getting the binary mask using visual stream are also shown in Fig. 2.

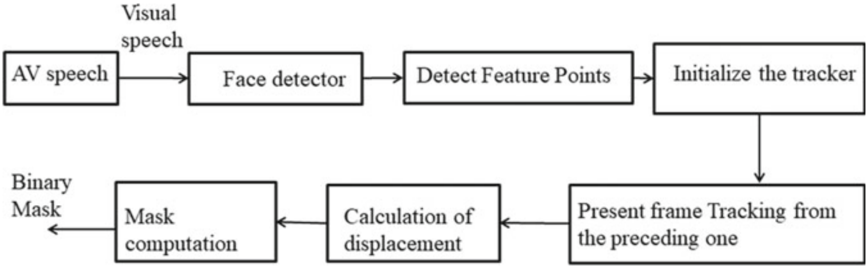

Fig. 2 Binary mask detection using video stream

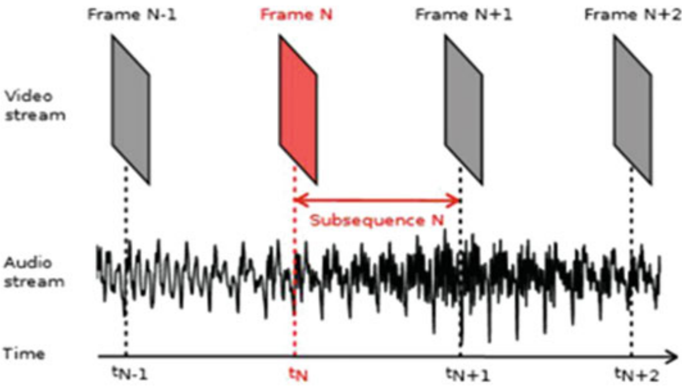

Fig. 3 Audio subsequence corresponding to a frame

2.3 Linear Interpolation

The number of video frames is significantly smaller than the number of determined audio frames. But the voice activity detection in audio stream and video stream is compared frame by frame for one-to-one correspondence. The original visual frames are interpolated to equate the number of audio stream frames. The onset/offset decision between two visual frames N and $N-1$ is the onset/offset decision of Nth frame as in Fig. 3. The general formula to determine the number of frames in a given audio stream is that

$$\mathrm{NbFr} = (\mathrm{length}(x[n] - \mathrm{FL} + \mathrm{FS})/\mathrm{FS} \qquad (2)$$

where NbFr = Number of audio frames, $x[n]$ = Audio samples, FL = Frame Length, FS = Frame Shift.

3 Speech Onset and Offset Detection Using Audio Stream

Varieties of information's are present in acoustic speech signal. In existing system, the voice activity detection using only audio is implemented as in [13]. The first step involved in conversion is framing. The input audio signal is segmented into 30 ms frames with 10 ms overlap between successive frames, and then Fourier transform is applied to obtain the frequency domain according to the Window, sidelobe attenuation, and FFT length properties. Using rectangular window will introduce high-frequency noise at the beginning and end of every frame. Hamming window is used to reduce this edge effect. The signal is then represented in power domain.

According to [14], the noise variance is estimated. According to [15], the posterior and prior SNR are estimated. The probability of speech present in the current frame is calculated using hidden Markov model (HMM) and log-likelihood ratio test according to [13]. Based on the probability of each frame, onset/offset decisions are made in the existing method.

4 Experimental Results

The experiment is first conducted by taking sample test videos from an audio–visual database like 'GRID Corpus,' where sample videos of both male and female speakers are available of the same duration. The noise samples with which the clean speech is mixed are taken from the 'Noisex92' database.

The frontview videos from the GRID corpus show the speakers uttering the same sentence consisting of voiced, unvoiced and silent segments to carry out the experiment. The mask using visual cues for voice activity detection is obtained using Viola–Jones and KLT algorithm. The target speech is separated using the mask obtained using visual cues. We have tested our system using three short videos from GRID corpus consisting of a total of (297 * 3) frames by mixing noise at $-$ 5 dB to $+$ 5 dB SNR conditions. The target speech separated using visual cues mask is compared with the target speech separated using mask obtained by the audio stream.

The experiment result is displayed in Table 1. The performance is evaluated using perceptual evaluation of speech quality (PESQ). PESQ score varies from 4.5 to $-$ 0.5, higher scores better is the quality.

5 Conclusion

We have presented a mask using visual cues for voice activity detection. The proposed system detects the face and the mouth region to effectively distinguish the speaking from the silent frames in low SNR conditions. The binary mask presented here is independent of noise and effective for speech enhancement.

On comparing, the proposed VAD using visual cues to VAD using audio cues for three videos each of 297 frames it is evident that at low SNR conditions the target speech enhanced using visual cues are having higher PESQ score than target speech enhanced using audio cue.

Extremely low lighting situations and faces that move away considerably from a frontal pose or are too far from the camera to provide enough information for the mouth would cause the system to perform poorly.

Table 1 PESQ improvement of the proposed system

Video	Type of noise	SNR (dB)	PESQ score visual-VAD (proposed)	PESQ score audio-VAD (existing)	PESQ score improvement
Video 1	Babble	5	1.96	1.38	0.58
		0	1.85	0.96	0.89
		− 5	1.79	1.48	0.31
	Factory	5	2.05	1.5	0.55
		0	1.72	0.91	0.81
		− 5	1.47	0.67	0.8
Video 2	Babble	5	1.67	1.23	0.44
		0	1.45	0.98	0.47
		− 5	1.28	0.93	0.35
	Factory	5	1.67	1.4	0.27
		0	1.5	0.95	0.55
		− 5	1.26	0.92	0.34
Video 3	BABBLE	5	1.87	1.17	0.7
		0	1.59	0.96	0.63
		− 5	1.53	0.82	0.71
	Factory	5	1.92	1.37	0.55
		0	1.64	0.69	0.95
		− 5	1.39	0.88	0.51
	Average		1.64	1.06	0.57

References

1. Shoba S, Rajavel R (2017, April) Adaptive energy threshold for monaural speech separation. In 2017 international conference on communication and signal processing (ICCSP), pp 0905–0908. IEEE
2. Shoba S, Rajavel R (2018) Image processing techniques for segments grouping in monaural speech separation. Circ Syst Signal Process 37(8):3651–3670
3. Sivapatham S, Ramadoss R (2018) Performance improvement of monaural speech separation system using image analysis techniques. IET Signal Proc 12(7):896–906
4. Shoba S, Rajavel R (2020) A new Genetic Algorithm based fusion scheme in monaural CASA system to improve the performance of the speech. J Ambient Intell Humaniz Comput 11(1):433–446
5. Shoba S, Rajavel R (2019) Improving speech intelligibility in monaural segregation system by fusing voiced and unvoiced speech segments. Circ Syst Signal Process 38(8):3573–3590
6. Chen T, Rao RR (1998) Audio-visual integration in multimodal communication. Proc IEEE 86(5):837–852
7. Viola P, Jones M (2004) Robust real-time face detection. Int J Comput Vision 57:137–154
8. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In Proceedings of IEEE CVPR, Kauai, HI, USA, December 9–14, vol 1, pp 511–518
9. Tomasi C, Kanade T (1991, April) Detection and tracking of point features, Computer Science Department, Carnegie Mellon University

10. Lienhart R, Maydt J (2002) An extended set of Haar like features for rapid object detection. In: IEEE ICIP, Rochester, NY, USA, September 22–25. vol 1, pp 900–903
11. Shi J, Tomasi C (1994) Good features to track. In: IEEE conference on computer vision and pattern recognition, pp 593–600
12. Kalal Z, Mikolajczyk K, Matas J (2010) Forward-backward error: automatic detection of tracking failures. In: Proceedings of the 20th International conference on pattern recognition, pp 2756–2759
13. Sohn J, Kim NS, Sung W (1999) A statistical model-based voice activity detection. Signal Process Lett IEEE 6(1)
14. Martin R (2001) Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Processing 9(5):504–512
15. Ephraim Y, Malah D (1984) Speech Enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 32(6):1109–1121