2013 6th International Congress on Image and Signal Processing (CISP 2013)

A Voice Activity Detection Algorithm With Sub-band Detection Based on Time-frequency Characteristics of Mandarin

Yinfeng Wang
School of Information Science and Engineering
Shandong University
Jinan, 250199, China

Shaoguang Huang
School of Information Science and Engineering
Shandong University
Jinan, 250199, China

Ying Wei
School of Information Science and Engineering
Shandong University
Jinan, 250199, China

***Abstract*—Voice activity detection algorithms are widely used in the areas of voice compression, speech synthesis, speech recognition, speech enhancement, and etc. In this paper, an efficient voice activity detection algorithm with sub-band detection based on time-frequency characteristics of mandarin is proposed. The proposed sub-band detection consists of two parts: crosswise detection and lengthwise detection. Energy detection and pitch detection are in the range of considerations. For a better performance, double-threshold criterion is used to reduce the misjudgment rate of the detection. Performance evaluation is based on six noise environments with different SNRs. Experiment results indicate that the proposed algorithm can detect the area of voice effectively in non-stationary environment and low SNR environment and has the potential to progress.**

Keywords- VAD, mandarin, sub-band detection, pitch

I. INTRODUCTION

Speech is one of the most important means in human's communication. At present, the digital speech processing has been widely used in many areas, such as voice compression, speech synthesis, speech recognition, and speech enhancement. When speech signals are transmitted or processed, noise is brought in inevitably. The quality and intelligibility of speech in recordings are degraded because of the existence of the background noise. Voice activity detection (VAD) algorithm is aimed at detecting the voice area from the noisy signals [1] [2]. Much effort has been invested into the research of VAD algorithms. There are mainly two types of algorithms: one is based on the time domain parameters of speech signals, such as short-time energy [3], zero-crossing rate [4] and duration parameters [5]. The other is based on frequency domain analysis of speech signals, such as entropy [6], the frequency band variance [7] and wavelet analysis [8]. Later, some scholars put forward the fusion VAD algorithms with several characteristics of speech signals [9], and the judgment of speech endpoint has developed from one single threshold to multiple thresholds and adaptive thresholds. Furthermore, linear prediction coefficients [10], cepstral coefficients [11], pitch [12] and neural network methods [13] have been applied to the VAD algorithm, which enriches the research of the VAD algorithms.

There are two main shortcomings in the currently common used algorithms. First, most of the current algorithms have good performance for high signal-to-noise ratio (SNR) and stationary noise environment, but do not perform so well in low SNR and non-stationary noise environment. Secondly, most of the noise reduction algorithms are based on English environment which cannot take advantage of the characteristics of mandarin. To solve these problems, in this paper, an algorithm aimed at a better performance in low SNR and non-stationary noise environment and based on mandarin is represented. The energy detection and the time-frequency characteristics on vowels and consonants of mandarin pronunciation, such as the pitch and the time length of the word, are involved in the detecting of the voice activity.

The paper is organized as following. In Section II the proposed algorithm, including the fundamentals and the details of the VAD algorithm is described. In Section III the simulation results of the proposed algorithm in different noise environments and different signal-to-noise rates are presented. Conclusion is drawn in Section IV.

II. THE PROPOSED ALGORITHM

A. Fundamentals

There are several important features of mandarin [14] [15]. Mandarin consists of consonant and vowel. Spectrogram of vowel consists of a fundamental frequency F0, called pitch, and its harmonics. Spectrogram of consonant is almost the same with the white noise. Both the vowel and the constant start or end almost at the same time. Furthermore, one mandarin word lasts about 300ms or more. These features are illustrated in Figure 1. Indication i1 is the area of consonant and its corresponding spectrum area is i2, which reflects the feature of consonant. Indication i3 and i4 reflect the feature of vowel in pitch.

This work is supported by the National Natural Science Foundation of China (Grant No. 61201372), Jinan Youth Star of Science and Technology Plan (Grant No. 20120112) and Shandong province science and technology development plan (Grant No. 2013GGX10103).
E-mail: eleweiy@sdu.edu.cn.

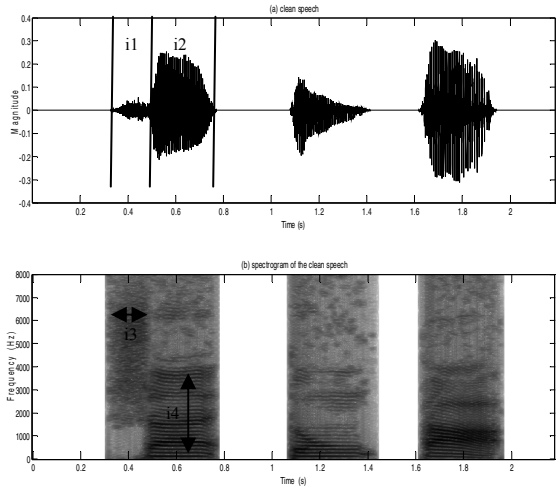

Figure 1. Spectrogram of the clean speech. (a) shows the waveform of the pure speech for three single mandarin words (shan dong da, in this case). (b) shows the corresponding spectrogram of the pure words.

B. *The proposed VAD algorithm*

The block diagram of the VAD algorithm is shown in Figure 2. The proposed voice activity detection algorithm consists of crosswise detection and lengthwise detection. After the acoustic filter bank, the noisy signals are divided into many channels. Crosswise detection determines whether the voice appears by detecting the variation of energy in one band. Lengthwise detection determines whether the voice appears by detecting the energy variation of one frame in all bands. Some other influencing factors are considered in the follow-up process, such as the pitch and corresponding harmonics, length of word, and etc.

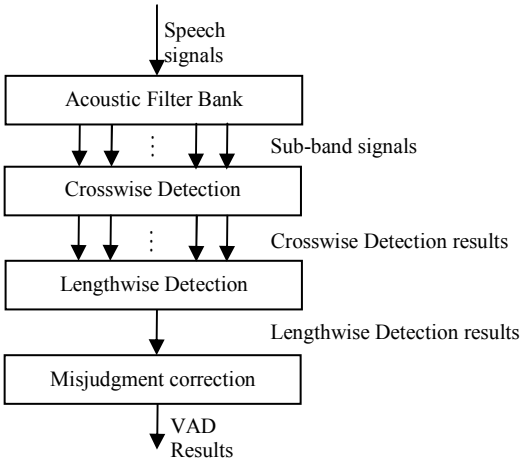

Figure 2. The block diagram of the VAD algorithm.

In this paper, the sampling rate of signals is 16 *KHz* and the word length is 16 bit. The signals first go though the acoustic filter bank, and are divided into 19 channels. The acoustic filter bank is 1/3 octave filter bank up to ANSI *S11.1* standard [16], and F21 to F39 channels are chosen to get the 19 channels signals. An example of the output of the filterbank is shown in Figure 3. The vowel signals are mainly concentrated in low frequency areas and the pitch appears in sub-band F22 in this case; the consonant signals are detected in high frequency areas in sub-bands F36 and F38.

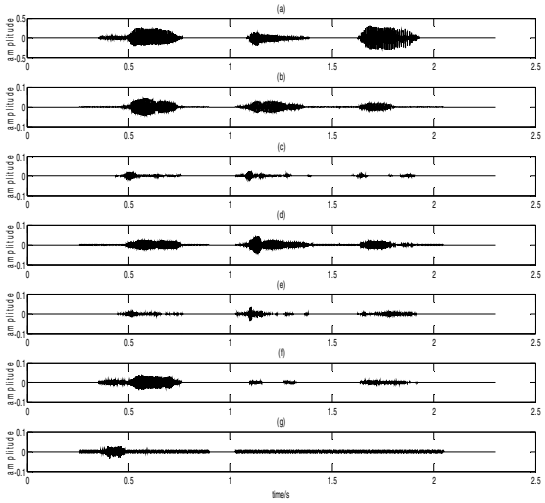

Figure 3. Simulation results of signal through the acoustic filter bank. (a) shows the waveform of the original speech signals. (b)-(g) show the decomposed input signals of sub-band F22, F23, F25, F26, F36 and F38.

1) *The Acoustic Filter Bank*

According to the ANSI *S11.1*: specification for octave, half-octave and third octave band filter sets [17], nominal midband frequencies for one-third-octave-band and octave-band filters are given in table Ⅰ.

TABLE Ⅰ. NOMINAL MIDBAND FREQUENCIES FOR ONE-THIRD-OCTAVE-BAND AND OCTAVE-BAND FILTERS IN THE AUDIO RANGE

Band number	**nominal midband frequency (Hz)**	**Band number**	**nominal midband frequency (Hz)**
19	80	30	1000
20	100	31	1250
21	125	32	1600
22	160	33	2000
23	200	34	2500
24	250	35	3150
25	315	36	4000
26	400	37	5000
27	500	38	6300
28	630	39	8000
29	800	40	10000

It can be seen that the nominal midband frequency of band *i* is about twice bigger than that of the band *i-3*. This characteristic is used to detect the pitch in the lengthwise detection.

2) *Crosswise Detection*

Crosswise detection is used to detect the speech area roughly by comparing the energy of the current frame with that of the estimated noise in one band, as shown in (1).

$$\mathrm{E}_{frame}(i,j)=\sum_{k=1}^{len_{frame}} sample(k,i,j)^2$$

$$\mathrm{E}_{noise}(j)=\frac{1}{m}\sum_{k=1}^{m} E_{frame}(i,j) \tag{1}$$

$$VAD_{crosswise}(i,j)=\begin{cases}1 & ,E_{frame}(i,j)>r_2*E_{noise}(j)\\ VAD_{crosswise}(i-1,j), & r_1*E_{noise}(j)\le E_{frame}(i,j)\le r_2*E_{noise}(j)\\ 0 & ,E_{frame}(i,j)<r_1*E_{noise}(j)\end{cases}$$

where, $sample(k,i,j)$ is the magnitude of sample k in frame i of the j-th band. $\mathrm{E}_{frame}(i,j)$ is the energy of this frame, where len_{frame} is the length of the frame. The mean energy of the starting m frames is used as the noise threshold of this band, denoted as $\mathrm{E}_{noise}(j)$.

Usually, when the energy of the current frame is bigger than the noise threshold, the current frame signal is regarded as a speech signal, otherwise noise signal. However, there is a middle fuzzy region where the misjudgments likely happen. To solve this problem, double-threshold criterion is used. In the j-th band, when the energy of the i-th frame is bigger than the absolute threshold $r_2*\mathrm{E}_{noise}(j)$, let the crosswise detection result be 1. When it is smaller than the protecting threshold $r_1*\mathrm{E}_{noise}(j)$, let the result be 0 ($r_2>r_1$). Or else, keep the result the same as the previous VAD result.

3) Lengthwise Detection

With the time characteristics of the vowels and consonants, lengthwise detection determines whether there is an emergence of the voice using the statistics results of the crosswise detection. The original VAD result of the i-th frame in lengthwise detection is shown in (2).

$$VAD_{lengthwise}(i)=\sum_{j=1}^{n} VAD_{crosswise}(i,j) \tag{2}$$

where, n is the number of bands, and equals 19 in this paper. Taking the time-frequency characteristics of mandarin into account, for bands F21 to F23, pitch detection is added to get a better performance, as shown in (3).

$$VAD_{lengthwise}(i)=\begin{cases}VAD_{lengthwise}(i)+ad1, & VAD_{crosswise}(i,j)=1\\ & \cap VAD_{crosswise}(i,j+3)=1\cap VAD_{crosswise}(i,j+6)=1\\ VAD_{lengthwise}(i)+ad2, & VAD_{crosswise}(i,j)=1\\ & \cap(VAD_{crosswise}(i,j+3)=1\oplus VAD_{crosswise}(i,j+6)=1)\end{cases} \tag{3}$$

where, *ad1* and *ad2* are the addends to indicate the impact of the pitch. According to the frequency distribution of the acoustic filter bank, for F21 to F23 bands, when the energy of signals in the j-th band and the $(j+3)$-th band is detected larger than the noise estimation, pitch appears. According to the possibility of the pitch appearance, let *ad1* be larger than *ad2*.

After the updating of $VAD_{lengthwise}(i)$, the VAD result of the i-th frame $VAD_{temp}(i)$ is calculated by (4),

$$VAD_{temp}(i)=\begin{cases}1 & ,VAD_{lengthwise}(i)>Thr1\\ VAD_{temp}(i) & ,Thr2\le VAD_{lengthwise}(i)\le Thr1\\ 0 & ,VAD_{lengthwise}(i)<Thr2\end{cases} \tag{4}$$

Also, double-threshold is used. Where *Thr1* is the absolute threshold and *Thr2* is the protecting threshold, which is smaller than *Thr1*.

4) Misjudgment correction

Misjudgment correction must be taken after sub-band detection to keep the speech continual and get a more accurate result. Two steps are done in this part. One is noise misjudgment correction in speech area and the other is speech misjudgment correction in noise area. To correct the noise misjudgment of the VAD result in speech area, if the number of continual zeros between the adjacent ones is less than r_3, 1 will be set to take the place of 0. To correct the speech misjudgment in noise area, the length of a word must be taken into account. If the number of continual ones is less than r_4, 0 will be set to take the place of 1. By testing, the VAD result will be good when r_3 ranges from 1 to 4 and r_4 ranges from 30 to 60.

Moreover, in order to adapt to the changing environment, noise estimation parameter need to be updated as (5):

$$\mathrm{e}_{noise}(i,j)=\begin{cases}\mathrm{e}_{noise}(i-1,j) & ,\ VAD_{temp}(i-1)=1\\ \mathrm{e}_{noise}(i-2,j)*(1-r_5)+\mathrm{e}_{noise}(i-1,j)*r_5 & ,\ VAD_{temp}(i-1)=0\end{cases} \tag{5}$$

where, $\mathrm{e}_{noise}(i,j)$ is the power estimation of noise for the i-th frame of the j-th band. While $i<m$, $\mathrm{e}_{noise}(i,j)$ equals to $\mathrm{E}_{noise}(j)$, otherwise $\mathrm{e}_{noise}(i,j)$ takes the place of $\mathrm{E}_{noise}(j)$ and is used in crosswise detection. The regulation parameter r_5 decides the update rate of noise estimation. In order to reduce the computation burden, the update rate is set as per 50 frames (0.2s).

III. Simulation Results

For mandarin speech database, 42 single words and 10 simple sentences are used to test the performance of the proposed algorithm. There are six types of background noise: white noise, pink noise, factory noise, f16 noise, destroyer noise and jet noise, which are provided by NOISEX-92 [18].

Figure 4 shows the simulation results of noisy signal through the acoustic filter bank, with the white background noise and the SNR is 3db. Compared with the waveforms in figure 3, the appearance of speech in low frequency bands is easily detected than that in high frequency bands, which fits to the characteristics of the consonant and vowel in frequency domain. The existence of noise can hardly influence the vowel detection, but not so ideal when it comes to the consonant.

The performance of the proposed VAD algorithm is evaluated by the accuracy of VAD with different SNR (-3db, 0db, 3db, 5db, 7db, 10db). To get the accuracy of VAD, manual classification on the pure speech is used as the standard result, and the accuracy of VAD is shown in (6) to (8).

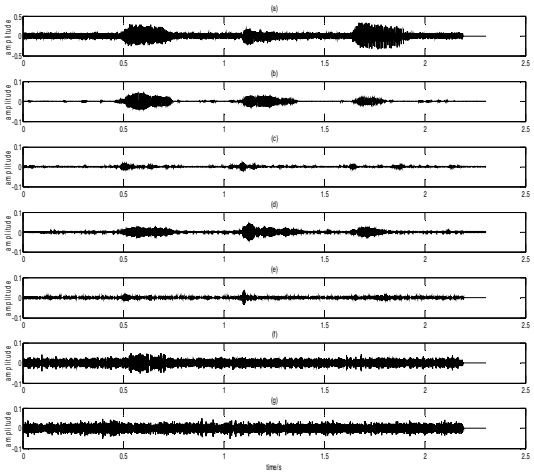

Figure 4. Simulation results of noise signal through the acoustic filter banr, with white bacrground noise and the SNR is 3db. (a) shows the waveform of the original noisy speech signals. (b)-(g) show the decomposed input signals of sub-band F22, F23, F25, F26, F36 and F38.

$$Accuracy\ of\ VAD = \frac{the\ number\ of\ correctly\ judged\ frames}{the\ number\ of\ frames} \quad (6)$$

$$Accuracy\ of\ VAD\ for\ speech\ signals = \frac{the\ number\ of\ correctly\ judged\ speech\ frames}{the\ number\ of\ speech\ frames} \quad (7)$$

$$Accuracy\ of\ VAD\ for\ noise\ signals = \frac{the\ number\ of\ correctly\ judged\ noise\ frames}{the\ number\ of\ noise\ frames} \quad (8)$$

The experiment parameters are set according to the characteristics of mandarin. The length of signal frame is 64, which means 4*ms* per frame. The noise power estimation parameter *m* is set to 10. Set r_1 to 1.2 and r_2 to 1.6. Set *ad1* to 4 and *ad2* to2. Set *Thr1* to 5 and set *Thr2* to 6. Set r_3 to 2 and r_4 to 40. Set r_5 to 0.1. The final simulation results of the single words and sentences with the proposed VAD algorithm are shown in Table II-IV.

TABLE II. THE ACCURACY OF THE PROPOSED VAD ALGORITHM FOR WORDS AND SENTENCES TABLE TYPE STYLES

SNR (db)	Accuracy of VAD (%)					
	White	Pink	Factory	F16	Destroyer	Jet
-3	91.68	89.56	79.91	75.27	87.47	76.43
0	93.35	91.84	84.38	79.22	90.00	83.86
3	94.15	93.70	86.06	81.66	91.73	86.98
5	95.56	94.59	87.89	82.98	92.32	88.71
7	96.36	95.62	88.79	83.53	93.09	91.57
10	96.58	95.42	90.30	85.05	93.22	93.51

TABLE III. THE ACCURACY OF VAD FOR SPEECH SIGNALS FOR WORDS AND SENTENCES

SNR (db)	The accuracy of VAD for speech signals (%)					
	White	Pink	Factory	F16	Destroyer	Jet
-3	97.39	96.54	84.53	66.58	88.84	99.59
0	96.41	95.66	84.86	66.50	88.04	99.59
3	96.49	95.01	83.16	66.21	87.01	99.59
5	96.18	94.63	82.02	65.80	87.03	99.38
7	96.00	92.74	82.02	64.97	86.46	99.35
10	95.30	90.24	81.56	63.99	84.84	99.30

TABLE IV. THE ACCURACY OF VAD FOR NOISE SIGNALS FOR WORDS AND SENTENCES

SNR (db)	The accuracy of VAD for noise signals (%)					
	White	Pink	Factory	F16	Destroyer	Jet
-3	88.06	85.20	76.99	80.76	86.60	61.80
0	91.40	89.43	84.07	87.26	91.24	73.93
3	92.66	92.87	87.89	91.42	94.71	79.02
5	95.17	94.57	91.60	93.83	95.66	81.97
7	96.59	97.44	93.08	95.27	97.26	86.65
10	97.39	98.69	95.84	98.34	98.52	89.85

As the test data shown in the tables, in white, pink and destroyer noise environments, the proposed VAD algorithm has good performance and the accuracy of VAD is above 90% even when the SNR is low. In factory, F16 and jet noise environments, the accuracy of VAD is above 80%. Compared with the common VAD algorithms, this result is satisfactory.

The accuracy of VAD for speech signals becomes smaller when the SNR becomes bigger, which means the proposed VAD algorithm performs better in low SNR environments. That's because the noise estimation in crosswise detection is not accurate enough when SNR is high, but much better when SNR is low. This protects the speech area in the largest extent in low SNR environments.

IV. CONCLUSION

An effective VAD algorithm with sub-band detection based on time-frequency characteristic of mandarin was proposed. After going through the acoustic filter bank up to ANSI *S11.1* standard, the signals are detected by crosswise detection and lengthwise detection in turn. Along with the whole process, misjudgment correction and noise updating are done. Experiments showed that the proposed algorithm could detect the speech area in low SNR environments well. The accuracy data provides a direction of optimization and the adjustment of the parameters could lead to a better performance.

ACKNOWLEDGMENT

Thanks to Jingru Huang for her assistance in research methods. Thanks to Xiaomei Zhang and Min Cao for the help in information collection.

REFERENCES

[1] Sangwan. A, M.C. Chiranth, R. Shah, V. Gaurav and R.V. Prasad, "Comparison voice activity detection algorithms for VOIP," Computers and Communications, 2002, pp. 530-535.

[2] B.F. Wu and K.C. Wang, "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments," IEEE Transactions on Speech and Audio Processing, 2005, pp. 762-775.

[3] L. Lamel, L. Labiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detect for isolated word recognition," Acoustics, Speech and Signal Processing, IEEE Transactions on 29.4, 1981, vol. 29,no. 4, pp. 777–785.

[4] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," in: *Speech Commun.*, 1989,vol. 8, pp. 45-60.

[5] H. Ney, "An optimization algorithm for determining the endpoints of isolated utterances," in: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'81, 1981, vol. 6, pp. 720–723.

[6] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," ICSLP, 1998, vol. 98, pp. 232-235.

[7] W.W. Hung and H.C. Wang, "On the use of weighted filter bank analysis for the derivation of robust MFCC," IEEE Signal Processing Letters, 2001, pp. 70-73.

[8] J. F. Wang and S. H. Chen, "A voice activity detection algorithm based on perceptual wavelet packet transform and teager energy operator," International Symposium on Chinese Spoken Language Processing, 2002, pp. 177–180.

[9] J.H. Chang, N.S. Kim and S.K. Mitra, "Voice activity detection based on multiple statistical models," IEEE Trans Signal Processing, 2006, pp. 1965-1976.

[10] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'77, vol. 2, 1977, pp. 323-326.

[11] J. A. Haign and J. S. Mason, "Robust voice activity detection using cepstral features," TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering, 1993, pp. 321–324.

[12] Morales-Cordovilla, J.A. , Ning Ma , Sanchez, V. , Carmona, J.L. , Peinado, A.M. , Barker, J., "A pitch based noise estimation technique for robust speech recognition with missing data," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 4808–4811.

[13] Kos M., "Noise reduction algorithm for robust speech recognition using minimum statistics method and neural network VAD," Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on. IEEE, 2007 , pp. 284-287.

[14] D. F. Rosenthal and H. G. Okuno, "Computational auditory scene analysis," Psychology Press, 1998.

[15] D. Wang and G. J. Brown, Eds., "Computational auditory scene analysis: principles, algorithms and applications," Wiley interscience, 2006, vol. 147.

[16] Y. T. Kuo, T. J. Lin, Y. T. Li, and C. W. Liu, "Design & implementation of low-power ANSI S1.11 filter bank for digital hearing aids," Circuits and Systems I: Regular Papers, IEEE Transactions, 2010, vol. 57, no.7, pp. 1684-1696.

[17] G. S. K. Wong, T. F. W. Embleton, S.B. blaeser, "ANSI *S11.1*: specification for octave, half-octave and third octave band filter sets", Ameriacan National Standard, 2004.

[18] A. Varga, H. J. M. Steenneken, M. Tomlinson, D. Jones. 1992. NOISEX-92. [Online]. Available: http://spib.rice.edu/spib/select_noise.html.