Voice Activity Detection in Non-stationary Noise

LI Ye, WANG Tong, CUI Huijuan, TANG Kun

Abstract—Existing voice activity detection algorithms degraded severely in low SNR or in non-stationary noise environment. This paper proposes a new fusion method in such environment . The method is based on the fusion of the SNR for selected sub-bands of the input speech and this fusion is implemented through a specific function called SAF (sum of activation function). The results show that the algorithm could give reliable voice activity detection result in low SNR and even in the presence of non-stationary noise.

Keywords: Voice activity detection, speech recognition, vocoder

I. Introduction

The process of separating conversational speech and silence is called the voice activity detection (VAD). VAD is very significant in some speech communication applications. Precise VAD could improve the performance of speech recognition systems and reduce the coding rate of vocoders. Various VAD algorithms have been proposed. The earlier algorithms are based on the Itakura LPC distance measure[1], energy levels[2], difference of energy and zero-crossing rate[3], cepspectral features[4], higher order statistics[5], high frequency energy and low frequency energy[6].

Unfortunately, the existing algorithms have some problems in low SNR environment especially in the presence of non-stationary noise. So the paper proposes a new voice activity detection algorithm based on the fusion of the SNR for all selected sub-bands in the form of a function called SAF. The algorithm makes the best use of the frequency feature and reduces the dependency of the VAD on the frequency energy distribution of the noise. Especially, the algorithm can still give reliable VAD in the presence of non-stationary noise and suits for realtime implementation.

II. Proposed VAD algorithm

The input speech at 8k sampling rate is segmented into 32ms frames primarily and a 256-point FFT is done on each frame.

$$X(k)=\sum_{n=0}^{255}x(n)e^{-j\frac{2\pi}{256}nk}\quad,\quad k=0,...,255 \tag{1}$$

The algorithm divides the whole frequency filed (0~4000Hz) into 16 bands, therefore every 8 points (250 Hz) make one band. For each band, $S(f_i)$ is calculated, which is the total spectral energy in band i (named as f_i).

$$S(f_i)=\sum_{k=8i-8}^{8i-1}X(k)^2\quad,\quad i=1,...,16 \tag{2}$$

The frequency characters of variant noises are different and VAD based on the whole frequency energy hasn't made the best use of this feature. Such as the pink noise, in very low SNR, for example, 0 dB, the VAD based on the whole frequency energy will give poor detection. It could be

observed that even in very low SNR, the speech frame will still have some sub-bands with high SNR while the noise frame does not. This could be utilized for VAD. Fig 1 shows the SNR of each sub-band in both the speech frame and the noise frame in pink noise with SNR of 5 dB and 0 dB.

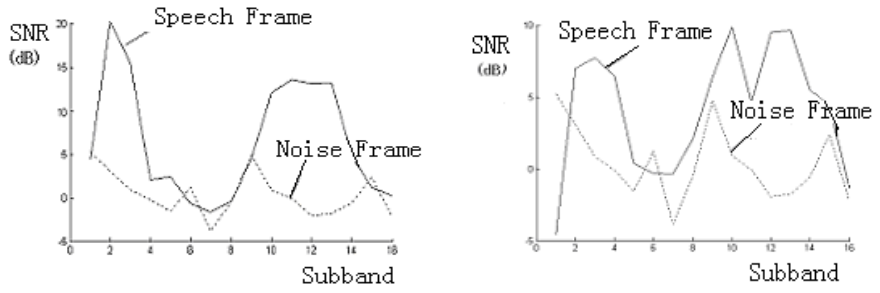

Fig.1. SNR of sub-bands for 5 dB (left) and 0 dB(right) in pink noise

It could be seen in Fig 1 that the speech frame still has some high SNR sub-bands even in 0dB SNR while noise frame does not. It also can be seen that SNR of the first frame and the last two frames are not so different whether they are speech frames or noise frames that they are not counted in the algorithm.

VAD algorithms usually assumed that the first n($n=5$ in the paper) input frames were noise and they could be used to estimate the initial noise character which would be updated later. So the initial energy of the sub-bands for noise frames, $N(f_i)$, could be simply estimated as the average of the $S(f_i)$ of the first n input frames. Since the n+1 frame, VAD is done for each frame as follows:

Step1: Compute the energy of selected sub-bands, $S(f_i)$, according to (2)

Step2: Compute the SNR of selected sub-bands

$$SNR(f_i) = S(f_i) / N(f_i), \qquad i = 2...14 \tag{3}$$

Step3: Compute the fusion feature H for VAD using the SAF function

$$H = SAF(SNR(f_2), SNR(f_3), ..., SNR(f_{14})) \tag{4}$$

$$SAF(SNR(f_2), SNR(f_3), ..., SNR(f_{14})) = \sum_{i=2}^{14} g(SNR(f_i)) \tag{5}$$

$$g(SNR(f_i)) = \begin{cases} 1, & if\ SNR(f_i) > T_a \\ 0, & else \end{cases} \tag{6}$$

The function SAF could be depicted as Fig 2. Then the feature H is smoothed,

$H = (H_{prev} + H)/2$, (H_{prev} was set zero before Step1) (7)

$$H_{prev} = H \tag{8}$$

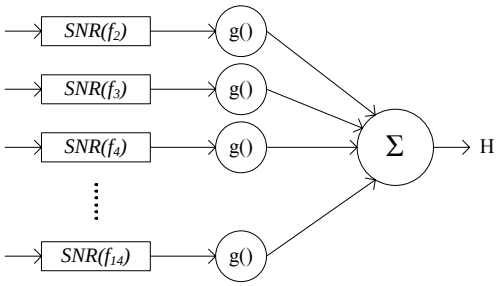

Fig.2. Process of SAF

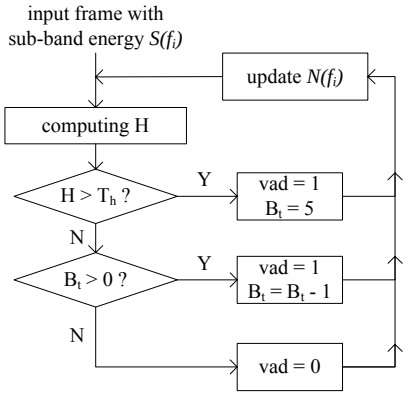

Fig.3. Process of VAD

Step4: VAD is made as follows, if $H > T_h$, then V_{ad}=1 and B_t is set as n; Else if B_t>0, then V_{ad}=1 and $B_t = B_{t-1} - 1$; Else V_{ad}=0. B_t is used for smoothing.

Step5: Update $N(f_i)$ and go step1, the update process of $N(f_i)$ is as follows:

$$N(f_i) = N(f_i) * (1-a) + S(f_i) * a \tag{9}$$

a was defined below, it would increase as the current frame resemble the noise frame more.

$$a = \max(0, \frac{\sum_{j=2}^{14} N(f_i) S(f_i)}{\sqrt{\sum_{j=2}^{14} N(f_i)^2} \sqrt{\sum_{j=2}^{14} S(f_i)^2}} - T_r) / 100 \tag{10}$$

Fig.3 shows the whole process of the proposed VAD. T_a、T_h、T_r and n were empirical constants.

III. Results

In the experiments, we evaluated the performance of the proposed algorithm in both white and tank noise. The evaluation file was recorded with 8 kHz sampling rates and stored as 16 bit integers. Fig.4. shows the VAD results (left) in white noise and (right) tank noise (SNR=0 dB).

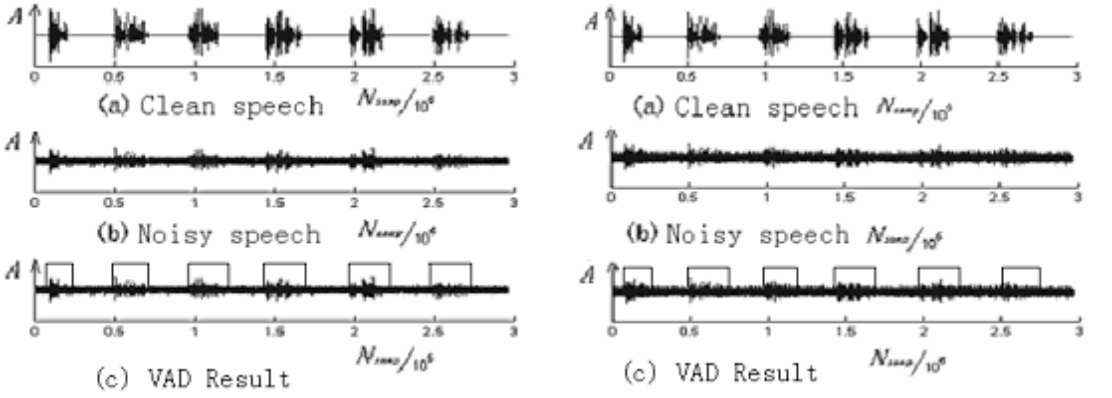

Fig.4. VAD results, (a) clean speech signal, (b) noisy speech signal, (c) speech signal after VAD，the part that VAD judges as speech was framed

The performance of the proposed algorithm was also evaluated in non-stationary tank noise, the SNR of which ranges from -5dB to 10dB. The VAD result of the proposed algorithm was compared with the common VAD algorithm that based on energy level and Woo's algorithm [6]. Fig.5. shows the VAD results.

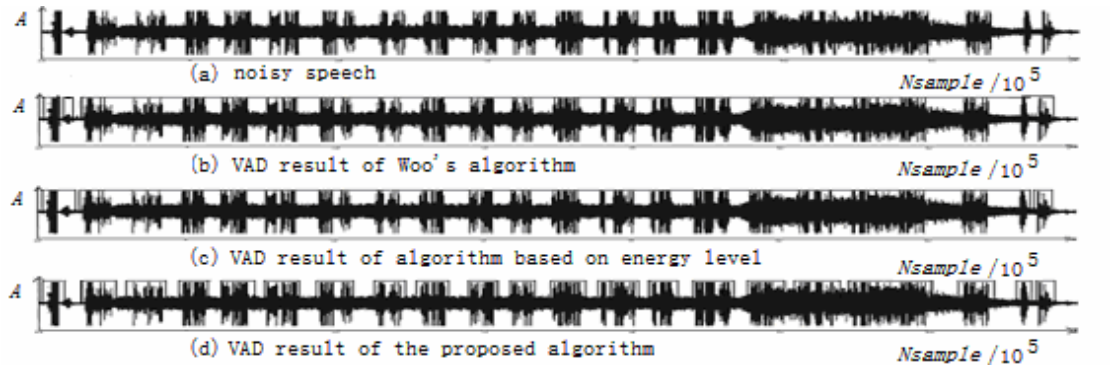

Fig.5. VAD results, the part which VAD judges as speech was framed, (a) Recorded noisy speech signal (tank noise), (b) Result of Woo's VAD, (c) Result of VAD based on energy level, (d) Result of the proposed VAD

Fig.5. shows that VAD algorithm based on energy level and Woo's VAD will judge noise as speech when the noise level rises quickly while the proposed algorithm could still give reliable VAD result and could be used in the presence of non-stationary noise.

IV. Conclusion

The commonly used VAD algorithms are shown to have problems in low SNR or in the presence of non-stationary noise. The paper proposed a new VAD algorithm based on the fusion of the SNR of selected sub-bands in the form of a function called SAF and tested the algorithm in both white and tank noise. The results show that the proposed algorithm could work well in low SNR or in non-stationary noise and suits for realtime implementation.

V. References

[1] Rabiner L R, Sambur M R. Voiced-unvoiced-silence detection using the Itakura LPC distance measure [A]. in Proc. Int. Conf. Acoust. Speech, Signal Processing, May 1977, pp. 323–326.

[2] Junqua JC, Mak B, Reaves B. A robust algorithm for word boundary detection in the presence of noise[J]. IEEE Trans on speech and Audio Processing, 1994, 2(3):406-412.

[3] Beritelli F, Casale S, Ruggeri G, et al. Performances evaluation and comparision of G.729/AMR/fuzzy voice activity detectors[J]. IEEE signal processing Letters,2002,9(3):85-88.

[4] Haigh J A, Mason J S. Robust voice activity detection using cepstral features [J]. in Proc. IEEE TENCON, China, 1993, pp. 321–324.

[5] Nemer E, Goubran R, Mahmoud S. Robust voice activity detection using higher-order statistics in the LPC residual domain [J] . IEEE Trans on Speech and Audio Processing,2001,9(3): 217—231.

[6] Woo K H, Yang T Y, Park K J，et al. Robust voice activity detection algorithm for estimating noise spectrum[J].Electronics Letters,2000,36(2):180—181.