**Summary** This letter introduces 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. Since 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 the 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.

2. The Proposed VAD Method

The block diagram of the implemented system for VAD using HSELT measure is shown in Fig. 1 and is introduced in next subsection.

2.1 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, such as the Mel scale. Equation (1) is the Hz to Mel warping used in the experiments [4]:

\[
\text{mel} = 2595 \cdot \log(1 + f/700)
\]

where \( \text{mel} \) 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 ≤ \( \xi \) ≤ 17, 0 ≤ \( k \) ≤ 127), over a frequency range of 0–4 KHz. 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, \( x_{\text{time}}(m, n) \), representing the magnitude of the \( m \)th point of the \( n \)th frame.

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

\[
x_{\text{freq}}(m, k) = \sum_{n=0}^{N-1} x_{\text{time}}(m, n) \cdot \exp(-j2\pi/N)^{kn},
\]

where \( x_{\text{freq}}(m, k) \) is the magnitude of the \( k \)th point of the spectrum of the \( m \)th frame. \( N \) is frame length and is 256 in our system), and \( M \) is the number of frames of the speech signal for analysis.

The spectrum \( x_{\text{freq}}(m, k) \) 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 \( \xi \) of the \( m \)th frame.

---

Kun-Ching WANG†, Member

---

1. Introduction

In mobile or portable environments, the VAD mechanism has to distinguish active speech from noise with low signal to noise ratio (SNR). Most 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 accurately distinguish speech from noise due to the colored background noise statistics are stationary over a longer period than those of noise. In general, no particular feature or specific set of features has been shown to perform accurately distinguish speech from noise due to the colored background noise statistics are stationary over a longer period than those of noise. In general, no particular feature or specific set of features has been shown to perform accurately distinguish speech from noise due to the colored background noise statistics are stationary over a longer period than those of noise.

Due to the fact that the HSELT measure can be used to discriminate noise from noisy speech signal, it 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 Mel) part, a low frequency (9–12 Mel) part, a high frequency (13–15 Mel) part and a highest frequency (16–17 Mel) 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 the 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.
where $f(\xi, k)$ also represents the weighting factor of the frequency energy at the $k$th point of the $\xi$th 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, $\tilde{x}(m, \xi)$

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

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

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

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

2.2 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

$$x(m, \xi) = \sum_{k=0}^{N-1} |x_{freq}(m, k)| \cdot f(\xi, k) \quad (3) \quad 0 \leq m \leq M - 1; \quad 1 \leq \xi \leq 17$$
is computed using the last \( R \) frame of the observed signal \( x(n) \) with respect to the current frame of interest. The HSELT, \( \text{HSELT}(m, \xi) \), at frequency subband \( \xi \) for the \( m \)th frame is computed as follows:

\[
\text{HSELT}(m, \xi) = \frac{1}{K_{\xi}} \sum_{\xi \in \xi_p} \text{HSELT}(m, \xi_p) \quad (7)
\]

(6)

where \( \text{HSELT}(m, \xi) \) is essentially an entropy measure on the normalized short-time spectrum computed at frequency subband \( \xi \) over \( R \) consecutive frames, ending at the \( m \)th 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 frames for specific frequency subband. Observing the Fig. 4, 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: 0–1 kHz (LL part band: 1–8 Mel), 1–2 kHz (LH part band: 9–12 Mel), 2–3 kHz (HL part band: 13–15 Mel) and 3–4 kHz (HH part band: 16–17 Mel). Consequently, the HSELT parameter at \( \xi_p \)th part band is computed as below:

\[
\text{HSELT}(m, \xi_p) = \frac{1}{K_{\xi_p}} \sum_{\xi \in \xi_p} \frac{X(n, \xi) - \sum_{l=m-R+1}^{m} X(l, \xi)}{X(n, \xi) - \sum_{l=m-R+1}^{m} X(l, \xi)} \times \log \left( \frac{X(n, \xi) - \sum_{l=m-R+1}^{m} X(l, \xi)}{X(n, \xi) - \sum_{l=m-R+1}^{m} X(l, \xi)} \right) \quad (6)
\]

where \( X(n, \xi) \) is the power spectrum magnitude at the \( \xi \)th frequency subband for the \( n \)th frame.

In Fig. 2, the view of HSELT measure: (a) Power spectrum. (b) Spectrogram.

In Fig. 3, the degrees of nonstationary of the non-speech frame and speech frame.

In Fig. 4, the transition between speech and non-speech enhanced by the degree of stationarity: (a) The waveform of speech. (b) The corresponding spectrogram. (c) The degree of nonstationary.
method[6] is used here, which is not constrained by any window length to update noise spectrum estimate.

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

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

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

\[
\text{wef}(m, \xi_p) = \frac{1}{1 + \exp \left[ -0.5 \cdot (\text{SNR}(m, \xi_p) - \eta(m, \xi_p)) \right]} 
\]
(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. Figure 5 shows the plots of the weighting coefficients from Eq. (10) depending on \( \eta \). Under the fixed value of a posterior SNR, the weighting coefficient decrease toward to zero when \( \eta \) 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 \( \eta \) (such as \( \eta = 20 \)) locate to highest frequency band (such as HH frequency part). On the contrary, let the sigmoid function with smallest \( \eta \) (such as \( \eta = 5 \)) locate to lowest frequency band (such as LL frequency part). So, the weighted HSELT measure is defined as below:

\[
\text{HSELT}_{\text{wef}}^\text{comb}(m) = \text{HSELT}(m, \xi_p) \times \text{wef}(m, \xi_p). 
\]
(11)
The combined-MLSIE, which comprises four part-bands, is expressed as below:

\[
\text{HSELT}_{\text{wef}}^\text{comb}(m) = \sum_{\xi_p=LL}^{HH} \text{HSELT}_{\text{wef}}(m, \xi_p). 
\]
(12)

It is found that each HSELT feature parameter accurately indicates the boundary of speech activity under \(-5 \text{dB} \) 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.

### 2.4 The VAD Decision

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

\[
\text{if } (\text{HSELT}_{\text{wef}}^\text{comb}(m) > Th) \\
\quad VAD(m) = 1; \\
\quad \text{update } Th; \\
\quad \text{else} \\
\quad VAD(m) = 0; 
\]
(13)

where \( Th \) mean the speech thresholds and can be recursively updated by using the mean and variance of the logarithmic combined HSELT to estimate the time-varying noise characteristics [14]. In fact, we assume that the first four frames contain noise only and then compute the initial noise mean and variance with the first five frames.

The scheme of adaptive threshold for the speech and noise can be computed by the following:

\[
Th = \mu_N + \alpha \cdot \sigma_N 
\]
(14)

Similarly, \( \mu_N \) and \( \sigma_N \) represent the mean and the variance of the logarithmic combined HSELT, respectively. In addition, \( \alpha_S \) is the adjustment constant which is used to determine the threshold value.

The mean and variance of the logarithmic combined HSELT are updated while the decision rule shows a noise period.

\[
\mu_N(m) = \gamma \cdot \mu_N(m-1) + (1 - \gamma) \cdot \text{HSELT}_{\text{wef}}^\text{comb}(m) \\
\frac{[\text{HSELT}_{\text{wef}}^\text{comb}]^2_{\text{mean}}(m)}{\sigma_N(m)} = \gamma \cdot \frac{[\text{HSELT}_{\text{wef}}^\text{comb}]^2_{\text{mean}}(m-1)}{\sigma_N(m)} + (1 - \gamma) \cdot \frac{[\text{HSELT}_{\text{wef}}^\text{comb}]^2_{\text{mean}}(m)}{\sigma_N(m)} 
\]
(15)

where \( \gamma = 0.5 \) is chosen by experiment. We then update the threshold using the updated mean and variance of the logarithmic combined HSELT.

### 3. 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 speaker; 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 \(-5 \text{dB} \) (from NOISEX database). All signals in the database were downsampled to 8-kHz, mono-channel
Table 1 Non-speech hit rate (HR0) from clean to −5 dB under the four types of noise.

| Noise type | SNR(dB) | Proposed | G729B | AMR-1 | AMR-2 | AFE | LTSD |
|------------|---------|----------|-------|-------|-------|-----|------|
|            |         | HR1 | HR0 | HR1 | HR0 | HR1 | HR0 | HR1 | HR0 | HR1 | HR0 |
| White noise | clean | 95.3 | 68.9 | 95.5 | 63.3 | 96.6 | 63.8 | 96.8 | 63.8 | 98.4 | 63.3 |
|            | −5     | 88.6 | 63.2 | 95.5 | 63.3 | 96.6 | 63.8 | 96.8 | 63.8 | 98.4 | 63.3 |
| Factory noise | clean | 99.1 | 78.4 | 96.1 | 70.7 | 99.3 | 72.9 | 97.6 | 64.6 | 96.3 | 64.2 |
|            | −5     | 99.1 | 78.4 | 96.1 | 70.7 | 99.3 | 72.9 | 97.6 | 64.6 | 96.3 | 64.2 |
| Car noise | clean | 99.2 | 81.1 | 97.4 | 72.6 | 99.5 | 72.3 | 98.4 | 66.7 | 96.6 | 68.9 |
|            | −5     | 97.8 | 59.4 | 99.7 | 52.2 | 96.8 | 43.2 | 94.7 | 38.3 | 95.2 | 35.7 |
| Roadable noise | clean | 98.6 | 71.8 | 93.3 | 41.1 | 98.7 | 45.6 | 94.2 | 37.6 | 94.3 | 38.3 |
|            | −5     | 94.3 | 43.4 | 91.2 | 24.1 | 95.5 | 25.7 | 89.5 | 35.3 | 93.7 | 32.7 |
| Average (%) |        | 94.6 | 55.4 | 88.5 | 34.2 | 94.2 | 36.1 | 89.5 | 44.1 | 92.5 | 43.3 |
| Error norm (%) |        | 40.9 | 66.7 | 64.1 | 56.7 | 57.2 | 54.3 |

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:

$$E_{norm} = \sqrt{(1 - HR1)^2 + (1 - HR0)^2} \quad (16)$$

Table 1 shows the HR0 and HR1 across SNR level from clean to −5 dB under four types of background noise. In addition, we present an average speech/non-speech hit rates and overall false error norm for SNRs from clean to −5 dB in the bottom of Table 1. 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.

4. Conclusion

Since the conventional VAD algorithm can not deal with unknown noises under in low SNR environments, we proposed a novel voice activity detection algorithm based on HSELT to overcome 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 evaluated at four types of noises under SNR levels between clean and −5 dB. We find that the accuracy of the proposed HSELT-based VAD scheme averaged over all noise and all SNRs is better than that the other VAD schemes considered 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.

Acknowledgments

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC 102-2221-E-158-006.

References

[1] P. Renevey and A. Drygajlo, “Entropy based voice activity detection in very noisy conditions,” Proc. EUROSEPHEE2001, pp.1887–1890, Sept. 2001.
[2] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing,” Proc. ICASSP, pp.4897–4900, April 2008.
[3] S.G. Tanyer and H. Ozer, “Voice activity detection in nonstationary noise,” IEEE Trans. Speech Audio Process., vol.8, no.4, pp.478–482, July 2000.
[4] D. O’Shaughnessy, Speech Communication. Reading, Addison-Wesley, MA, 1987.
[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 applications,” IEEE Commun. Mag., vol.35, no.9, 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.271–287, 2004.
[10] G. Evangelopoulos and P. Maragos, “Multiband modulation energy tracking for noisy speech detection,” IEEE Trans. Audio, Speech, Language Process., vol.14, no.6, pp.2024–2038, Nov. 2006.
[11] P.K. Ghosh, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” IEEE Trans. Audio, Speech, Language Process., vol.19, no.3, pp.600–613, 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] Y.D. Cho and A. Kondoz, “Analysis and improvement of a statistical model-based voice activity detector,” IEEE Signal Process. Lett., vol.8, no.10, pp.276–278, 2001.
[14] S.V. Gerven and F. Xie, “A comparative study of speech detection methods,” Proc. Eurospeech, vol.3, pp.1095–1098, 1997.