2014

ISSN : 0974 - 7435 Volume 10 Issue 20

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [12342-12348]

A voice activity detection algorithm based on spectral entropy analysis of sub-frequency band

Zhang Yuxin*, Ding Yan
School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022
E-mail: zyx@cust.edu.cn

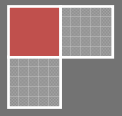
ABSTRACT

This paper proposes an effective voice activity detection (VAD) algorithms in low SNR noise environment. Traditional short-term energy and zero-crossing rate can only get high performance at high SNR environment. The spectral entropy algorithm is used to detect stationary noise signal, which is based on the inherent steady characteristics of noise signal. The whole spectrum is divided into some sub-bands, and then, the entropy value of sub-bands are computed separately. Since the voice change is stronger in some frequency bands, the sub-frequency band is extracted for detecting endpoint. The experimental results show that proposed method greatly improves performance of VAD at low SNR environment.

KEYWORDS

Voice activity detection; Spectral entropy; Sub-frequency band.

© Trade Science Inc.



INTRODUCTION

The human's speech is discontinuous. Thus, the ASR system begins to work when speech is detected. Usually, only the voice activity detection (VAD) programming runs in order to reduce the calculation cost of ASR system, when speech signal is nothing. Furthermore, the end-points of speech are accurately detected is important to improve the recognition accuracy of ASR system. Thus, VAD is a very important technique problem, especially in high ambient noise environments[1-3]. The accurate endpoint detection of speech is a simple problem in the most benign circumstances. In practice, one or more problems usually make accurate VAD difficult in the noisy background (e.g., fans or machinery running). In nonstationary environments (e.g., the presence of door slams, irregular road noise, car horns) with speech interference (as from TV, radio). Other factors are that the distortion introduced by the transmission system when the speech is sent, (e.g., cross-talk, intermodulation distortion, and various types of tonal interference arise to various degrees in the communications channel).

Many VAD methods have been proposed in speech recognition systems. VAD algorithm typically relies on the short-time energy and zero-pass ratio[4,5]. The associated techniques use different features of syllables in the time-domain and are low computational complexity. In the clean environment, these algorithms can be achieved very good results for endpoint detection. However, in low SNR environment, the detection accuracy greatly reduce, not even judge. In recent years spectral entropy method is more used. Be-cause there is big difference the human voice spectrum and the noise. We can distinguish speech and noise according to changes in the characteristics of the voice spectrum and it has noise robustness. Based on the traditional spectral entropy algorithm, the accuracy of detection greatly will decrease in nonstationary noise environments.

In this paper, we propose an algorithm which is based on short sub-band spectral entropy analysis. We divide the whole spectrum into sub-bands and they are limited. Then, we calculate entropy of every sub-band, from the result we can determine changes in signal intensity. Because changes of voice in the individual frequency domain is stronger, for endpoint detection, we may extract the sub-band entropy if it is strong. From the experiments results, this method is successful that it improves the VAD in low SNR environment performance.

TRADITONAL VAD METHOD

Short-time energy algorithm

Since the speech signal is a nonstationary processing, the way cannot be used to process speech signal, which is used to process stationary signal. The produced processing of speech signal is closely-related with physical working of phonatory organ. This physical working is slower than vibrations of sounds. The speech signal in $10 \sim 30$ms time can be as a quasi-steady signal (as short-time steady state), because the parameters of spectrum and physical characteristics are almost invariant. Thus, a speech signal can be divided into many short frames and every frame is as a detecting unit. According to the energies of speech and nonspeech frame, short-time energy based VAD approach can identify endpoints of any speech signal, because the energy of speech frame is larger than that of nonspeech frame[6].

The samples of a waveform of input speech signal is defined as $x(m)$, m is the sample index. The short-time square energy of speech signal $E(n)$ is defined as

$$E(n) = \sum_{m=-\infty}^{+\infty} [x(m)\omega(m-n)]^2 \quad \textbf{(1)}$$

Zero-crossing rate algorithm

Sometime, aforesaid short-time energy algorithms are inaccurate for VAD. The human's pronunciation include the surd and sonant. The sonant is produced by the vibration of the vocal chords. The amplitude of sonant is high and periodicity is apparent. The surd is without vibration of the vocal chords, it is produced by the friction, impact or plosive that the suction of air into the mouth. Thus, the short-time energy is lower than that of sonant. It can be identified into nonspeech easily by short-time energy method. The waveform of surd segment goes up and down so quickly around zero level value, and the number of crossing zero level value for nonspeech segment is fewer. The number of crossing zero level value can be used to distinguish the endpoint of speech signal. The method is described as zero-crossing rate (ZCR)[7]. The zero-crossing rate is defined as

$$Z(n) = \frac{1}{2n} \sum_{m=-\infty}^{+\infty} |sgn[x(m)] - sgn[x(m-1)]| \, \omega(m-n) \quad \textbf{(2)}$$

Where the $sgn[\cdot]$ is symbol function, it is defined as

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

The $\omega(n)$ usually uses the rectangular window function.

Double thresholds algorithm based on short-time energy and zero-crossing rate

The double thresholds algorithm sets two thresholds for speech signal. The starting of speech signal is detected by the hither threshold, and then the other threshold is used to accurately detect the real starting point of speech signal. The algorithm is described as follow.

Firstly, the threshold of zero-crossing rate is defined as

$$IZCT = \min(IF, \overline{IZC} + 2\zeta_{IZC}) \quad \textbf{(4)}$$

Where $\overline{IZC}$ is the average of ZCR of first five frames, ζ_{IZC} is the standard deviation of ZCR, IF is empirical value, usually $IF = 25$.

The short-time energy $E_i(n)$ of the first N frames is calculated. The maximum among short-time energies of all frames is defined as IMX, and the minimum is defined as IMN.

$$I_1 = 0.03 \times (IMX - IMN) + IMN \quad \textbf{(5)}$$

$$I_2 = 4 \times IMN \quad \textbf{(6)}$$

so ITL and ITU are defined as

$$ITL = min(I_1, I_2) \quad \textbf{(7)}$$

$$ITU = 5 \times ITL \quad \textbf{(8)}$$

Then, we detect the $E_i(n)$ of each frame which is from No $N + 1$ frame. If $E_i(n)$ of a frame is more than ITL, then the frame number is recorded as p_1, detecting continues. If $E_i(n)$ of a frame is lower than ITL, and all $E_i(n)$ are less than ITU, which frames are until current frame, then p_1 is updated to current frame number. Otherwise, the p_1^{th} frame is as the starting of speech signal.

Finally, we forward compare the ZCR of each frame from the p^{th} frame, if $ZCR(n)$ of continuous three frames are more than, then the p_1 is updated to first frame number of three frames. Otherwise, the starting of speech signal is still the p_1^{th} frame.

VAD based on spectral entropy

Recently spectral entropy becomes a research focus[8-10]. And the signal processing is performed in the frequency domain. Generally, we think that there is a certain link that the uncertainty of events and the probability of occurrence of the event. If there is small probability of the event, the uncertainty of the event is a big. Conversely, it is very small. If the uncertainty that the event has is big, the amount of information that it provide is big. The small uncertainty event provides less information. In speech recognition, noise energy generally is flat. And over each frequency it is relatively stable distribution. Because the speech signal undulate greatly, its energy is concentrated in few bands to the different pronunciation. The average amount of information on the noise signal spectral entropy is small. But the amount of information of the speech signal is big. So we can use this distinction to detect both speech and nonspeech section.

Algorithm description

Firstly, we frame noisy speech signal extracted. The frame length is 23.2ms (256 points). Offset frame is 11.6ms (128 points). Then we transform FFT for each frame. After that we obtain spectral components frame $v_i(f_k)$. Spectrum for each frame of each normalized spectral density is:

$$p_{i,k} = \frac{v_i(f_k)}{\sum_{k=1}^{N} v_i(f_k)} \; k = 1,2,\cdots,N \quad \textbf{(9)}$$

$p_{i,k}$ the spectral probability density of k^{th} frequency component of the frame. N FFT length (512 points). i is frame number. f_k is frequency components. $p_{i,k}$ is structured according to the magnitude of value.

Then, If the power spectrum of the energy structure can make the distribution of spectral entropy more smoothly. It can reflect the distribution of each component at each discrete points.

$$p'_{i,k} = \frac{|v_i(f_k)|^2}{\sum_{k=1}^{N} |v_i(f_k)|^2} \; k = 1,2,\cdots,N \quad \textbf{(10)}$$

The human's speech frequency focus from 250Hz to 3500Hz, to reduce the impact of noise on speech signal, we increase constraints $250\text{Hz} \le f_k \le 3500Hz$. If the frequency is not within it, $v_i(f_k) = 0$. In addition, if a frequency band is excessive, its possible reasons are what some noise frequency is too concentrated, which will lead to the probability density is too large. So we set an upper limit $\vartheta = 0.9$:

$$p'_{i,k} = 0 \; if \; p'_{i,k} > \vartheta \quad \textbf{(11)}$$

Finally, On the basis of the above definition, corresponding spectral entropy of each frame or entropy is:

$$\mathrm{H}_i = -\sum_{k=1}^{N} p'_{i,k} \cdot log p'_{i,k} \quad \textbf{(12)}$$

In practical situations, the environmental noise is varied. If the distribution of noise is not stable, in short time noise fluctuations is big, noise spectral entropy maybe over the value of speech. Which makes mistakes of VAD recognition. We define a smooth function, which cannot make noise spectral entropy values change larger because of short-term volatility fluctuation. The smooth function is:

$$H'_i = \psi H'_{i-1} + (1-\psi) H_i \quad \textbf{(13)}$$

ψ is balance coefficient $0 \le \psi \le 1$, and empirical coefficients. If the energy change of noise is big, its value should become small. Generally its value is $0.8 \sim 0.95$, $H'_0 = 0$. Simultaneously, when we detect that from H'_i of the first frame, there are M consecutive frames (generally $M = 10$) are greater than the threshold value. We take the i^{th} frame as a start end. So this is also a way to avoid these effects.

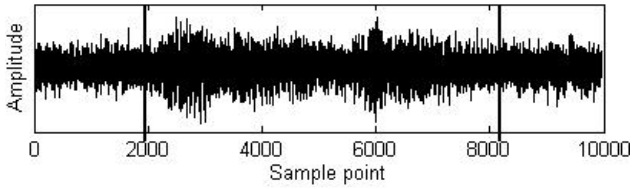


Figure 1 : Noisy speech with 0dB white noise

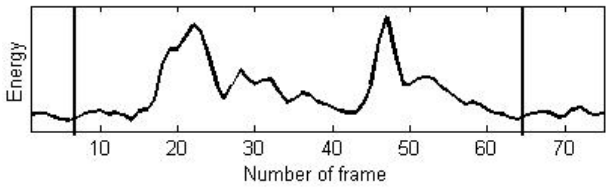


Figure 2 : VAD for STE

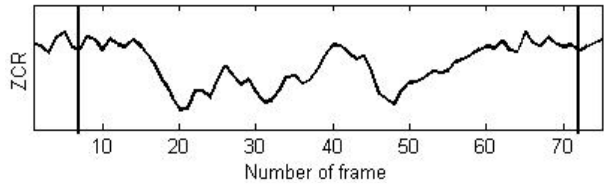


Figure 3: VAD for ZCR

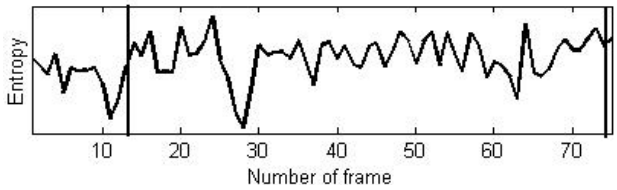


Figure 4 : VAD for spectral entropy

VAD BASED ON ENTROPY ANALYSIS OF SUB-FREQUENCY BAND

Figure 2-3 show The VAD based on STE and ZCR is suitable in high SNR environment. Figure 4 shows spectral entropy method is not well adapted to changes in the noise. Because the frequency of noise and speech is different, generating speech signal in its frequency energy will increase. However, the energy of the start and end portions of the speech signal is lower. So it is difficult to judge by the noise signal transition boundaries of the speech signal. From Figure 5, we found that the energy of the speech signal is not evenly distributed throughout the frequency spectrum. But there is large changes in the energy of certain frequencies, in other frequencies which is small. The high power channel of different short frame is different.

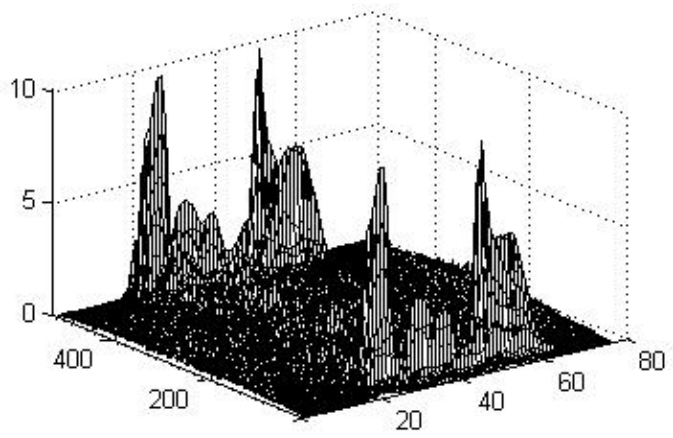


Figure 5 : 20dB the energy frequency of speech signal in white noise environment

As such, if we extract the bands that speech signal energy have big changes. Its energy variance must be much more pronounced than the entire frame fluctuation variance. VAD is even more favorable. If we calculate the energy changes of each band, the overall analysis is more difficult. The entire spectrum may be divided for several frequency bands for analysis. Based on the above analysis, we propose VAD algorithm that is short spectral entropy based on the frequency band.

Frequency of the speech signal is from 250Hz to 3500Hz. The band of sub-frequency band is 250Hz. The whole frequency is divided into 15 sub-bands. The length of short frame is 23.2ms (256 points), the shift of frame is 11.6ms (128 points). The transform point of FFT is 512 point. Each sub-frequency band is 34 points. Therefore, the energy of i^{th} sub-frequency band is defined as:

$$e_x(j) = \sum_{k=34j-33}^{34j} |v_i(f_k)|^2 , 1 \le j \le 15 \quad \textbf{(14)}$$

Each coming with a change in the energy spectrum of the entropy of the band can be represented. Since the spectral entropy band with VAD applications, the number of frames is creasing. Band spectral entropy is difficult to determine. Therefore, we define a sliding window and it only counts the sub- band spectral entropy falling into the sliding window frame. As the window slides forward constantly, there is a new frame falling into the window. At the forefront of the window frame will leave the window. When the noise signal is a stationary distribution and sliding window is in the noise section, its entropy is little change. When the energy of the speech signal into the window of some great bands may increase, it leads entropy increase. This apparent change is beneficial speech judgment.

Algorithm description

Step 1: Frame noisy speech signal, then make FFT for each frame.

Step 2: The spectrum is divided into 15 sub-bands for each frame, and then calculate the energy of each sub-frequency band.

Step 3: Set the value of the sliding window, $w = 10$. Usually, the first frame is considered as non-speech, so the first 10 frames are in the window. And then calculate the value of each sub-band spectral entropy. Then take it as a noise spectral entropy valuation as a determination reference of the rear frame. Short sub- band spectral entropy is:

$$p_x(j) = \frac{e_x(j)}{\sum_{x=1}^{w} e_x(j)} \quad \textbf{(15)}$$

$$H(j) = -\sum_{x=1}^{w} p_x(j) \cdot log p_x(j) \quad \textbf{(16)}$$

Step 4: Setting sub- band spectral entropy threshold $\overline{H(j)} = \omega H(j)$, ω is experience.

Step 5: About endpoint detection. Sliding window starts from the initial state, in turn slide forward a frame size and calculate the value of the sliding window $\mathcal{W}_i$.

The value of the sliding window is defined: The number of which sub- band spectral entropy value is greater than the threshold value of $H(j) > \overline{H(j)}$ is calculated. In the initial state $\mathcal{W}_0 = 0$. If $\mathcal{W}_i < 2$, the window slides forward. If $\mathcal{W}_i \ge 10$, then endpoint is in the window. If 5 consecutive values satisfy window is $2 \le \mathcal{W}_i \le 10$, so the endpoint is in the first one window of continuous 5 windows.

Step 6: Calculate the sum of all the sub- band energy for each frame of each frame $H(j) > \overline{H(j)}$ of the window. Find forward from the first 10. If sub-band energy continuously decreases, the last frame is the starting point.

Step 7: Determining an end point is similar to the method described above. After finding the window, a look back from the first sub-band energy continuously decreasing, the last frame is the end.

EXPEREMENT AND RESULT

In the paper we use a standard library Noisex 92 white and babble noise. Which is added to the clean speech and test SNR under 20dB, 10dB, 0dB three cases.

TABLE 1 : Detection accuracy in 0dBwhite noise (%)

VAD	SNR		
	20dB	**10dB**	**0dB**
Double threshold method	93	72	40
Spectral entropy method	94	83	71
Sub- band spectrum entropy method	96	89	83

TABLE 2 : Detection accuracy in 0dB babble noise (%)

VAD	SNR		
	20dB	**10dB**	**0dB**
Double threshold method	92	70	32
Spectral entropy method	94	82	69
Sub- band spectrum entropy method	95	89	81

TABLE 1 and TABLE 2 show the detection accuracy of the method based on short-term energy and zero-crossing rate double threshold, traditional spectral entropy method and sub-band analysis method based on entropy. Experiments show that the method of short-term energy and zero-crossing rate are almost undetectable in SNR $<$ 0dB. The spectral entropy method has strong noise immunity. In low SNR we can get better recognition rate. However, the method we proposed here is better than other algorithms in 0～20dB.

CONCLUSIONS

The VAD methods of short-term energy and zero-crossing rate only apply in high SNR environment. Though spectral entropy detection algorithm is effective in low SNR environment, spectrum entropy only consider the difference of each band of internal frame. And it didn't take into the difference between adjacent frames. If the noise spectrum is very similar to the spectrum of the speech, it is difficult to distinguish between speech and nonspeech by entropy. Therefore, by dividing the band we can analysis energy change of the main band of speech. Obviously, the change can be get. Setting sliding window reflects the continuity of speech data, which can avoid identification errors of instantaneous increase of entropy that the sudden emergence of short-term noise leading. And we only calculate the frequency bands which has large energy changes. At the same time it increases the speed of calculation. Final, though the sliding window algorithm reflects the correlation between adjacent frames, it increase the amount of computation. Our future work will attempt to improve the speed of the moving window quickly.

REFERENCES

[1] J.Dong, X.H.Zhao; Ou Shifeng; Robust Endpoint Detection Algorithm of Chinese, Paper presented at 4th WiCOM'08, Dalian, China, 01-05 **(2008)**.

[2] H.Ghaemmaghami, R.Vogt, S.Sridharan; Speech Endpoint Detection Using Gradient Based Edge Detection Techniques, Paper presented at the 2nd ICSPCS 2008. Australia, 01-08 **(2008)**.

[3] L.Rabiner, B.H.Juang; Fundamentals of speech recognition, 1st ed. Upper Saddle River, New Jersey, USA: Prentice Hall PTR, **(1993)**.

[4] Xu Dawei,Wu Bian, Zhao Jianwei et al.; A real time algorithm for voice activity detection in noisy environment, Computer Engineering and Application, **24(1)**, 115-117 **(2003)**.

[5] Liang-Sheng, Huang, Chun-Ho Yang; A Novel Approach to Robust Speech Endpoint Detection in Car Environments. Paper presented at ICASSP200, 1751-1754 **(2000)**.

[6] N.Erdol, C.Castelluccia, A.Zilouchian; Recovery of missing speech packets using the short-time energy and zero-crossing measurements, IEEE Transactions on Speech and Audio Processing, **1(3)**, 295–303 **(Jul. 1993)**.

[7] Y.K.Lau, C.K.Chan; "Speech recognition based on zero crossing rate and energy," IEEE Transactions on Acoustics, Speech and Signal Processing, **33(1)**, 320–323 **(Feb. 1985)**.

[8] C.Jia, B.Xu; An improved entropy-based endpoint detection algorithm, Paper presented at ICSLP 2002, Taipei, 285-288 **(2002)**.

[9] Fan Yingle, Li Yi, Wu Chuanyan; Speech Endpoint Detection Based on Speech Time-Frequency Enhancement and Spectral Entropy, Paper presented at 27th Annual International Conference of the Engineering in Medicine and Biology Society, 4682-4684 **(Jan. 2006)**.

[10] N.Majstorovic, M.Andric, D.Mikluc; Entropy-based algorithm for speech recognition in noisy environment, Telecommunications Forum (TELFOR), 670 **(Nov. 2011)**.