A New Voice Activity Detection Method Using Maximized Sub-band SNR

Weiwu Jiang, Wai Kit Lo and Helen Meng
The Chinese University of Hong Kong
E-mail: {wwjiang, wklo, hmmeng}@se.cuhk.edu.hk

Abstract

This paper presents a novel voice activity detection (VAD) method using Maximum Values of Sub-band SNR (MVSS) as the detection feature. The proposed new feature MVSS has different distributions between speech and non-speech signal, which is helpful for separating the speech signal from heavy noise. An adaptive threshold is applied to improve VAD accuracies and track the noisy signal rapidly without complex computation. Experimental results show that the proposed method achieves better performance than the conventional ETSI AMR VADs under the NOISEX-92 database.

1. Introduction

Voice activity detection (VAD) in noisy environments is an important and challenging research problem for speech processing. It is a critical process for robust speech recognition, speech coding, speech enhancement, etc. Existing VAD methods are often proposed based on general speech features, such as short-time energy, linear prediction coefficients (LPC) [1], spectral features, entropy [2], etc. These features have different VAD performance under conditions with different signal-to-noise ratios (SNR) — while they may achieve good performance in clean condition, performance degrades significantly in low SNR conditions.

Recently, alternative features such as power envelopes [3] and sub-band energy [4, 5] have been used for VAD. These features demonstrate that they can sustain performance in low SNR and require only a small number of parameters for optimization. However, they still cannot perform well for VAD if different types of noise are involved, even with the use of an adaptive threshold [6]. Consequently, several new features based on the SNR or channel SNR were proposed, such as the ETSI Adaptive Multi-Rate (AMR) VAD [7] and AFE [9]. The ETSI AMR VAD has two categories: VAD option1 uses sub-band energy as the detection feature and improves its accuracy by pitch detection technology. The VAD option2 considers channel SNR as the detection feature and also applies an adaptive background noise estimator.

In this paper, we propose a novel VAD method using maximum values of sub-band SNR (MVSS) as a detection feature. The idea behind MVSS is that high SNR points in sub-bands offer the evidence of speech. This novel method can separate the speech and non-speech signal well, especially global SNR of the utterance becomes low. Experimental results demonstrate that the proposed method is effective in speech/non-speech discrimination compared with standard works under the NOISE-92 database.

This paper is organized as follows: Section 2 presents the proposed VAD method. Section 3 describes experimentation with the proposed method and compares it with other standard VAD techniques. Section 4 summarizes our work and suggests future directions.

2. Approach for VAD

It is well known that the energy distribution of the speech signal does not spread evenly over all frequencies. For example, the energy of the voiced speech mainly concentrates in the formant regions. This means local SNR values of those formant regions become significantly higher compared to other regions. When clean speech is corrupted by the strong background noise (See Figure 1), some frequency bins of speech frame may still have high SNR values (indicated by arrows). On the other hand, when speech is absent, SNR values of all frequency bins tend to be homogeneously low. Thus it is better to use SNR-based features than energy-based features to represent the speech signal. In many cases, since the noise signal contains low-frequency components and masks the speech content, using only the global SNR may not

work very well. Therefore, we extract features based on SNR points to reduce noise.

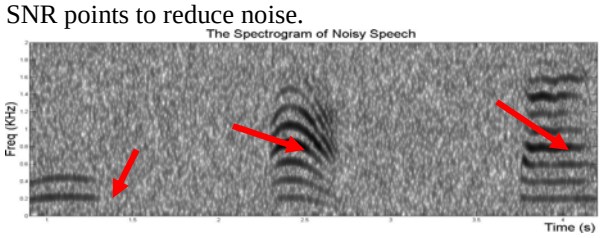

Figure 1. Short-time spectrogram of clean speech corrupted by Gaussian white noise (global SNR = 0db)

Figure 2 shows the block diagram of the proposed VAD method. Given the spectrum of a speech utterance transformed by DFT, we first divide the whole spectrum into several sub-bands. Secondly, sub-band SNRs are estimated and maximum values of sub-band SNR (MVSS) are extracted as detection features. The background noise estimation and final VAD decision are made by comparing feature value with an estimated threshold. During initialization, the estimated noise spectrum and threshold are calculated by assuming that speech always follows an initial period of noise. Details of proposed method are described in the following sections.

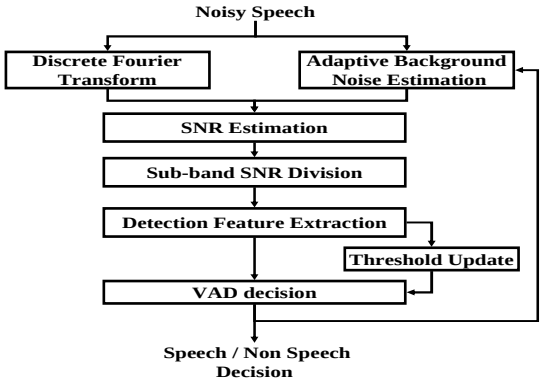

Figure 2. The block diagram of the proposed VAD algorithm

2.1 Point estimation for SNR

Given an additive noisy speech, let $Y(\lambda,k)$, $S(\lambda,k)$ and $N(\lambda,k)$ denote the k^{th} spectral component L-point FFT of noisy speech, clean speech and noise signal, respectively:

$$Y(\lambda,k) = S(\lambda,k) + N(\lambda,k) \qquad (1)$$

where λ =1,2,...,N is the frame number and k=1,2,…,L. We extract a posteriori SNR of each spectral component for the current frame λ as follows:

$$G_{SNR}(\lambda,k) = 10\log_{10}\frac{P_y(\lambda,k)}{P_n(\lambda,k)} \qquad (2)$$

where $P_y(\lambda,k)$ is the signal power spectrum of the current frame λ and $P_n(\lambda,k)$ is the noise power spectrum. However, $P_n(\lambda,k)$ could not be obtained directly. The expected value of the noise power spectrum $\bar{P}_n(\lambda,k)$ during non-speech periods is estimated instead. Therefore, the point estimate for SNR becomes:

$$\bar{G}_{SNR}(\lambda,k) = \frac{P_y(\lambda,k)}{\bar{P}_n(\lambda,k)} \qquad (3)$$

2.2 Sub-band division and SNR estimation

Since most sonorant regions of speech (e.g. vowels) contain harmonic structures, we adopted the ETSI VAD standard [7] to divide $\bar{G}_{SNR}(\lambda,k)$ points into 9 bands (between 0Hz ~ 4000Hz). Cut-off frequencies are shown in Table 1. Therefore, the set of estimated SNR values $G(\lambda)_i$ for the i^{th} sub-band in the frame λ is given by:

$$G(\lambda)_i = \{\bar{G}_{SNR}(\lambda,k) \mid k_{ibegin} \le k \le k_{iend}\} \qquad (4)$$

where k_{ibegin}, $k_{iend} \in \{1,2,\ldots,L\}$ are the cut-off indexes of each sub-band according to frequency.

Table 1. Cut-off frequencies of the 9 sub-bands in the ETSI VAD standard

Band Num.	Freq (Hz)	Band Num.	Freq (Hz)
1	0 – 250	6	1500 – 2000
2	250 – 500	7	2000 - 2500
3	500 – 750	8	2500 - 3000
4	750 – 1000	9	3000 – 4000
5	1000 – 1500		

2.3 MVSS and feature distance calculation

As we know, low-SNR-value regions most likely correspond to non-speech parts while high-SNR-value regions are obvious evidences of speech when noise is stationary. In other words, points of SNR values in harmonic regions become larger if speech is present. Therefore, it makes sense to locate the M largest values as sub-band SNR key values from each sub-band SNR set $G(\lambda)_i$. We define the maximum value of sub-band SNR $G_{\max}(\lambda)_i$ (MVSS) for the i^{th} sub-band set $G(\lambda)_i$ as follows:

$$G_{\max}(\lambda)_i = \frac{1}{M}\sum_{r=1}^{M} P_{(r)}(\lambda)_i \qquad (5)$$

where $P_{(r)}(\lambda)_i$ is the r^{th} largest value in the value set $G(\lambda)_i$ (M=6 in our experiments).

Finally we denote the feature distance gain $D(\lambda)$ as follows:

$$D(\lambda) = \sum_{i=1}^{9} G_{\max}(\lambda)_i + \sum_{i=1}^{9} (G_{\max}(\lambda)_i - \overline{G_{\max}(\lambda)})^2 \quad (6)$$

where $\overline{G_{\max}(\lambda)} = \sum_{i=1}^{9} G_{\max}(\lambda)_i / 9$.

Since the high SNR points in the value set $G(\lambda)_i$ are considered as speech information with little noise, MVSS indicates the distribution of speech given by an utterance. In other words, the distance gain $D(\lambda)$ shares the same distribution under different noise levels. Figure 3 depicts feature distance gain $D(\lambda)$ under different noise level. From the figure, the curves of normalized feature distance gains look similar under different noise levels. This verifies that the proposed MVSS feature is robust.

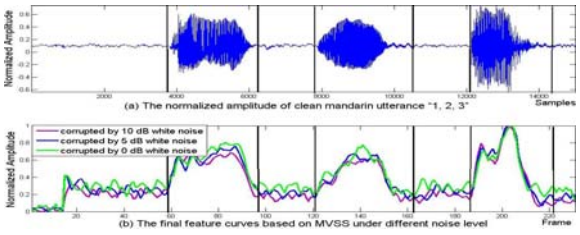

Figure 3. (a) The normalized amplitude of a clean Mandarin utterance "1, 2, 3" (b) Final feature curves based on MVSS under different noise levels (White noise at 0, 5, 10dB)

2.4 Feature interpretation

The detection feature MVSS used in this work is based on the assumption that the speech harmonic structure is robust to strong noise conditions. Since the conventional methods employ sub-band energy or global SNR as features, they may include much noise information, which degrades performance. In a low SNR environment, it is impossible to get rid of noise by extracting energy as a feature, especially when the energy of noise appears in the lower frequency band. However MVSS still works well because it reduces the effect of noise corruption by using a set of high SNR points. Moreover, important speech harmonic information is mainly kept to indicate the presence of speech. If the current frame is speech, the distance gain $D(\lambda)$ will become large; otherwise the gain value will become very small.

Experimental results demonstrate the potential advantage of using MVSS as the detection feature. Figure 4 shows the comparison of different features in low frequency band. From the figure, we see the plot for energy varies slowly near the boundary of speech and non-speech regions, which makes the speech detection difficult. On the contrary, the plot for MVSS changes significantly between speech and non-speech regions, which makes it easier for detection.

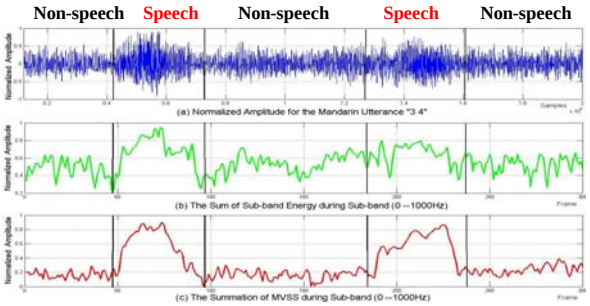

Figure 4. (a) The normalized amplitude of noisy mandarin utterance "3, 4" (b) The sum of sub-band energy (0~1000 Hz) (c) The sum of MVSS (0~1000 Hz).

2.5 Threshold calculation and VAD decision

In order to solve the case that speech is corrupted by sudden strong noise, we use previous frames information to prevent the threshold from degenerating suddenly. The final threshold is defined as follows:

$$E_{th}(\lambda) = \frac{1}{K}\sum_{j=0}^{K-1} E(\lambda - j) \quad (7)$$

$$E(\lambda) = \begin{cases} D(\lambda) & \text{if last frame is non-speech} \\ E_{th}(\lambda - 1) & \text{if last frame is speech} \end{cases} \quad (8)$$

where K is the number of frames to be averaged. A lower limit E_{th_MIN} is applied to improve the performance of the threshold. During the initialization phase, we assume that the first N (10~20) frames are always in non-speech state.

The final VAD decision is made by comparing the distance gain $D(\lambda)$ again the threshold $E_{th}(\lambda)$:

$$F(\lambda) = \begin{cases} 1 & D(\lambda) \geq E_{th}(\lambda) \\ 0 & D(\lambda) < E_{th}(\lambda) \end{cases} \quad (9)$$

The indicator function F indicates whether the current frame is a speech-like frame ($F(\lambda)$=1) or a noise-like frame ($F(\lambda)$=0).

A hangover is necessary before the final VAD decision is made [10]. Figure 5 shows the applied VAD hangover scheme. The flag $STATE(\lambda)$ indicates the final VAD decision of the current frame λ. If $STATE(\lambda)$=1, the current frame is speech; otherwise speech is absent. Suppose that the previous frame is noise ($STATE(\lambda$-1)=0), the current frame will remain in non-speech state unless there are more than m consecutive frames with

the following condition $D(\lambda) \geq E_{th}(\lambda)$ ($F(\lambda)$=1). Otherwise it will change to the speech state ($STATE(\lambda)=1$). On the other hand, suppose that the previous frame is in the speech state ($STATE(\lambda\text{-}1)$=1), the current frame will remain in speech state until there are n consecutive frames satisfying the condition $D(\lambda) < E_{th}(\lambda)$ ($F(\lambda)$=0). Otherwise it will switch to the non-speech state. Initially, we assume that it is in the noise state ($STATE(0)$=0).

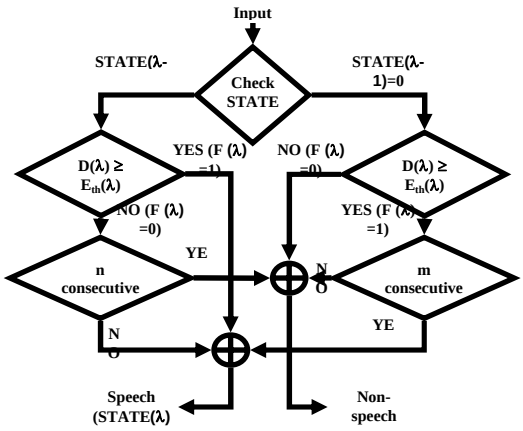

Figure 5. VAD hangover algorithm

2.6 Parameter update

In order to estimate the non-stationary noise, we continuously update the parameters according to the following equations:

$$\bar{G}(\lambda)_i = \alpha_1 G(\lambda)_i + (1-\alpha_1)\bar{G}(\lambda-1)_i \quad (10)$$

$$\bar{E}_{th}(\lambda) = \alpha_1 E_{th}(\lambda) + (1-\alpha_1)\bar{E}_{th}(\lambda-1) \quad (11)$$

$$\bar{D}(\lambda) = \alpha_1 D(\lambda) + (1-\alpha_1)\bar{D}(\lambda-1) \quad (12)$$

$$\bar{P}_y(\lambda,k) = \alpha_1 P_y(\lambda,k) + (1-\alpha_1)\bar{P}_y(\lambda-1,k) \quad (13)$$

If the update flag $STATE(\lambda)$ is set to be "1", we will update the background noise estimate in the next frame by:

$$\bar{P}_n(\lambda,k) = \alpha_2 \bar{P}_n(\lambda-1,k) + (1-\alpha_2)\bar{P}_y(\lambda-1,k) \quad (14)$$

where the parameters α_1, α_2 are fixed relevance factors for update processes in the whole system.

3. Experimental Results

The testing database used in this paper was collected from 20 individual speakers (10 male and 10 female) in our lab. Each speaker read 5 utterances of ten isolated Chinese digits "1, 2,…, 10" in a quite environment and the speech is recorded. A variety of noise from the NOISEX-92 database [11] were used, including the white noise, pink noise, volvo (car) noise and the military vehicle noise. The input signal was digitized at 8000Hz. We then extract frames of 32ms long (with 24ms overlap, or 8ms frame rate) and apply Hamming-windowed. Finally a 256-point FFT was applied. The update factors α_1, α_2 were both set to 0.95 and the number of frames for threshold estimation K is set to be 40. The lower limit of threshold E_{th_MIN} is set to be 4~7 and the number m, n are set to be 3, 8 respectively.

The proposed VAD is evaluated in terms of its ability to identify segments of speech or non-speech (background noise) at different SNR levels. Reference decisions on clean speech were made by manually labeling samples. Detection performance is assessed in terms of speech hit rate (SHR) and non-speech hit rate (NSHR), which are defined by a function of all actual speech and non-speech samples accurately detected.

Table 2. Comparison performances of the proposed VAD and standard AMR VADs (Accuracy rate %)

Noise type	SNR (dB)	AMR VAD1		AMR VAD2		MVSS	
		SHR	NSHR	SHR	NSHR	SHR	NSHR
White noise	15	**96.9**	71.3	96.4	75.7	95.6	**89.4**
	10	**96.7**	64.0	96.3	70.3	95.0	**86.0**
	5	**94.5**	50.7	95.2	62.4	90.3	**86.6**
	0	**92.4**	46.4	94.5	50.8	86.2	**84.8**
Volvo (car) noise	15	99.4	86.7	99.7	88.9	**99.6**	**90.4**
	10	99.3	82.7	99.3	85.7	**99.4**	**88.0**
	5	97.8	76.6	98.8	83.9	**99.0**	**82.9**
	0	95.5	68.5	98.2	78.6	**98.0**	**86.7**
Pink noise	15	93.7	**91.4**	93.8	92.2	**96.3**	89.5
	10	89.8	72.9	91.1	**89.2**	**94.2**	87.5
	5	88.5	69.4	80.4	87.1	**93.8**	**85.0**
	0	81.7	58.9	73.6	72.3	**89.8**	**85.6**
Military vehicle noise	15	97.5	74.2	**99.6**	**88.9**	99.5	86.8
	10	98.1	70.7	98.6	**82.8**	**99.1**	81.2
	5	91.8	63.9	93.5	**81.7**	**98.6**	77.4
	0	86.7	60.8	89.7	76.0	**94.1**	**78.5**

Table 2 shows the detection results of the novel VAD based on the MVSS feature, the standard AMR VAD1 and the AMR VAD2. The performances of different methods are compared in terms of the SHR and NSHR ranging from 15 dB to 0 dB in global SNR. We observe that AMR VAD1 performances steadily with high SHR for the whole range of SNRs. However it performs poorly in terms of NSHR when the noise level increases. VAD2 demonstrates considerable improvements over VAD1 with better NSHR. Unfortunately, it still suffers

fast degradation in the speech detection performance under certain unfavorable noisy conditions (e.g. pink noise). The proposed method can handle the above problems robustly even when SNR decreases and achieves best performance on average in term of SHR and NSHR.

4. Conclusions

In this paper, a new VAD method is proposed based on maximum values of Sub-band SNR (MVSS). In addition, an adaptive threshold updating scheme is used in our method. The experimental results prove that the proposed method achieves better performance compared with AMR VADs under the low SNR condition. Finally the proposed method is suitable for real-time application since it its computational complexity is relatively low.

Acknowledgment

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. The author also would like to thanks Dr. Lu Wei for the useful talks and suggestions on the research topic.

References

[1] L.R. Rabiner and M.R. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem*," IEEE Trans. Acous., Speech, Signal Proc.*, 1977.

[2] B.F.Wu and K.C.Wang, "Noise spectrum estimation with entropy-based VAD in non-stationary environments," *IEICE Trans. Fund. Elec. Comm. & Comp. Sci.,* 2006.

[3] M.Mark and K.Birger, "Speech pause detection for noise spectrum estimation by trackingpower envelope dynamics," *IEEE Trans. Speech Audio Proc.,* vol.10, pp.109-118, 2002.

[4] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

[5] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. Speech Audio Proc.*, 2002.

[6] J. Ramirez, J. C. Segura, M. C. Benitez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. of EUROSPEECH 2003*

[7] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[8] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm, compression algorithms," *ETSI ES 200 050 Recommendation*, 2002.

[9] A.Davis, C.Nordholm, and T. Roberts, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech and Langrage Proc.,* vol. 14, no. 2, pp. 412–424, March 2006.

[10] A.P. Varga, H.J.M Steeneken, M. Tomlinson, D.Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition", In Technical Report, *DRA Speech Research Unit*, 1992.