Innovative VAD Based on Horizontal Spectral Entropy with Long-Span of Time

Kun-Ching Wang, Chiun-Li Chin and Chun-Ming Wang

***Abstract*—This paper shows innovative VAD based on horizontal spectral entropy with long-span of time (HSELT) feature sets to improve mobile ASR performance in low signal-to-noise ratio (SNR) conditions. Due to the signal characteristics of nonstationary noise change with time, we need long-term information of the noisy speech signal to define a more robust decision rule yielding high accuracy. We find that HSELT measures can horizontally enhance the transition between speech and non-speech segments. Based on the above finds, we can use the HSELT measures to achieve high accuracy for detecting speech signal form various stationary and nonstationary noises.**

***Index Terms*—voice activity detection, horizontal spectral entropy, long-term, Mel-scaled filter bank**

I. Introduction

In fact, in a mobile or portable environment, VAD mechanism has to distinguish active speech from noise with low signal to noise ratio (SNR). Most of these VAD algorithms assume that the background noise statistics are stationary over a longer period of time than those of noise. In general, no particular feature or specific set of features has been shown to perform uniformly well under different noise conditions. For example, energy-based features do not work well in low SNR [1] and, similarly, under colored noise, entropy measure fails to distinguish speech from noise with good accuracy due to the colored spectrum of speech [2]. Also SNR estimation is a critical component in many of the existing VAD schemes, which is particularly difficult in non-stationary noise [3]. Ramirez *et al*. [9] proposed the use of long-term spectral divergence between speech and noise for VAD, although they assign the VAD decision directly to the frame in the middle of the chosen long analysis window.

In this letter, we propose innovative VAD based on horizontal spectral entropy with long-span of time (HSELT). Due to that the HSELT measure can be used to discriminate noise from noisy speech signal and, hence, can be used as a potential feature for voice activity detection (VAD). First, the 17 log-energies are derived through Mel-scaled filter bank and are composed of a lowest frequency (1-8 bark) part, a low frequency (9-12 bark) part, a high frequency (13-15 bark) part and a highest frequency (16-17 bark) part. Due to the signal characteristics of nonstationary noise change with time, we need long-term information of the noisy speech signal to define a more robust decision rule yielding high accuracy. We find that HSELT measure can enhance the transition from non-speech to speech-only or from speech-only to non-speech. So, the HSELT measure can be used to detect the endpoint of speech signal.

Manuscript received Jan 8, 2013; This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC 101-2221-E-158-005.

K. C. Wang is with the Department of Information Technology & Communication, Shin Chien University, 200 University Road, Neimen, Kaohsiung 84550, Taiwan (e-mail: kunching@mail.kh.usc.edu.tw)

C. L. Chin and C. M. Wang are with the Department of Applied Information Sciences, Chung Shan Medical University, No.110, Sec.1, Jianguo N.Rd., Taichung City Taiwan (e-mail: ernestli@csmu.edu.tw); (e-mail: q3721451@yahoo.com.tw)

II. The Proposed VAD Method

A. Mel-Scale Filter Bank

In fact, human ear perceives speech along a nonlinear scale in the frequency domain. Based on the finding, we use a filter bank, spaced uniformly on a nonlinear, warped frequency scale frequency and frequency (hertz), and described by the following equation [4]:

$$mel = 2595 \cdot \log(1 + f/700) \tag{1}$$

where mel is the mel-frequency scale and f is in hertz. The mel-scale filter bank of 17 bands are approximated by simulating 17 triangular bandpass filters, $f(\xi,k)$ $(1 \le \xi \le 17,\ 0 \le k \le 127)$, over a frequency range of 0-4KHz. With the mel-scale frequency bank, the energy of each frequency band for each time frame of a speech signal can be calculated. Consider a given time-domain noisy speech signal, $x_{time}(m,n)$, representing the magnitude of the nth point of the mth frame.

The spectrum, $x_{freq}(m,k)$, of this signal is first calculated by discrete Fourier transform (256-point DFT)

$$x_{freq}(m,k) = \sum_{n=0}^{N-1} x_{time}(m,n) \cdot \exp(-j2\pi/N)^{kn}, \quad 0 \le k \le N-1;\ 0 \le m \le M-1 \tag{2}$$

where $x_{freq}(m,k)$ is the magnitude of the kth point of the spectrum of the mth frame, is 256 in our system, and M is the number of frames of the speech signal for analysis.

The spectrum $x_{freq}(m,k)$ is then multiplied by the weighting factors $f(\xi,k)$ on the mel-scale frequency bank. We can sum the products for all k to get the energy $x(m,\xi)$ of each frequency band ξ of the mth frame

$$x(m,\xi) = \sum_{k=0}^{N-1} \left| x_{freq}(m,k) \right| \cdot f(\xi,k), \quad 0 \le m \le M;\ 1 \le \xi \le 17 \tag{3}$$

where $f(\xi,k)$ also represents the weighting factor of the frequency energy at the kth point of the ξth band.

In fact, some undesired noise is resulted from our experiments that the energy $x(m,\xi)$ obtained in Eq.(3). Hence, a three-point median filter is further used to get the smoothed energy, $\hat{x}(m,\xi)$

ISBN: 978-988-19252-6-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

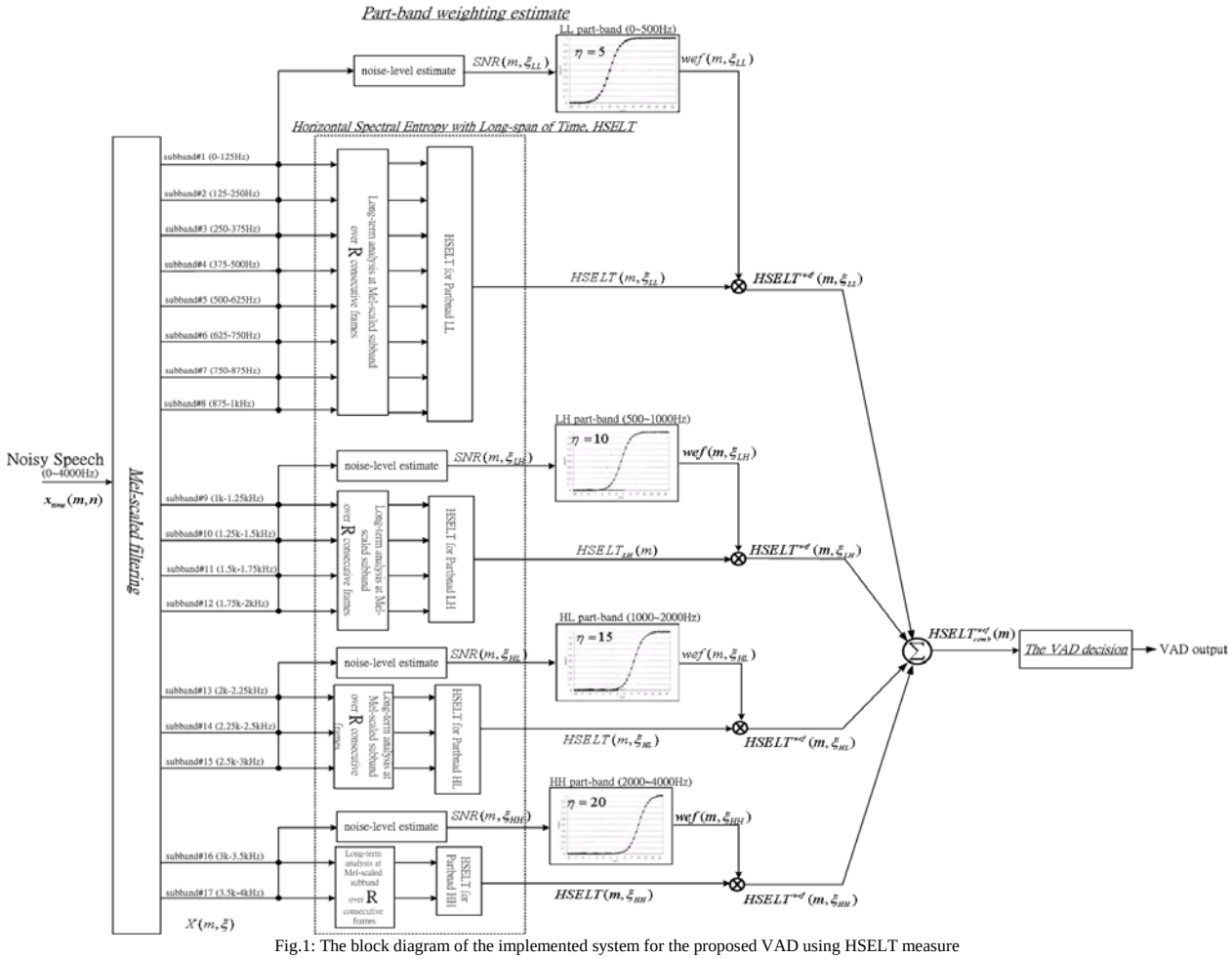

Fig.1: The block diagram of the implemented system for the proposed VAD using HSELT measure

$$\hat{x}(m,\xi)=\frac{x(m-1,\xi)+x(m,\xi)+x(m+1,\xi)}{3}. \quad (4)$$

Finally, the energy, $X(m,\xi)$, can be normalized by removing the frequency energy of the beginning interval, BGN, from the smoothed energy, $\hat{x}(m,\xi)$

$$X(m,\xi)=\hat{x}(m,\xi)-\mathrm{BGN}=\hat{x}(m,\xi)-\frac{\sum_{j=0}^{4}\hat{x}(j,\xi)}{5}. \quad (5)$$

where BGN is the energy of the beginning interval estimated by averaging the frequency energy of the first five frames of the recording.

B. Definition of the HSELT

This subsection derives a parameter, which can estimate the degree of nonstationary of the signal. We find that HSELT measure can enhance the transition from non-speech to speech-only or from speech-only to non-speech. So, the HSELT measure can be used to enhance the endpoint of speech/non-speech signal. The HSELT measure at any time is computed using the last R frame of the observed signal $x(n)$ with respect to the current frame of interest. The HSELT, $HSELT(m,\xi)$, at frequency subband ξ for the mth frame is computed as follows:

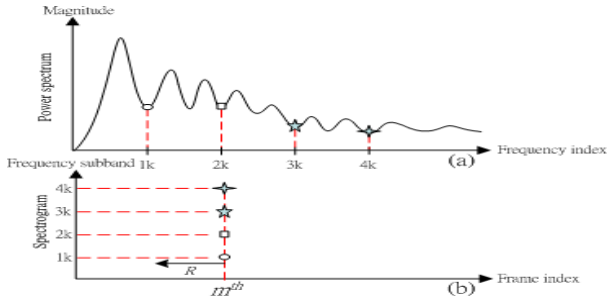

Fig.2: The view of HSELT measure: (a) Power spectrum amplitude for 4 kHz bandwidth signal. (b) Spectrogram for an entropy measure on the normalized short-time spectrum computed at frequency ξ over R consecutive frames, ending at the mth frame.

$$HSELT(m,\xi)=\sum_{n=m-R+1}^{m}\frac{X(n,\xi)}{\sum_{l=m-R+1}^{m}X(l,\xi)}\times\log\left(\frac{X(n,\xi)}{\sum_{l=m-R+1}^{m}X(l,\xi)}\right) \quad (6)$$

where $X(m,\xi)$ is normalized spectrum energy and is defined later, and $HSELT(m,\xi)$ is essentially an entropy measure on the normalized short-time spectrum computed at frequency subband ξ over R consecutive frames, ending at the mth frame (as shown in Fig.2).

In Fig.3, it shows the degrees of nonstationary between the non-speech frame and speech frame over R consecutive

ISBN: 978-988-19252-6-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

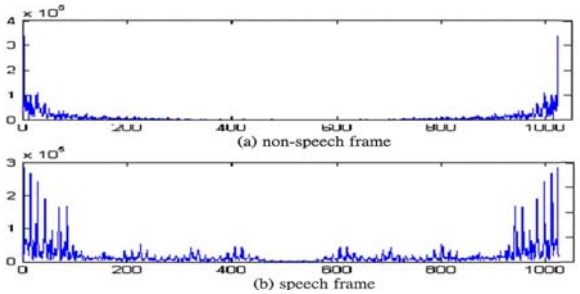

Fig.3: The degrees of non-stationary of the non-speech frame and speech frame.frames for specific frequency subband. We can find the degree of nonstationary during speech segment is larger than

that during non-speech segment, especially at transition between non-speech and speech. So, we can detect the endpoint of speech/non-speech by horizontally get the entropy value over R consecutive frames at specific frequency subband.

In order to further describe the degree of nonstationary of the signal, we only check the four part-bands and reduce the complexity to determine a reliable HSELT value. So, we merge 17 critical subbands into four part-bands: 0~1kHz (LL part band: 1-8 bark), 1~2kHz (LH part band: 9-12 bark), 2~3kHz (HL part band: 13-15 bark) and 3~4kHz (HH part band: 16-17 bark). Consequently, the HSELT parameter at ξ_pth part band is computed as below:

$$HSELT(m,\xi_p)=\frac{1}{K_{\xi_p}}\sum_{\xi\in\xi_p} HSELT \qquad (7)$$

$$\xi_p\in\{LL,LH,HL,HH\}.$$

where K_{ξ_p} is subband number at ξ_pth part-band.

C. Part-Band Weighting Estimation

Due to that the influence of noise upon to the detection performance, we need a parameter will help us sense how much the current part-band is corrupted by noise. In order

to determine the part-band utility rate on ξ_p part for mth frame, a posterior SNR, $SNR^{pot}(m,\xi_p)$ is required, and it is formulated as:

$$SNR^{pot}(m,\xi_p)=10\cdot\log_{10}\frac{P_{N+S}(m,\xi_p)}{P_N(m,\xi_p)} \qquad (8)$$

where $P_{N+S}(m,\xi_p)$ means power energy range from each part for the observed noisy speech signal. $P_N(m,\xi_p)$ is the estimated noise power on ξ_pth part for mth frame.

Observing the Eq.(8), we know that the subband noise power spectrum has to be estimated while determining the value of a posterior SNR. Various methods [5] were proposed for tracking the minimum of the noisy speech power spectrum energy over a fixed search window length in order to estimate the noise-level quickly and accurately. To speed up the determination of local minimum of noisy speech spectrum over a search window size, Doblinger's efficient method [6] is used here, which is not constrained by any window length to update noise spectrum estimate.

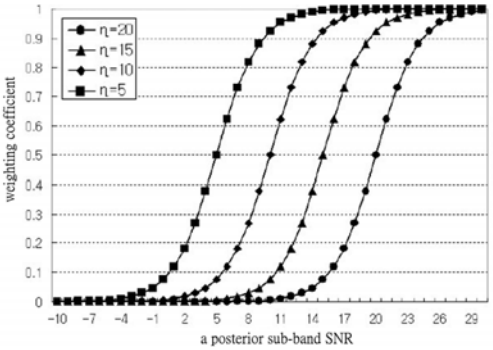

Fig.4: The plots of weights coefficients against a posterior SNR.

$$\begin{aligned}&\text{If } P_{\min}(m-1,\xi_p)<P_{N+S}(m,\xi_p),\\&\text{then } P_{\min}(m,\xi_p)=\gamma\cdot P_{\min}(m-1,\xi_p)+\\&\qquad\frac{1-\gamma}{1-\beta}\left[P_{N+S}(m,\xi_p)-\beta\cdot P_{N+S}(m-1,\xi_p)\right],\\&\text{else } P_{\min}(m,\xi_p)=P_{N+S}(m,\xi_p),\end{aligned} \qquad (9)$$

where $P_{\min}(m,\xi_p)$ denotes the local minimum of power energy of the noisy speech. γ and β are constants determined experimentally.

After the value of a posterior SNR obtained, the part-band weight coefficient, $wef(m,\xi_p)$, is calculated by applying a sigmoid function:

$$wef(m,\xi_p)=1/1+\exp\left[-0.5\cdot\left(SNR(m,\xi_p)-\eta(m,\xi_p)\right)\right] \qquad (10)$$

where $\eta(m,\xi_p)$ is a center-offset of the sigmoid function and is depended on part-band index. Therefore, we will use this information to weight each part-band. Fig.4 shows the plots of the weighting coefficients from Eq.(10) depending on η. Under the fixed value of a posterior SNR, the weighting coefficient decrease toward to zero when η is increasing. In addition, the values of the all parameter are determined by experimental test. According the fact that the speech components almost focus in lower frequency band, let the sigmoid function with largest η (such as $\eta=20$) locate to highest frequency band (such as HH frequency part). On the contrary, let the sigmoid function with smallest η (such as $\eta=5$) locate to lowest frequency band (such as LL frequency part). So, the weighted HSELT measure is defined as below:

$$HSELT^{wef}(m,\xi_p)=HSELT(m,\xi_p)\times wef(m,\xi_p). \qquad (11)$$

The combined-MLSIE, which comprises four part-bands, is expressed as below:

$$HSELT^{wef}_{comb}(m)=\sum_{\xi_p=LL}^{HH} HSELT^{wef}(m,\xi_p). \qquad (12)$$

It is found that each HSELT feature parameter accurately indicates the boundary of speech activity under -5dB factory noise, especially at transition between speech and non-speech segments. Summing the four HSELT as a combined HSELT, we can determine an accuracy detection result.

D. The VAD Decision

Then, the voice activity is defined by the decision rules as shown below:

ISBN: 978-988-19252-6-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\begin{aligned} &\text{if } (HSELT_{comb}^{wef}(m) > Th) \\ &\quad VAD(m)=1; \\ &\quad \text{update } Th; \\ &\text{else} \\ &\quad VAD(m)=0; \end{aligned} \tag{13}$$

The threshold value Th is updated by recursive equation.

III. Experimental Results

To evaluate the advantages of the proposed HSELT feature sets for speech detection, we used a set of 12 sentences (about 107 seconds) from 4 different speakers: two males and females from TIMIT database. The utterances as speech or non-speech frames are corrupted by four different types of background noise: white noise, factory noise, car noise and babble noise at different average SNR levels between clean and -5dB (from NOISEX database).

All signals in the database were downsampled to 8-kHz, mono-channel and 16-bits per sample. These experiments were analyzed using speech pause hit-rate (HR0) and the speech hit-rate (HR1) (i.e., the fraction of all actual pause or speech frames, respectively). The proposed HSELT VAD is compared in terms of the average hit-rates (with optimal parameters) to state-of-the-art VAD methods, such as G.729B [7], AMR1 [13], AMR2 [8], ETSI AFE [12] and LTSD [9]. The optimal parameters for the proposed VAD were: $\eta_{HH}=5$, $\eta_{HL}=10$, $\eta_{HL}=15$, $\eta_{HH}=20$, and $R=5$, while the filter bank decomposed the signal in $\xi_{num}=17$ for Mel-scaled subband.

In order to quantify the speech/non-speech hit rates, we use the error norm of false alarm rates defined as:

$$E_{norm}=\sqrt{(1-HR1)^2+(1-HR0)^2} \tag{14}$$

Finally, we present an average speech/non-speech hit rates and overall false error norm for SNRs from clean to -5dB in Table I. It is found that the average value for HR1 of LTSD VAD is only comparable to the proposed HSELT VAD. The other VADs are inferior to the proposed HSELT VAD. In terms of HR0, the HR0 of proposed HSELT VAD is obviously superior to other VADs. So, the proposed HSELT achieved the minimum false alarm error norm, with a 40.90% value.

Table I
Average speech/non-speech hit rates and overall false error norm for SNRs from clean to -5dB

	Proposed	G.729	AMR1	AMR2	AFE	LTSD
HR1(%)	*94.60%*	88.50 %	94.20 %	89.50 %	92.50 %	95.70 %
HR0(%)	*59.40%*	34.20 %	36.10 %	44.10 %	43.30 %	45.90 %
Error norm(%)	*40.90%*	66.70 %	64.10 %	56.70 %	57.20 %	54.3%

IV. Conclusion

Since the conventional VAD algorithm could not deal with the unknown noises under in low SNR environments, we proposed a novel voice activity detection based on HSELT to improve the drawback. HSELT VAD is composed of four components: mel-scaled filter bank, HSELT feature extraction, part-band weighting estimation, and the VAD decision. It is found that the proposed method use HSELT feature sets can increase accuracy of ASR in mobile communication corrupted by unknown noises. The proposed HSELT-based VAD method is evaluate at eleven types of noises and five types of signal to SNR conditions. We find that the accuracy of the proposed HSELT-based VAD scheme averaged over all noise and all SNRs is better than that other considered VAD when the error norm of false alarm rates is 40.9%. Experiments in a mobile environment showed the proposed HSELT method obtain the best behavior in detecting non-speech with a 59.40% HR0 average value. In addition, the proposed VAD also attains a 94.60% HR1 average value in speech detection.

Acknowledgment

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC 101-2221-E-158-005.

References

[1] P. Renevey and A. Drygajlo, “Entropy Based Voice Activity Detection in Very Noisy Conditions,” in *Proc EUROSPEECH2001*, pp.1887-1890, September 2001.

[2] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing,” in *Proc. ICASSP*, Apr. 2008, pp. 4897–4900.

[3] S. G. Tanyer and H. Ozer, “Voice activity detection in nonstationary noise,” *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, July 2000.

[4] D. O’Shaughnessy, *Speech Communication*. Reading, MA: Addision-Wesley, 1987, p. 150.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process*, vol. 9, no. 5, pp. 504-512, 2001.

[6] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Proc. Eurospeech 2*, pp. 1513-1516, 1995.

[7] A. Benyassine, E. Shlomot, and H. Su, “ITU-T recommendation G.729, annex B, a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data spplications,” *IEEE Commun. Mag.*, pp.64–72, Sept. 1997.

[8] “Digital cellular telecommunications system (Phase 2+); Adaptive multi rate (AMR) speech; ANSI-C code for AMR speech codec,” 1998.

[9] J. Ramirez, J. C. Segura, C. Benitez, A. Torre, and A. Rubio, ¡§Efficient voice activity detection algorithms using long-term speech information,” *Speech Commun.*, vol. 42, pp. 271¡V287, 2004.

[10] G. Evangelopoulos and P. Maragos, “Multiband Modulation Energy Tracking for Noisy Speech Detection,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, November 2006.

[11] P.K. Ghosh, A. Tsiartas and S. Narayanan, "Robust Voice Activity Detection Using Long-Term Signal Variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, March 2011.

[12] Speech Processing, Transmission, and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms, 2002. ETSI, ETSI ES 201 108 Recommend.

[13] Cho, Y.D., Kondoz, A., “Analysis and improvement of a statistical model-based voice activity detector,” *IEEE Signal Process. Lett.*, vol.8, no.10, pp.276–278, 2001.