Discriminating Speech and Nonspeech from Video Signals using SFF VAD

Avani S Babu[1], Amrutha V Nair[2]

[1]M .Tech student, Department of Communication Engineering, Sree Buddha College of Engineering for Women, Elavumthitta, Pathanamthitta, Kerala, India

[2]Assistant Professor, Department of Electronics and Communication Engineering, Sree Buddha College of Engineering for Women, Elavumthitta, Pathanamthitta , Kerala, India.

Abstract: ***An image processing approach is used for speech/nonspeech discrimination. The approach is based on single frequency filtering (SFF) and visual VAD. SFF is the amplitude envelope of the signal is obtained at each frequency with high temporal and spectral resolution where visual VAD is a classifier to determine whether a speaker is silent or not in a frame using the associated video signal. The high resolution property of SFF helps to exploit the resulting high signal-to-noise ratio (SNR) regions in time and frequency. But in SFF method, nonspeech is also considered as speech in the audio signal at particular situations. To avoid this issue, a technique is proposed with the combination of SFF and Visual VAD in which the speech is extracted from the video signals by the lip movement. In this method uses lip shape and degree of lip opening as visual features representing a subject's lip motion. After the lip movement analysis, the audio analyzed output and video analyzed output is combined together to distinguish the voiced/unvoiced region with a SVM classifier.***

Keywords: Single Frequency Filtering (SFF), Voice Activity Detection (VAD), spectral resolution, lip motion, Support Vector Machine (SVM)

1. Introduction

Voice activity detection (VAD) is a technique used in speech processing in which the presence or absence of human speech is detected. The objective of VAD is to discriminate speech and nonspeech in the acoustic signal. In the case of human, they are able to differentiate speech and nonspeech. But it is not possible for a machine to discriminate the voiced/unvoiced regions [1]. For such reasons VAD is used. It can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session. One of the applications of VAD is that it can avoid unnecessary coding/transmission of silence packets in Voice over Internet Protocol applications. VAD is an important enabling technology for a variety of speech-based applications. Therefore various VAD algorithms have been developed that provide varying features. It also compromises between sensitivity, accuracy, latency and computational cost.

The typical design steps of a VAD algorithm are the following:

1) There may first be a noise reduction stage.
2) Then some features or quantities are calculated from a section of the given input signal.
3) A classification rule is applied to classify the section as speech or non-speech.

Single Frequency Filtering (SFF) is the method of extracting the energy at a single frequency in the audio signal. It is the amplitude envelope of the signal that is obtained at each frequency with high temporal and spectral resolution. The high resolution property helps to enslave the resulting high signal to noise ratio (SNR) regions in time and frequency. This method does not use training data to derive the characteristics of speech or nonspeech. This SFF method gives better performance than the Adaptive Multi-rate (AMR) method. But there is a demerit in the discriminating process when we use the audio signal. The demerit is that sometimes speech will consider as nonspeech and vice versa. To avoid this video signal is used instead of audio signal. In this proposed method, visual VAD is used along with the SFF VAD. Visual VAD is a classifier used to discriminate speech and noise in a frame using the associated video signal. It mainly uses the visual features of the mouth region. The position of the lip and the motion of the lip are the features used for discrimination. In the proposed system, the output of the audio and video frame is combined together to identify the voiced/unvoiced region.

Section II Describes the method used in the Proposed System. Section III gives results of evaluation of the proposed system based on the corresponding VAD. Section IV gives a summary.

2. Discrimination between Speech and Nonspeech

The flow chart of the proposed system is shown in Figure 1. In this process a video signal is given as input signal. The proposed system consists of two sections. One of the section is distinguishing the speech and noise in the audio section of the video and the second is the differentiating the speech from noise using the video frame. In the audio section, SFF VAD algorithm is used and in the video frames, visual VAD algorithm is used. A SVM classifier is used to classify visual features in the face shot. The details are explained below:

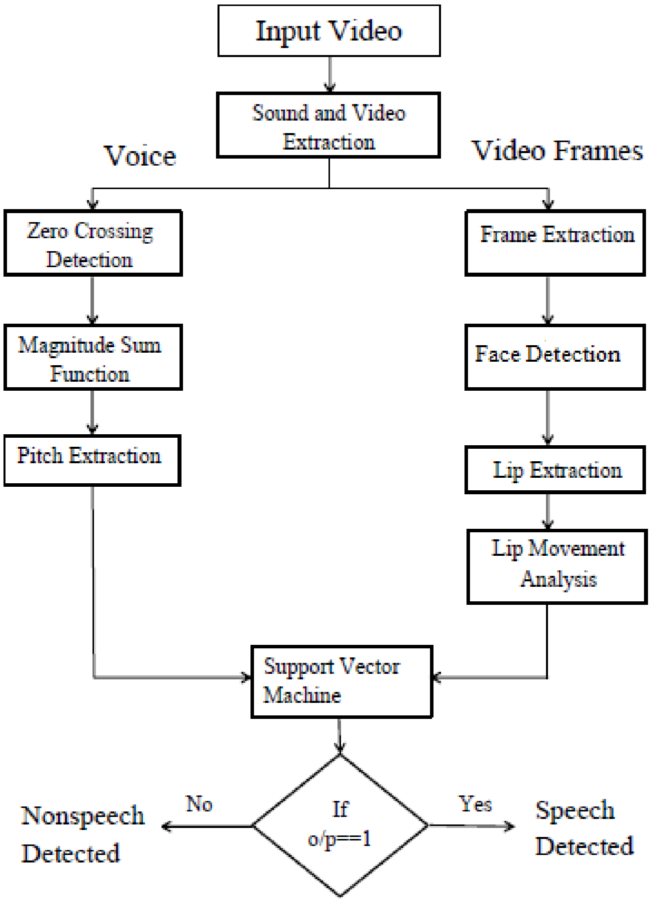


Figure 1: Flow chart of the proposed system

2.1 Audio features Extraction

In the proposed system, the left portion of the flowchart is the extraction of audio features from the input video signal. First step in this is the calculation of the envelope of the speech signal at each frequency. Weighted component of envelop is considered and a decision logic is applied to it.

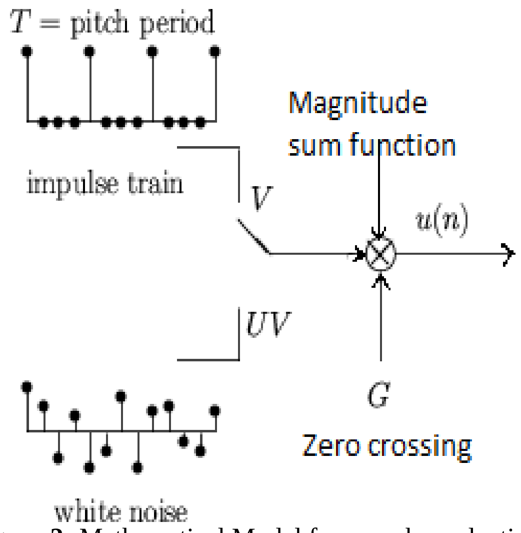


Figure 2: Mathematical Model for speech production

There are three features are considered in the audio signal. They are the following:

a) Zero crossing is the point at which a function crosses the horizontal axis as its value passes through zero and changes sign. It changes the sign from positive to negative.
b) Magnitude Sum Function is the sum of the absolute values of the given signal.
c) Pitch is the degree of highness or lowness of a tone.

Figure 2 shows the mathematical model of the speech production [2]. These three features are combined together to get the speech. A threshold value is given to separate differentiate the speech and the noise. The regions above the threshold is considered as speech and the below region as the nonspeech in the audio signal.

2.2 Visual features Extraction

Visual features include the facial expressions in the face of a speaker. Visual VAD is used in this section. Different processes are take place in this section. First process is the extraction of the each frame. After that the face of the speaker is detected. From the face, the position of the eyebrows, mouth, eyes and nose are obtained. The position of the lip is extracted. The analysis of the lip is based on the two factors: a) lip opening and b) lip shape [4].

a) Lip Shape: aspect ratio and its time derivative.
b) Degree of lip opening: area of lip region and its derivative.

Support Vector Machine (SVM) is used to classify these features after the lip movement analysis. The aspect ratio, degree of lip opening and their derivatives are given to the classifier.

Final process is the combining process of the output of the audio stream and the video frame after the analysis. Decision logic is applied to this process. If the output of the audio stream and video stream is 1, then t is considered as speech and otherwise as nonspeech.

3. Simulation Results

Matlab R2013a is used as the simulation tool to perform the task. MATLAB (Matrix Laboratory) is a programming language developed by Math Works.

In the proposed method, video signal is given as the input. Sound and video is extracted from the video signal and they are analyzed separately. For the analysis of the audio, the SFF VAD algorithm is used and for the visuals, visual VAD algorithm along with certain process is used.

In the audio section, zero crossing is applied and the threshold value is calculated for the signal. Similarly, the magnitude sum function and pitch extraction is done to the extracted audio. After the process, their corresponding threshold value is calculated. This threshold value will determine the whether the speech is occurred or not. Figure 3 shows the output of the audio after the analysis. The three processes are plotted separately. In the graph, the blue colors represent the audio from the video signal and the final output

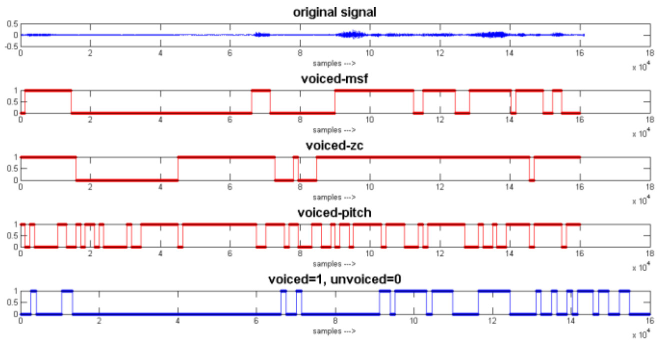
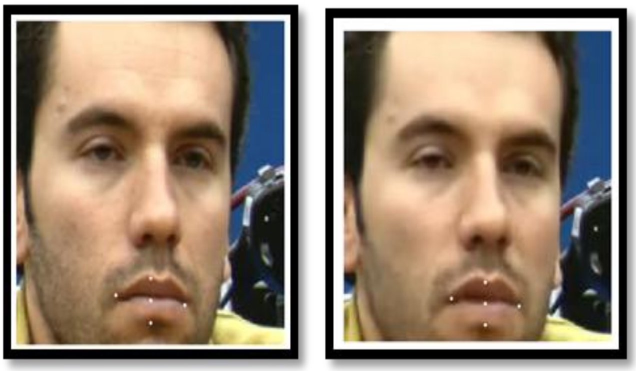


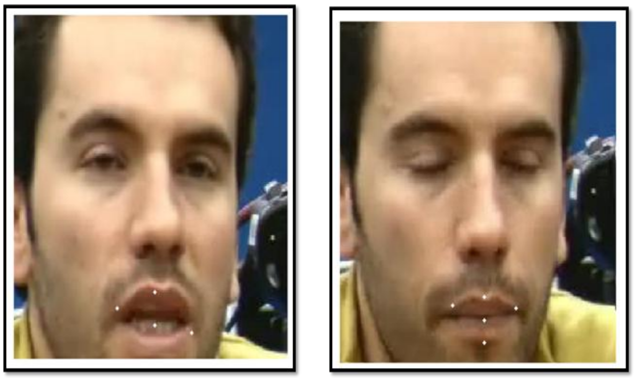
Figure 3: Audio output

after the analysis. The red colors represent the zero crossing (zc), magnitude sum function (msf) and the pitch. The value above the threshold is logic 1 and the value below the threshold is the logic 0.These three features are combined together to obtain the final output. Logic AND operation is used in the final process. Speech is obtained when the three features give 1 as output otherwise it is considered as nonspeech for logic 0. Audio analysis will not provide the actual speech and nonspeech. So the visual analysis of the audio signal is also needed.

In the video frame, the first process is the extraction of each frame in the video. After that face of the speaker is detected and the position of the features in the face such as mouth, eyes, eye brows and lips are obtained. The next process is the lip extraction in the mouth region. The output of the lip extraction is shown in the Figure 4. The lip region is represented by the four positions. The left and right side of the lip is represented as e and d and the top and bottom is represented as the s and I respectively. The figures 4(a) – 4(d) shows the lip positions of the speaker. Figure 4(a) represents the silence of the speaker. Figure 4(b) shows the starting of the speech and the other figures 4(c) and 4(d) shows the continuation of the speech. Lip movement analysis is the next step after the lip extraction. In the analysis, the features used are the lip shape and the degree of the lip opening. The aspect ratio v_1 *(n)* is used to calculate the lip shape with the width and height of the lip. Time derivative of the aspect ratio is also considered in the analysis and it is represented by v_2 *(n)*. The degree of lip opening is obtained by the area of the lip, v_3 *(n)* and its time derivative, v_4 *(n)*.



(a) (b)



(c) (d)

Figure 4: Lip Extraction of the speaker. (a) The speaker is silent, (b) Speaker starts to speak, (c) and d) shows the variation of the lip during the speech.

The analyzed features are represented in a matrix called feature matrix. This feature matrix is given to the SVM classifier. This classifies the features and store in another matrix. And this matrix only has the value 0 and 1.Target matrix is created on the basis of the input video and the target matrix is different for the different videos. This is considered as the truth value. For the final evaluation of the video frame, the class matrix and the target matrix is compared and the compared value is used for the final stage.

The combining of the audio and video output is the final process in the proposed system. Figure 5 shows the final output. The audio stream use 48000 samples per second and the video stream uses 25 frames per second. It is difficult to combine these features with different unit. So the samples are converted according to the 25 frames per sec. The graph shows the audio output and video output and the combined output. Logic AND is used to combine the audio and visual output. In this, the logic 1 is considered as speech and the logic 0 is used as the nonspeech.

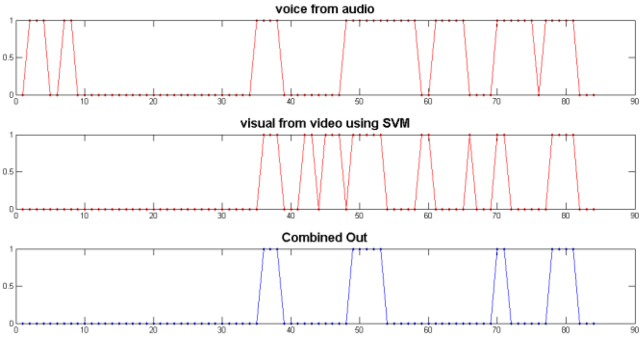


Figure 3: Output of the proposed system

4. Conclusion

The SFF method exploits the fact that speech has high SNR regions at different frequencies and at different times. The variance of speech across frequency is higher than that for noise, after compensating for spectral characteristics for noise. The spectral characteristics of noise are determined using the floor of the temporal envelope at each frequency, computed by the SFF approach. But it has certain limitations in extracting the speech and nonspeech from the audio signal. In this case nonspeech is also considered as speech and vice-versa.

To avoid this issue, a technique is proposed in which the speech is extracted from the video signals. This process is distinguishing the speech from the noise by the lip movement of the speaker along with the audio. In this, audio and video frames are differentiated from the video signal and according to the movement of the lip and audio, the speech is detected and the remaining is considered as non speech or noise.

5. Acknowledgment

I would like to express profound gratitude to our Head of the Department, Prof. Cherian Schariah, for his encouragement and for providing all facilities for my work. I express my highest regard and sincere thanks to my guide, Asst. Prof. Ms. Amrutha V Nair, who provided the necessary guidance and serious advice for my work.

References

[1] G. Aneeja and B. Yegnanarayana, “Single Frequency Filtering Approach for Discriminating Speech and Nonspeech”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 705-717, April 2015.

[2] Amol R. Madane, Zalak Shah, Raina Shah and Sanket Thakur, “Speech Compression Using Linear Predictive Coding”, IEEE Trans. Speech Audio Process., pp. 119-122, 2009.

[3] Qingju Liu, Andrew J. Aubrey and Wenwu Wang, “Interference Reduction in Reverberant Speech Separation with Visual Voice Activity Detection”, IEEE Transactions on Multimedia, vol. 16, No. 6, October 2014, pp. 1610-1623.

[4] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide and H. Murase, “Detection of Inconsistency Between Subject and Speaker Based on the Co-occurrence of Lip Motion and Voice Towards Speech Scene Extraction from News Videos”, Multimedia (ISM), 2011 IEEE International Symposium on, Dana Point CA, 2011, pp. 311-318.

[5] K. Han and D. Wang, “An SVM based classification approach to speech separation”, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, Prague, 2011, pp. 4632-4635.

[6] www.mathworks.in

Author Profile

Avani S Babu received the B-Tech degrees in Electronics and Communication Engineering from M.G University, Kerala at Sree Buddha college of Engineering for women in 2014. And now she is pursuing her M-Tech degree in Communication Engineering under the same university in Sree Buddha college of Engineering for women, Elavumthitta, Pathanamthitta.

Amrutha V Nair working as Assistant Professor in department of Electronics and Communication, Sree Buddha college of Engineering for women, Elavumthitta, Pathanamthitta.