Voice Activity Detection Based on The Bispectrum

Hui-jing Dou Zhao-yang Wu
School of Electronic Information and Control Engineering
Beijing University of Technology, Beijing, China
dhjing@bjut.edu.cn wuzhaoyang2007@163.com

***Abstract*—In order to improve the performance of voice activity detection under multiple noise environments, a new voice activity detection algorithm based on the bispectrum was presented. This method detect the speech signal by using the special slice of the bispectrum. The performance of the algorithm was compared with the voice activity detection algorithm in standard G.729 annex B of ITU in experimental simulation. Experimental results show that the algorithm has high classification accuracy and stability.**

Keywords-voice activity detection; bispectrum; noise environment;

Ⅰ. INTRODUCTION

With the arrival of digital informational era and the rapid development of mobile communication technology, high quality speech communication is desired. Speech signal processing has become a very important component of voice communication systems. Moreover, Voice activity detection (VAD) still plays an important role in all kinds of speech processing system, for example, it can improve the channel capacity for voice communications, reduce co-channel interference, and extend the battery life of portable equipments of a cellular radio system.

Voice activity detection (VAD) has a long history, From the earliest based on short-term energy and zero crossing rate to the voice-based models and statistical knowledge of the various complex algorithms, voice activity detection algorithm theory and implementation methods are constantly updated. With the rapid development of communication technology, the VAD algorithm based on the short term energy and crossing rate has not meeted the demand of the people. Subsequently, various of the VAD algorithm such as based on the cepstral coefficients[1], entropy[2] has been proposed.

In recent year, some scholars have proposed voice activity detection based on the statistical model[3][4]. There are also have the VAD algorithm based on support vector machine[5], wavelet theory[6][7], neural network[8], Cyclic statistics[9]. Hence, how to ensure the robustness of the algorithm while reducing the complexity of the algorithm is a challenging issue.

In this paper, a voice activity detection algorithm is presented based on the bispectrum of speech signal. The paper is organized as follows: The method of bispectrum estimation is introduced in Section Ⅱ. The feature extraction is introduced in section Ⅲ. The proposed algorithm is introduced in section Ⅳ. The experimrntal results are given in section Ⅴ. The conclusion is drawn in section Ⅵ.

Ⅱ. THE ESTIMATION OF BISPECTRUM

In the definition of power spectrum, which need the autocorrelation function is absolutely sum. Similarly, in order to ensure the Fourier transform existence of higher-order moments and cumulants, also require higher-order moments and cumulants is absolutely sum, namely

$$\sum_{\tau_1=-\infty}^{\infty}\cdots\sum_{\tau_{k-1}=-\infty}^{\infty}\left|m_{kx}(\tau_1,\tau_2,\cdots,\tau_{k-1})\right|<\infty \qquad (1)$$

$$\sum_{\tau_1=-\infty}^{\infty}\cdots\sum_{\tau_{k-1}=-\infty}^{\infty}\left|c_{kx}(\tau_1,\tau_2,\cdots,\tau_{k-1})\right|<\infty \qquad (2)$$

Hence, the definition of the higher-order moment spectra and higher-order cumulants spectra as follows:

$$M_{kx}(\omega_1,\cdots,\omega_{k-1})=\sum_{\tau_1=-\infty}^{\infty}\cdots\sum_{\tau_{k-1}=-\infty}^{\infty}m_{kx}(\tau_1,\tau_2,\cdots\tau_{k-1})e^{-j(\omega_1\tau_1+\cdots+\omega_{k-1}\tau_{k-1})} \qquad (3)$$

$$S_{kx}(\omega_1,\cdots,\omega_{k-1})=\sum_{\tau_1=-\infty}^{\infty}\cdots\sum_{\tau_{k-1}=-\infty}^{\infty}c_{kx}(\tau_1,\cdots,\tau_{k-1})e^{-j(\omega_1\tau_1+\cdots+\omega_{k-1}\tau_{k-1})} \qquad (4)$$

Where $m_{kx}(\tau_1,\tau_2\cdots,\tau_{k-1})$ is the higher-order moments of $x(n)$, $c_{kx}(\tau_1,\tau_2\cdots,\tau_{k-1})$ is the higher-order cumulants of $x(n)$.

In this article, in order to reduce the complexity of the algorithm, the author adopt the bispectrum. The definition of bispectrum as follows:

$$B_x(\omega_1,\omega_2)=\sum_{\tau_1=-\infty}^{\infty}\sum_{\tau_2=-\infty}^{\infty}c_{3x}(\tau_1,\tau_2)e^{-j(\omega_1\tau_1+\omega_2\tau_2)} \quad (5)$$

Where $c_{3x}(\tau_1,\tau_2)$ is the third-order cumulants of $x(n)$.

In practice, the estimation of the bispectrum usually by the direct estimation approach. Assume that $x(0),x(1),\cdots,x(N-1)$ is the data that have zero mean, and the sample frequency is f_s. The estimation approach of the bispectrum has four steps:

The first step is dividing the data into K segment, each segment has M samples, namely $x^{(k)}(0)$, $x^{(k)}(1),\cdots$, $x^{(k)}(M-1)$, where $k=1,2,\cdots,K$. Each segment was allowed to have overlap.

The second step is computing the DFT coefficient.

$$X^{(k)}(\lambda)=\frac{1}{M}\sum_{n=0}^{M-1}x^{(k)}(n)e^{-j2\pi n\lambda/M} \quad (6)$$

Where $\lambda=0,1,\cdots,M/2;k=1,2,\cdots,K..$

The third step is computing the triple correlation of the DFT coefficient:

$$\hat{b}_k(\lambda_1,\lambda_2)=\frac{1}{\Delta_0^2}\sum_{i_1=-L_1}^{L_1}\sum_{i_2=-L_1}^{L_1}X^{(k)}(\lambda_1+i_1)X^{(k)}(\lambda_2+i_2)X^{(k)}(-\lambda_1-\lambda_2-i_1-i_2)$$
$$k=1,2,\cdots,K;0\le\lambda_2\le\lambda_1,\lambda_1+\lambda_2\le f_s/2 \quad (7)$$

Where $\triangle_0=f_s/N_0$, but the select of the N_0 and L_1 should satisfy $M=(2L_1+1)N_0$.

The last step is computing the mean of the bispectrum estimation. Hence, the bispectrum estimation of $x(0),x(1),\cdots,x(N-1)$ can express as follows:

$$\hat{B}_D(\omega_1,\omega_2)=\frac{1}{K}\sum_{k=1}^{K}\hat{b}_k(\omega_1,\omega_2) \quad (8)$$

Where $\omega_1=\frac{2\pi f_s}{N_0}\lambda_1,\omega_2=\frac{2\pi f_s}{N_0}\lambda_2$.

Figure1, figure 2 are the simulation of bispectrum estimation, figure 1 is the bispectrum estimation of clean speech, figure 2 is the bispectrum estimation of de-noised speech under carinterior noise, which the SNR is 0dB.

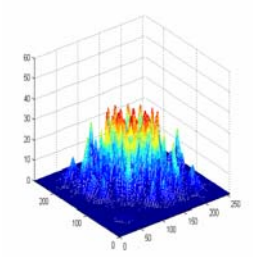


Figure 1. The bispectrum estimation of the clean speech

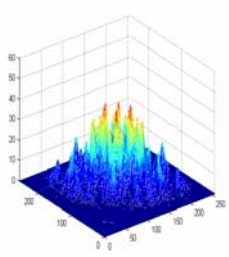


Figure 2. The bispectrum estimation of the de-noised speech(carinterior 0dB)

From figure1 and figure 2, it is difficulty to find the difference of bispectrum between the clean speech and de-noised speech. In order to reduce the complexity, only analyse the diagonal slice.

III. SPEECH FEATURE EXTRACTION

From the above analysis, it was easy to find that there is no difference between the clean speech and the de-noised speech except the amplitude. In order to find the difference between the clean speech and the de-noised speech, let's us discuss the diagonal slice. In this paper, the diagonal slice is gained by extracting the diagonal of the bispectrum estimation. Figure 3 show the diagonal slice of clean speech bispectrum estimation, the horizontal axis is the length of fast fourier transform, the vertical axis is the amplitude. At the same time, for subsequent analyse convenient, normalized the horizontal axis which show by the figure 4. From the figure 3 and figure 4, it is easy to find the diagonal slice of clean speech bispectrum estimation is symmetrical. Hence, in this article, the author analysed the half.

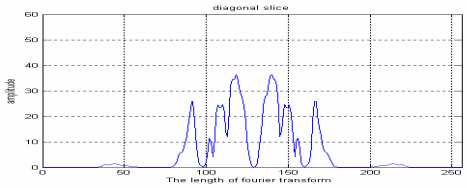


Figure 3. The diagonal slice of clean voice bispectrum estimation

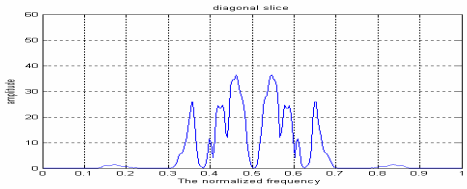


Figure 4. The diagonal slice of de-noised voice bispectrum estimation

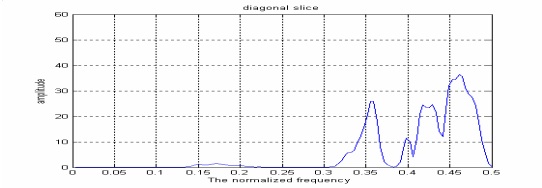


Figure 5. The diagonal slice of clean voice bispectrum estimation

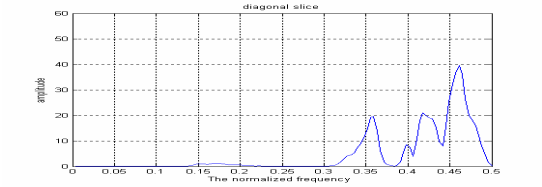


Figure 6. The diagonal slice of de-noised voice(carinterior 0dB) bispectrum estimation

Figure 5 represent the simulation results the diagonal slice of clean voice bispectrum estimation, and figure 6 represent the diagonal slice of de-noised voice bispectrum estimation under in carinterior noise which the SNR is 0dB. From the figures, it is easy to find that the noise affect the range from 0.4375 to 0.5, besides, there is no affection in other range. From the figures 7 and figures 8, it also can be found that the un-voice has the same case, the noise affect the range from 0.4375 to 0.5, besides, there is no affection in other range.

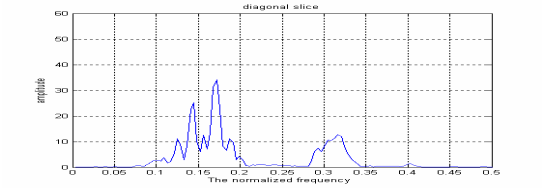


Figure 7. The diagonal slice of clean un-voice bispectrum estimation

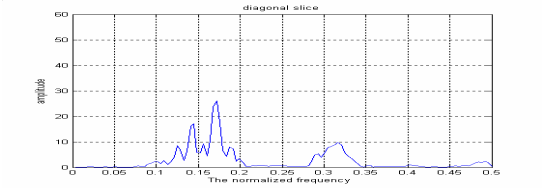


Figure 8. The diagonal slice of de-noised un-voice(carinterior 0dB) bispectrum estimation

From the above analyzing, it is easy to find that as to voice, the range from 0.125 to 0.25 and from 0.3125 to 0.4375 have little change. As to un-voice, there is little change from 0.125 to 0.375. Hence, the author choose the intersection and defined E_l as the energy from the range 0.125 to 0.25, defined E_h as the energy from the range 0.3125 to 0.375, the equation as follows:

$$E_l = \sum_{i=0.125}^{0.25} \left| \hat{B}_D(i) \right|^2 \qquad (9)$$

$$E_h = \sum_{i=0.3125}^{0.375} \left| \hat{B}_D(i) \right|^2 \qquad (10)$$

Ⅳ. THE VAD ALGORITHM USING BISPECTRUM

The proposed algorithm is described as follows.

Firstly, subtracting mean and dividing standard deviation from $x(n)$. The standardized signal $x'(n)$ is gained, namely:

$$x'(n) = \frac{x(n) - E(x(n))}{std(x(n))} \qquad (11)$$

Where $std(x(n))$ represent the sample standard deviation of the data in $x(n)$.

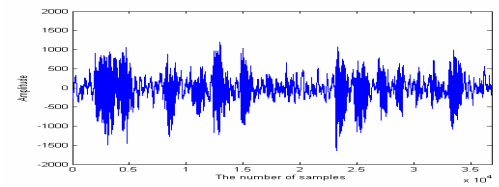
Secondly, computing the bispectrum based on the direct method, the calculation steps are expressed in section 2.

Thirdly, computing E_l and E_h. The equations are expressed in equation (9) and equation (10).

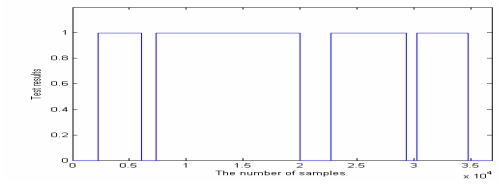
Finally, make the voice activity detection decision. The rule is if each of E_l and E_h is above the threshold, then detected as speech, else detected as silence or noise.

Ⅴ. THE EXPERIMENTAL RESULTS

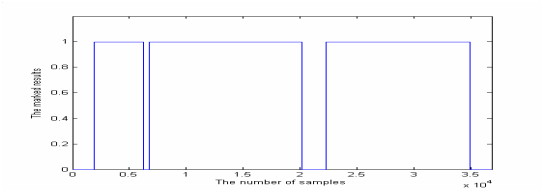
In the experiments, 1350 frames of Chinese speech are used for test. The speech signals were sampled at 8000Hz with 16-bit resolution. Each speech signal is divided into frames whose length is 160 samples(20ms) with 0 samples overlap between adjacent frames. Figure 9(a) represents the de-noised speech which under the carinterior noise and the SNR is 0dB, figure 9(b) gives a graphical representation of the proposed VAD algorithm in operation under car noise conditions at 0dB SNR, figure 9(c) represents the marked results.



(a) The de-noised speech(SNR=0dB)



(b) The test results



(c) The marked results

Figure 9. The speech waveforms and test results

In order to further evaluate the performance of the algorithm, compare this algorithm to G.729 Annex B VAD algorithm. Test environment were selected factories noise, babble noise and carinterior noise and the SNR were -5dB, 0dB, 5dB, 10dB, 15dB and 20dB six different SNR. In this paper, the performance of the algorithm is measured by P_{cs}, P_{cs} can be defined as the ratio of the correctly classified frames to the total number of frames.

Table Ⅰ. Performance evaluation in different carinterior conditions

	Proposed VAD	G.729 B VAD
-5dB	92.5%	88.9%
0dB	94.1%	90.5%
5dB	94.5%	91.6%
10dB	95.5%	92.3%
15dB	96.2%	93.7%
20dB	97.4%	95.4%

Table Ⅱ. Performance evaluation in different factory conditions

	Proposed VAD	G.729 B VAD
-5dB	89.2%	85.7%
0dB	90.1%	87.1%
5dB	92.1%	88.8%
10dB	94.5%	90.1%
15dB	96.1%	93.1%
20dB	97.2%	95.4%

Table Ⅲ. Performance evaluation in different babble conditions

	Proposed VAD	G.729 B VAD
-5dB	87.2%	84.5%
0dB	88.3%	85.7%
5dB	91.9%	87.8%
10dB	92.3%	89.1%
15dB	96.0%	91.5%
20dB	96.8%	93.7%

Tables I-III show the performance of the proposed voice activity detection algorithm and the G.729 B VAD algorithm under carinterior noise, factory noise and babble noise conditions. From table I-III, it can be seen that our algorithm achieves better accuracy than the G.729 B VAD. Specially, the proposed algorithm has the good performance even if under in low SNR.

Ⅵ. CONCLUSIONS

In this paper, the author presented a robust voice activity detection algorithm, which based on the bispectrum estimation. Firstly, the author analysed the difference of speech signal and noise signal in bispectrum. Secondly, the author extracted the feature. Finally, the author proposed a voice activity detection algorithm and simulated the VAD algorithm. Experimental results show that the algorithm has high classification accuracy and stability, which is an effective voice activity detection algorithm.

REFERENCES

[1] J.A. Haigh and J.S. Mason, “Robust Voice Activity Detection Using Cepstral Features”, In Proc. of IEEE TENCON'93, vol. 3, pp. 321-324, 1993, Beijng.

[2] Kun-Ching Wang and Yi-Hsing Tasi. “Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy”. 2008 Second International Symposium on Universal Communication. pp:423-428.

[3] Joon-Hyuk Chang, Nam Soo Kim and Sanjit K. Mitra. “Voice Activity Detection Based on Multiple Statistical Models”. IEEE Transactions On Signal Processing, 2006, 54(6):1965-1976.

[4] Oliver Gauci, Carl J. Debono and Paul Micallef. “A Maximum Log-Likelihood Approach to Voice Activity Detection”. 2008 3rd International Symposium on Communications, Control, and Signal Processing, ISCCSP 2008. pp:383-387.

[5] Chen, Shi-Huang, Guido, Rodrigo Capobianco and Chen, Shih-Hao. “Voice activity detection in car environment using support vector machine and wavelet transform”. ISM Workshops 2007 - 9th IEEE International Symposium on Multimedia – Workshops.252-255.

[6] Kh. Aghajani, M.T. Manzuri, M. karami, H. tayebi. “A Robust Voice Activity Detection Based on Wavelet Transform”. 2008 Second International Conference on Electrical Engineering (ICEE),2008, Lahore, Pakistan.

[7] J. Shaojun, G. Hitato, Y. Fuliang, “A New Algorithm For Voice Activity Detection Based On Wavelet Transform”. proc. of int .symposium of intelligent multimedia, video and speech processing, pp. 222-225, Oct. 2004, Hong Kong.

[8] Pham, T.V. Tang, C.T.and Stadtschnitzer, M. “Using artificial neural network for robust voice activity detection under adverse conditions”. 2009 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF). Research, Innovation and Vision for the Future.pp:1-8.

[9] Huijing Dou, Changchun Bao and Ruwei Li. “A voice activity detection using cyclic statistics based on sinusoidal speech model”. 2008 9th International Conference on Signal Processing, ICSP 2008. pp:1239-1242.