Improved Voice Activity Detection Based on Iterative Spectral Subtraction and Double Thresholds for CVR

Xiangbin Li[1], Guo Li[1], Xueren Li[2]
1 AAC Department. Engineering College,
Air Force Engineering University, Xi' an, Shaanxi, 710038, China
2 SR Department. Air Force Engineering University, Xi' an, Shaanxi, 710051, China
xianlxb@163.com

Abstract

Cockpit voice recorder (CVR) in aircraft black box records many cockpit voices, such as speaker voices, noises and background sounds with special meanings. Cockpit voices' complexity exacerbates analysis difficulty through traditional differentiating and hearing methods, so that fresh cockpit voices are not captured easily from non-stationary sounds. In this paper, by analyzing firstly thoroughly characteristics of cockpit voices, we develop an improved voice activity detection scheme based on iterative spectral subtraction and double thresholds Finally, to demonstrate the effectiveness of the proposed scheme, we make simulations with a section of speech (SNR=8) from standard voice bank and a section of true cockpit voices and compare the probabilities P_{cs} and P_{cn} of the three algorithms, where P_{cs} denotes probability of correctly detecting speech frames P_{cn} probability of correctly detecting noise frames. Simulations results are presented to demonstrate the effectiveness of the improved algorithm.

1. Introduction

Cockpit voice recorder (CVR) in aircraft black box records many cockpit voices, such as speaker voices, noises and background sounds with special meanings. They are complex non-stationary signals with characteristics of mutation, instantaneousness and singularity. In some special cases, speaker voices combined in background sounds play an important role in air accident investigation (AAI). However, when the aircraft is flying, especially when an accident is happening, the record condition of CVR is becoming worse. Therefore, fast extraction of speaker voices from cockpit voices is an important work in AAI. The conventional measure of AAI is based on "differentiating and hearing" method and simple audio processing, so more exact speaker voices can't be obtained easily.

Voice activity detection (VAD) is just detecting the beginning and ending of a section of speech signal, and achieves the goal of distinguishing speaker voice from background sounds. VAD in AAI requires proper detection and low computing cost with the purpose of gaining more time for AAI. The traditional double thresholds VAD based on short time energy and zero crossing rates acquires wonderful performance results in case of high signal to noise ratio(SNR), but has total failure when speaker voices are submerged in strong background sounds or low SNR, for example, when the accident happens.

In this paper, we develop an improved scheme based on double thresholds VAD with spectral subtraction. The paper is organized as follows: Section 2 analyzes characteristics of cockpit voices. Section 3 and 4 describe respectively the scheme of traditional VAD based on double thresholds and basic spectral subtraction. In section 5 and 6 present the improved scheme, and simulation results are also given in this section. Finally, some conclusions are drawn in section 7.

2. Voice characteristics of CVR

The frequency scope of cockpit voice recorded by CVR is very wide, about 150Hz to 6800 Hz, which brings some difficulty to sound separation. With the aim of facilitating sound separation, the cockpit voice is classified to three kinds: aviation noises, speaker voices and background sounds [1][2].

Aviation noises include additive noises and non-additive noises. Additive noises contain periodic noises, impulse, broad band noises, and speech interference. Non-additive noises are mainly sound residue and circuit noises. Non-additive noises can be transformed to additive ones by means of a particular transformation. However, more specifically, aviation noises include engine sound, exterior air current noise while flying, skating noise while takeoff and landing, circuitry noise in electrical equipments and circuits, motor noise droved by power when manipulating aircraft and so on.

978-0-7695-3342-1/08 $25.00 © 2008 IEEE
DOI 10.1109/PEITS.2008.84

Table 1 Characteristics of sounds of CVR

Sound	Characteristics in time domain	Characteristics in frequency domain	Short time energy	Zero crossing rate
surd	less obvious	Obvious, energy mainly locates in high frequency band	weak	high
sonant	obvious and periodic	has formant, energy mainly locates in low frequency band	strong	low
silence	less obvious	less obvious	weak	high
impulse noise	transient		strong	high

Background sounds mainly consist of various sounds except cockpit voices and aviation noises. Different background sound implies that special event has happened [3].Cockpit voices involve conversations between pilot and co-flyers, communication from control tower and speech for navigation and identification. Voice signals are a time-varying and non-stationary random process, but its characteristic keep unchangeable in a short time 10-30 ms because of relatively stability of vocal cords sound channel. Chinese language includes surd and sonant. Table 1 shows some characteristics of sounds of CVR.

3. Basic VAD algorithm based on double thresholds

3.1. Basic conception

(1) Short Time Energy (STE): The sound intensity of a speech series $x(n)$ is described by short time energy, which is defined as follows:

$$E_n = \sum_{m=-\infty}^{m} \left[x(m)w(n-m)\right]^2 = \sum_{m=n}^{n+N-1} \left[x(m)w(n-m)\right]^2 \quad (1)$$

Especially, STE of surd is very weak and STE of sonant is quite strong.

(2) Zero Crossing Rate (ZCR): The ZCR of a speech series $x(n)$ is defined as follows:

$$z_n = \sum_{m=-\infty}^{m} \left|\mathrm{sgn}\left[x(n)\right] - \mathrm{sgn}\left[x(n-1)\right]\right| w(n-m) = \left|\mathrm{sgn}\left[x(n)\right] - \mathrm{sgn}\left[x(n-1)\right]\right| w(n) \quad (2)$$

where $\mathrm{sgn}[x]$ is a sign function and $w(n)$ is a window function, which are defined as follows:

$$\mathrm{sgn}[x] = \begin{cases} 1 & (x \geq 0) \\ -1 & (x < 0) \end{cases} \quad (3)$$

$$w(n) = \begin{cases} 1/2N & (0 \leq n \leq N-1) \\ 0 & (\mathrm{others}) \end{cases} \quad (4)$$

3.2. VAD based on double thresholds

Generally, we use ZCR to detect sonant and STE to surd in practical applications [4]. The whole VAD process is divided to four sections: silence section (status=0), transition section (status=1), speech section (status=2) and end section. At the beginning of VAD, we set two thresholds for STE and ZCR each other, for example, high threshold T_{amp1} and T_{zcr1}, low threshold T_{amp2} and T_{zcr2}. Besides, we define a variable *count* as a speech counter, *silence* as silence counter, *minlen* as a minimum time threshold. Figure 1 shows flow of VAD based on double thresholds.

Many practices prove that this method can separate speeches from background noises effectively and efficiently in high SNR according to table 1. However, the aviation condition with low SNR and awful environment for record causes the method loses its own performance, because the speeches are submerged in strong aviation background noises. Therefore, former noise reduction and speech enhancement are becoming extremely important.

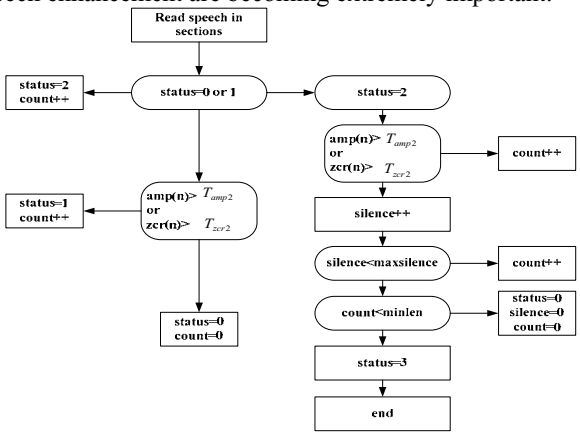

Figure 1 Flow chart of VAD based on double thresholds

4. Scheme of basic spectral subtraction

The basic spectral subtraction (SS) method is described briefly in this section. Assume that a noisy speech signal is expressed as

$$y(i) = s(i) + d(i), \quad 0 \leq i \leq N-1 \quad (5)$$

where $s(i)$ and $d(i)$ are a frame of clean speech and noise, respectively. Considering human's ear be no-sensitive to phase distortion, phase signal of noise is implemented when phase is restored [5][6]. In the frequency domain, equation (5) is expressed as

$$|Y_k|^2 = |S_k|^2 + |N_k|^2 + S_k N_k^* + S_k^* N_k \quad (6)$$

where Y_k , S_k and N_k are discrete-time Fourier transforms (DFT) of $y(i)$, $s(i)$ and $d(i)$, respectively. Because $s(i)$ and $d(i)$ are independent and N_k is gauss distribution, equation (6) is expressed as (7) in frequency domain.

$$\left|\hat{S}_k\right|^2 = \left[|Y_k|^2 - E\left[|N_k|^2\right]\right]^{1/2} = \left[|Y_k|^2 - \lambda_n(k)\right]^{1/2} \quad (7)$$

For a frame of speech signal, we have $|Y_k|^2 = |S_k|^2 + \lambda_n(k)$, so estimate of original speech is expressed as

$$\left|\hat{S}_k\right|^2 = \left[|Y_k|^2 - E\left[|N_k|^2\right]\right]^{1/2} = \left[|Y_k|^2 - \lambda_n(k)\right]^{1/2} \quad (8)$$

where $\lambda_n(k)$ is statistical mean of unvoiced speech, $\left|\hat{S}_k\right|^2$ is amplitude of enhanced speech.

However, basic SS can generate much musical noises in residual noises. Some modified SS are proposed to reduce effect. Weighting factor α and power coefficient β are introduced into SS, so equation (8) is modified as

$$\left|\hat{S}_k\right| = \left[|Y_k|^{\alpha} - \beta\lambda_n^{\alpha}\left[|N_k|^2\right]\right]^{1/\alpha} \quad (9)$$

Modified SS is degraded to basic SS when α =2 and β =1[5]. Other modified SS is showed in relative references [6][7]. Better enhancement performance can be gained by adjusting two parameters suitably, but voice distortion becomes severer as the degree of noise reduction is larger.

5. Proposed VAD based on improved spectral subtraction

5.1. Improved spectral subtraction

In this paper, we propose iterative spectral subtraction to formerly reducing noise and enhancing speech. This method uses basic SS or modified SS for appropriate times. The former enhanced speech becomes latter input signal, so music noise is seen as input noise to reduce again.

5.2. Proposed VAD based on improved spectral subtraction

On the basis of section above, we can firstly apply spectral subtraction for noisy sound of CVR to reducing noise and enhance speech, and then enhanced signal is filtered by a preceding filter, finally cockpit voice is extracted by means of double thresholds VAD. Figure 2 shows the flow chart of proposed VAD. The preceding filter is a high-pass filter, such as $1-0.9375z^{-1}$, which can filter low-frequency interference, especially interference of frequency 50Hz or 60Hz, and advance spectrum of high frequency which is useful for cockpit voice.

6. Experiment and simulation

6.1. Evaluation standard of VAD

For a wonderful VAD, two requirements must be taken into considered comprehensively: to detect more speech sections and more unvoiced speech sections. However, when VAD tries to detect more speech frames, it misjudges silence as speech or otherwise. The latter is ever worse than the former for accident investigation. Therefore, two evaluation standards are compared to weighing quantificationally the performance of VAD: probability of correctly detecting speech frame P_{cs} and probability of correctly detecting noise frame P_{cn}, which are expressed as

$$P_{cs} = \frac{N_1}{N_1^{hand}}, \quad P_{cn} = \frac{N_0}{N_0^{hand}} \quad (10)$$

where N_1^{hand} and N_0^{hand} are relatively the overall number of hand-labeling speech frames and noise frames by hand-labeling, N_1 and N_0 are relatively number of being detected correctly by VAD.

6.2. Experiment results

In this paper, a section of speech in car (SNR =8) from standard voice bank Aurora2 and a section of true cockpit sound are used, simulation experiments based on traditional double thresholds VAD only and the proposed VAD are carried out. Figure 3 and table 2 compare the performance of various methods in different environment. Due to former spectral subtraction, the SNR increases, the curves of STE and ZCR become smoother and proper probability increases.

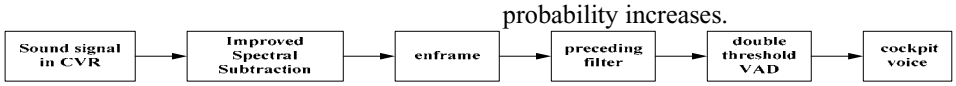

Figure 2 Flow chart of proposed VAD

Table 2 Experiment results

Method / Evaluation standard / Test samples	Traditional double thresholds VAD			Proposed VAD (Iterative number=1)			Proposed VAD (Iterative number=2)		
	SNR of original signal	P_{cs} (%)	P_{cn} (%)	SNR of enhanced signal	P_{cs} (%)	P_{cn} (%)	SNR of enhanced signal	P_{cs} (%)	P_{cn} (%)
In car(SNR=8)	8	85.4	86.2	10	91.2	93.1	14	95.2	96.5
True cockpit sound(SNR=1)	1	65.4	66.6	9	85.4	86.6	11	92.1	93.3

7. Conclusion

According to the characteristics of sound signal recorded in CVR, the objective of this paper proposes iterative spectral subtraction to improve the performance of traditional double thresholds VAD. Basic spectral subtraction has its own flaws: Because the current noise frame is replaced by the statistical mean of the whole noise, there are much musical noises in residual noises. To reduce the effect, we introduce iterative spectral subtraction. It can reduce efficiently and effectively music signal by adjusting appropriately the iterative number. The proposed algorithm can increase the SNR and reduce computational complexity for runtime requirement. Experiment results demonstrate that spectral subtraction which is used to noise reduction and speech enhancement to enhance SNR can provide feasible condition for traditional double thresholds VAD. Proposed VAD can also gain wonderful performance even if in the low SNR.

Reference

[1] James R. Cash. Group Chairman's Report of Investigation Sound Spectrum Study of Cockpit Voice Recorder, American Airlines Flight 587, DCA02MA001, Belle Harbor, NY, Nov. 12 ,2002.

[2] Ronald Stearman. Signal Analysis of Cockpit Voice Recorder Data, Report No. ASE46Q-FP1 [R]. The University of Texas, 2003.August. 16.

[3] YANG Lin, Cockpit Voice Recorder (CVR) and Laboratory Processing Methods [J]. China Civil Aviation, 2003, 29(12):21-22.

[4] An Improved Voice Activity Detection Using Higher Order Statistics [J]. IEEE Trans. on Acoustic, Speech, Signal Process, VOL.13.NO.5, SEPTEMBER 2005:965-974

[5] Sim B L, Tong Y C, Chang J S, and Tan C T. A parametric formulation of the generalized spectral subtraction method. IEEE Trans. on. Speech and Audio Processing, 1993, 6(4):325-337

[6] Boll S F. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on Acoustic, Speech, Signal Process, 1979, ASP-27(2):113-120

[7] Y. Ephraim and D.Malah, "Speech enhancement using a minimum mean-square log-spectral amplitude estimator," IEEE Trans. Acoustic, Speech, Signal Processing, vol. 33, no. 2, pp 443–445, 1985.

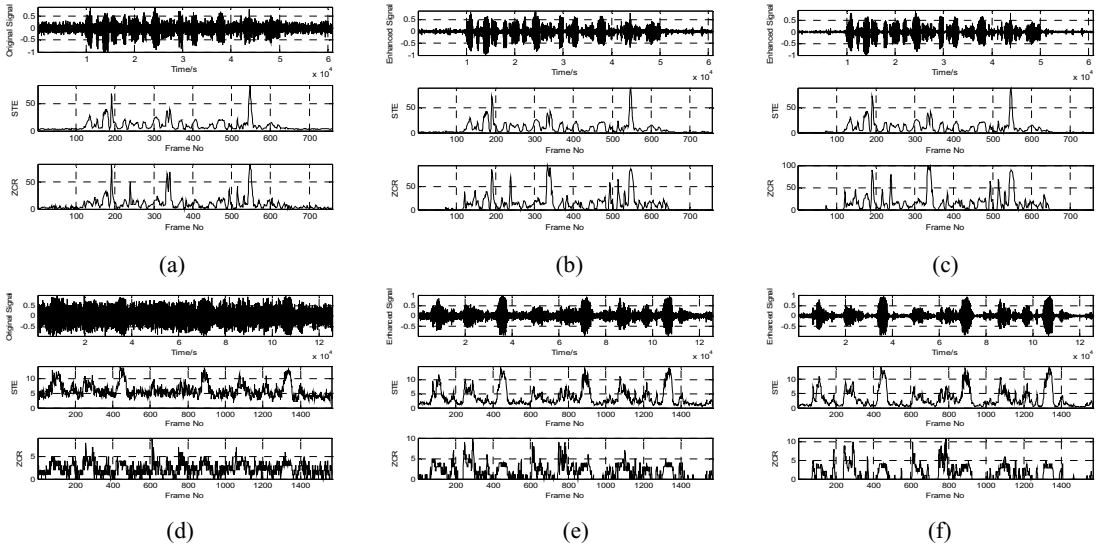

Figure 3 Performance of various methods in different environment. (a)(b)(c) a section of speech in car (SNR =8). (a) Traditional double thresholds VAD. (b) Proposed VAD (Iterative number=1). (c) Proposed VAD (Iterative number=2). (d)(e)(f) A section of true cockpit sound. (d) Traditional double thresholds VAD. (e) Proposed VAD (Iterative number=1). (f) Proposed VAD (Iterative number=2)