*Sensors* **2013**, *13*, 16533-16550; doi:10.3390/s131216533

OPEN ACCESS

**sensors**

**ISSN 1424-8220**

www.mdpi.com/journal/sensors

*Article*

# A Novel Voice Sensor for the Detection of Speech Signals

**Kun-Ching Wang**

Department of Information Technology & Communication, Shih Chien University,
200 University Road, Neimen, Kaohsiung 84550, Taiwan; E-Mail: kunching@mail.kh.usc.edu.tw;
Tel.: +886-07-667-8888-5723; Fax: +886-07-667-8888-4332

*Received: 3 November 2013; in revised form: 26 November 2013 / Accepted: 27 November 2013 / Published: 2 December 2013*

**Abstract:** In order to develop a novel voice sensor to detect human voices, the use of features which are more robust to noise is an important issue. Voice sensor is also called voice activity detection (VAD). Due to that the inherent nature of the formant structure only occurred on the speech spectrogram (well-known as voiceprint), Wu *et al.* were the first to use band-spectral entropy (BSE) to describe the characteristics of voiceprints. However, the performance of VAD based on BSE feature was degraded in colored noise (or voiceprint-like noise) environments. In order to solve this problem, we propose the two-dimensional part-band energy entropy (TD-PBEE) parameter based on two variables: part-band partition number upon frequency index and long-term window size upon time index to further improve the BSE-based VAD algorithm. The two variables can efficiently represent the characteristics of voiceprints on each critical frequency band and use long-term information for noisy speech spectrograms, respectively. The TD-PBEE parameter can be regarded as a PBEE parameter over time. First, the strength of voiceprints can be partly enhanced by using four entropies applied to four part-bands. We can use the four part-band energy entropies for describing the voiceprints in detail. Due to the characteristics of non-stationary for speech and various noises, we will then use long-term information processing to refine the PBEE, so the voice-like noise can be distinguished from noisy speech through the concept of PBEE with long-term information. Our experiments show that the proposed feature extraction with the TD-PBEE parameter is quite insensitive to background noise. The proposed TD-PBEE-based VAD algorithm is evaluated for four types of noises and five signal-to-noise ratio (SNR) levels. We find that the accuracy of the proposed TD-PBEE-based VAD algorithm averaged over all noises and all SNR levels is better than that of other considered VAD algorithms.

**Keywords:** voice sensor; two-dimensional; part-band energy entropy; long-term information analysis; Mel-scaled filter bank

## 1. Introduction

So far, user-friendly voice interfaces have been widely used in consumer devices, such as interactive digital TV, personal digital assistants and cellular phones [1–3]. Voice sensor (also called voice activity detection, VAD) refers to the problem of distinguishing speech from non-speech regions. It is found that VAD is a critical component in voice-command application. However, the use of features which are more robust to noise is an important issue. Various types of different approaches to VAD have been proposed recently. In early VAD algorithm designs, short-term energy, zero-crossing rate and LPC coefficients [4] were used as feature parameters for detecting voices. In addition, some algorithms further used cepstral features [5], formant shape [6], and least-square periodicity measures [7]. Others have used correlation coefficients [8], wavelet coefficients [9], entropy measures [10], and a set of metrics [11]. Remirez *et al.* recently formulated long-term spectral divergence (LTSD) between speech and non-speech segments as a discriminative speech feature [12]. Ma *et al.* further proposed a long-term spectral flatness measure (LSFM) to improve speech detection robustness for lower SNR [13]. More complex algorithms use statistical model-based features [14,15], which have decision rules derived from the likelihood ratio test.

In fact, a robust VAD algorithm in the presence of different types of noises is necessary and critical. Depending on the characteristics of the human voice, a variety of parameters has been proposed for VAD. 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 at low SNR [16]. Similarly, entropy measures fail to distinguish speech from noise with good accuracy due to the colored spectrum of speech [17]. SNR estimation is also a critical component in many of the existing VAD schemes, which is particularly difficult for non-stationary noise [18]. The use of features which are more robust to noise is an important issue for develop a robust VAD algorithm. Due to the fact that the inherent nature of the formant structure only occurred on speech spectrograms and is the well-known as the "voiceprint", Wu *et al.* were the first to use band-spectral entropy (BSE) to describe the characteristics of voiceprints [19]. However, the performance of BSE-based features for VAD was degraded under colored noise environment conditions.

In order to solve this problem, we propose a two-dimensional part-band energy entropy (TD-PBEE) method in this paper to improve the robustness of the proposed VAD method in colored noisy environments. The TD-PBEE parameter can be regarded as the relation of spectral entropy *versus* time index. In summary, the TD-PBEE is based on two variables: part-band number ($N$) upon frequency index and long-term size ($R$) upon time index. First, the four part-bands (the optimal is $N = 4$) derived from 17 log-energies by a Mel-scaled filter bank are partitioned as 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. Consequently, the strength of voiceprints can be more enhanced by four PBEEs than that by BSE. Secondly, we will use long-term information processing to refine the PBEE due to the non-stationary characteristics of speech and various noises. Each part-band has different long-term window *R* sizes. Through different $R$ values, the TD-PBEE dependent on each part-band will be determined to efficiently represent the voiceprint characteristics in each critical frequency band. Consequently, the voice-like noise can be distinguished from noisy speech through the concept of PBEE with long-term information. Our experiments show that the proposed feature extraction of TD-PBEE is quite insensitive to background noise. The proposed TD-PBEE-based VAD scheme is evaluated for four types of noises and five signal-to-noise ratio (SNR) levels. We find that the accuracy of the proposed TD-PBEE-based VAD method averaged over all noise and all SNR levels is better than that of other considered VAD algorithms.

The remainder of this paper is organized as follows: in Section 2, the procedure of determining the TD-PBEE parameter is described. In Section 3, the proposed VAD based on TD-PBEE is schematically introduced. In Section 4, experimental results demonstrate the effectiveness of the proposed TD-PBEE VAD method. Finally, Section 5 concludes the paper.

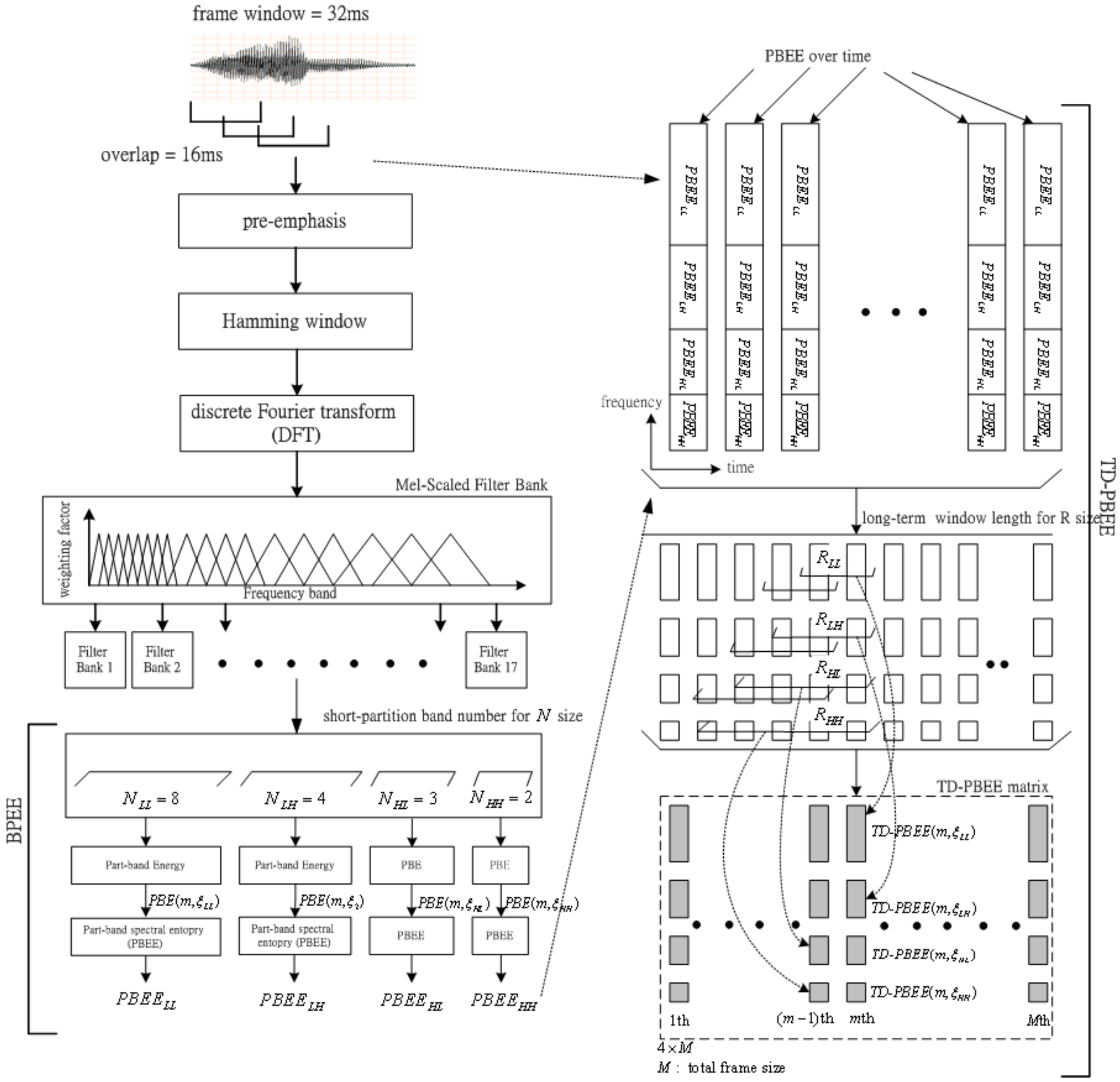
## 2. The Proposed Two-Dimensional Part-Band Energy Entropy (TD-PBEE) Measure

According to the findings from [18], Wu *et al.* were the first to use BSE to describe the voiceprint characteristics of speech-only spectrograms. It is found that the BSE can detect the human-voice signals. In this subsection, we further improve the BSE and propose a novel feature extraction of the TD-PBEE parameter. The definition of the TD-PBEE will be shown in detail. Figure 1 shows the procedure of feature extraction of TD-PBEE. Observing Figure 1, we can find the procedure of the TD-PBEE is based on (*R*, *N*). The input speech signal is frame windowed (32-ms length and 16-ms shift) using the Hamming window. In order to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing, the digitized speech signal is first put through a first-order pre-emphasis filter with pre-emphasis coefficient 0.97:

$$H(z) = 1 - 0.97z^{-1} \tag{1}$$

After the pre-emphasis process, a speech signal is divided into frames by multiplying a Hamming window. In order to avoid sharp changes, we make the windows overlap with each other. Hence, the utterance is segmented into a sequence of overlapped frames. Secondly, a Discrete Fourier Transform (DFT) is applied to obtain the short time spectrum of each frame. We then multiply the spectrum by the common Mel-scale filter bank weighting factors and compute the energy of each frequency band. We generate the output energy of each filter of the 17-channel Mel-scale filter bank. Then, the short-partition band number, *N*, is used in the paper. The value of *N* is four and comprises a set of $N_{LL}$, $N_{LH}$, $N_{HL}$ and $N_{HH}$ ($N = \{N_{LL}, N_{LH}, N_{HL}, N_{HH}\}$), so the four part-band energy (PBE) is denoted as $PBE(m,\xi_{LL})$, $PBE(m,\xi_{LH})$, $PBE(m,\xi_{HL})$ and $PBE(m,\xi_{HH})$. Each of the short-partition bands shows a lowest frequency (1–8 Mel) part ($N_{LL} = 8$), a low frequency (9–12 Mel) part ($N_{LH} = 4$), a high frequency (13–15 Mel) part ($N_{HL} = 3$) and a highest frequency (16–17 Mel) part ($N_{HH} = 2$). The strength of voiceprints can be partly enhanced by using four part-band energy entropies (PBEE) applied to the four part-bands. Consequently, the voiceprint-like noise will not be detected in noisy speech. The inherent characteristic of voiceprints can be better characterized by PBEE than by the BSE parameter.

**Figure 1.** The block diagram of the feature extraction for TD-PBEE measurements.



Finally, collecting a sequence of PBEE coefficients along the time axis, we can get a PBEE over time. Applying the long-term spectral information processing for R size, the value of each TD-PBEE is depended on different $R$: ($R_{LL}$, $R_{LH}$, $R_{HL}$ and $R_{HH}$). The value of $PBEE_{LL}$ over $R_{LL}$ consecutive frames is determined at the specific $LL$th part-band. Similarly, the value of $PBEE_{LH}$ is determined over $R_{LH}$ consecutive frames at the specific $LH$th part-band. The value of $PBEE_{HL}$ is determined over $R_{HL}$ consecutive frames at the specific $HL$th part-band. The value of $PBEE_{HH}$ is determined over $R_{HH}$ consecutive frames at the specific $HH$th part-band. Consequently, the TD-PBEE parameters are chosen from the set of $TD\text{-}PBEE_{LL}$, $TD\text{-}PBEE_{LH}$, $TD\text{-}PBEE_{HL}$ and $TD\text{-}PBEE_{HH}$ coefficients over long-term average processing. The TD-PBEE parameter can be regarded as the relation of spectral entropy *versus* time index, so we also call it the TD-PBEE matrix. In this section, we will first introduce the definition of the PBEE based on *N*. Then, the TD-PBEE based on *R* will be presented later.

### *2.1. Definition of the PBEE Based on* $N$

In order to further improve the advantage of characterizing voiceprints though band-spectral entropy (BSE), we adopt a novel concept of part-band spectral entropy (PBEE). This concept lets full-bands be partitioned into some little part-bands. Through spectral entropy determined from each part-band, the voiceprint can be more partially described.

Figure 2 shows the partition structure of the Mel-scaled filter bank. It is found that higher sub-band numbers are focused on the lower frequencies. Inversely, the lower sub-band numbers are focused on higher frequencies. Observing the Figure 2, each part-band has a different band number. Although many part-band numbers can clearly describe the voiceprint, this will need more computer power. In Table 1, we observe the fact that a higher number of part-band partitions can achieve higher VAD accuracy, but we need more computing time to run the VAD algorithm. Inversely, a lesser number of part-band partitions leads to lower VAD accuracy. Considering the trade-off between accuracy and real-time requirements, the number of part-band partitions, $N$ equal four is best compromise. The numbers of each part-band are $N_{LL}=8$, $N_{LH}=4$, $N_{HL}=3$ and $N_{HH}=2$, respectively. The four part-bands comprise 0~1 kHz (LL part-band $\xi_{LL}$: 1–8 Mel), 1~2 kHz (LH part- band $\xi_{LH}$: 9–12 Mel), 2~3 kHz (HL part-band $\xi_{HL}$: 13–15 Mel) and 3~4 kHz (HH part-band $\xi_{HH}$: 16–17 Mel). Consequently, the PBEE parameter at each part-band is computed as below:

$$PBEE(m,\xi_{LL})=-\sum_{\xi=1}^{N_{LL}}P_{LL}(m,\xi)\log P_{LL}(m,\xi) \tag{2}$$

$$PBEE(m,\xi_{LH})=-\sum_{\xi=N_{LL}+1}^{N_{LL}+N_{LH}}P_{LH}(m,\xi)\log P_{LH}(m,\xi) \tag{3}$$

$$PBEE(m,\xi_{HL})=-\sum_{\xi=N_{LH}+1}^{N_{LH}+N_{HL}}P_{HL}(m,\xi)\log P_{HL}(m,\xi) \tag{4}$$

$$PBEE(m,\xi_{HH})=-\sum_{\xi=N_{HL}+1}^{N_{HL}+N_{HH}}P_{HH}(m,\xi)\log P_{HH}(m,\xi) \tag{5}$$

where $P_{LL}(m,\xi)=E_b(m,\xi)\Big/\sum_{k=1}^{N_{LL}}E_b(m,k)$ is the probability corresponding to 17 band-energies. Band-energy for each frame is denoted as $E_b(m,\xi)=|X(m,\xi)|^2$.

Similar, the other probabilities are defined as below:

$$P_{LH}(m,\xi)=E_b(m,\xi)\Big/\sum_{k=N_{LL}+1}^{N_{LH}}E_b(m,k) \tag{6}$$

$$P_{HL}(m,\xi)=E_b(m,\xi)\Big/\sum_{k=N_{LH}+1}^{N_{HL}}E_b(m,k) \tag{7}$$

$$P_{HH}(m,\xi)=E_b(m,\xi)\Big/\sum_{k=N_{HL}+1}^{N_{HH}}E_b(m,k) \tag{8}$$

Figure 3 shows the four PBEE values determined from four part-bands. We can find that the PBEE value is dependent on the different frequency band numbers $N$. Due to the fact that the voiceprints mostly focus on middle or low frequency band, more band numbers are required. Inversely, less band numbers are assigned to the higher frequency band due to the fact that the higher band is almost always dominated by noise components.

**Table 1.** The number of part-band partitions related to the VAD accuracy and delay time.

| The Number of Part-Band Partitions $N$ | VAD Accuracy (%) | Delay Time (s) |
|---|---|---|
| 2 | 68.3% | 0.64 |
| ***4*** | ***82.3%*** | ***0.94*** |
| 6 | 82.5% | 1.24 |
| 8 | 83.4% | 1.96 |

**Figure 2.** The structure of a four part-band partition.

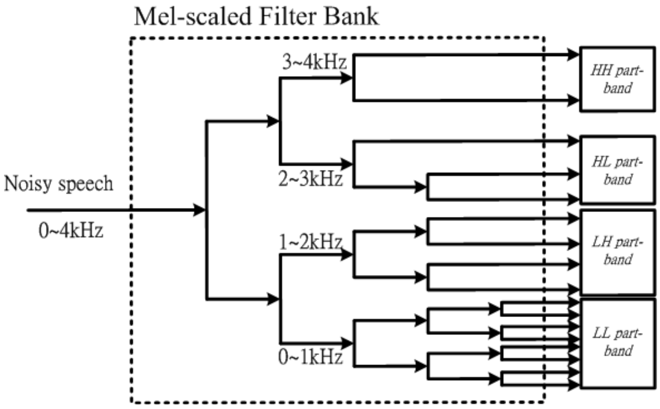


**Figure 3.** The block diagram of four PBEE values determined from four part-bands.

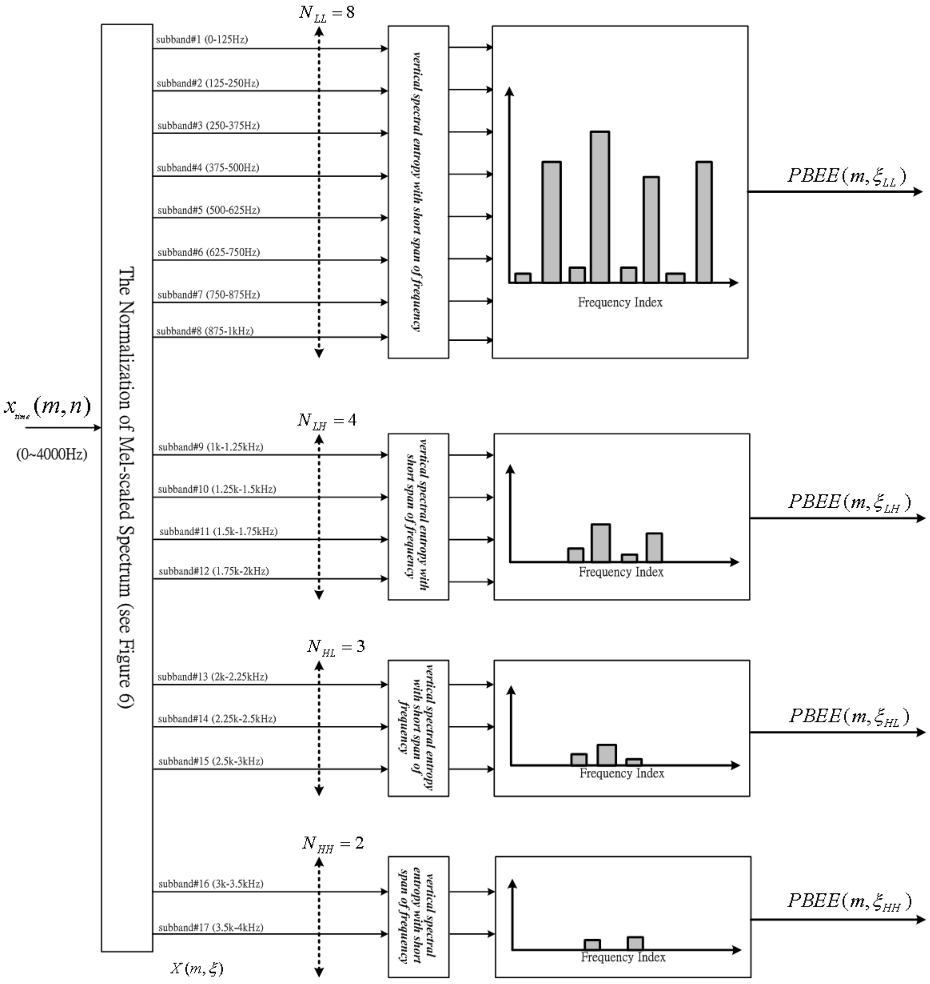
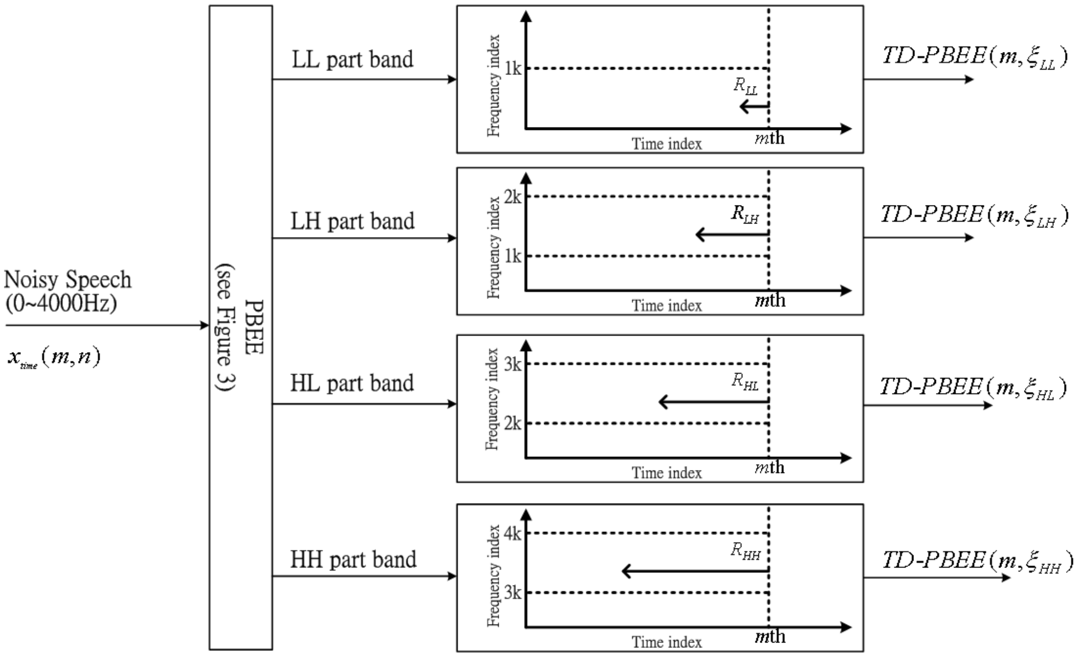


**Figure 4.** The block diagram of four TD-PBEE values determined from four PBEEs over time.



### *2.2. The TD-PBEE Based on* $R$

In order to further refine the PBEE parameter, long-term information processing is used to determine a reliable evaluation for the strength of voiceprint on part-band. In this subsection, each part-band has different long-term windows size corresponding to $R_{LL}$, $R_{LH}$, $R_{HL}$ and $R_{HH}$. Due to the fact that voiceprint-like noise can often focus on high frequency bands, a concept of long-term information is required, so the assumption is expressed as $R_{LL} < R_{LH} < R_{HL} < R_{HH}$ for four PBEE parameters of each part-band. In addition, this assumption also reduces the search time decreasing computing power for the low frequency band and increasing the accuracy of voiceprint evaluation for the entire speech signal.

Consequently, the definition of two dimensions for PEBB parameter means that the one dimension is the time index and the other dimension is the frequency index. The computation of the TD-PBEE is shown as below:

$$TD\text{-}PBEE(m,\xi_{LL})=\sum_{n=m-R_{LL}}^{m}PBEE(n,\xi_{LL})\Big/R_{LL} \tag{9}$$

$$TD\text{-}PBEE(m,\xi_{LH})=\sum_{n=m-R_{LH}}^{m}PBEE(n,\xi_{LH})\Big/R_{LH} \tag{10}$$

$$TD\text{-}PBEE(m,\xi_{HL})=\sum_{n=m-R_{HL}}^{m}PBEE(n,\xi_{HL})\Big/R_{HL} \tag{11}$$

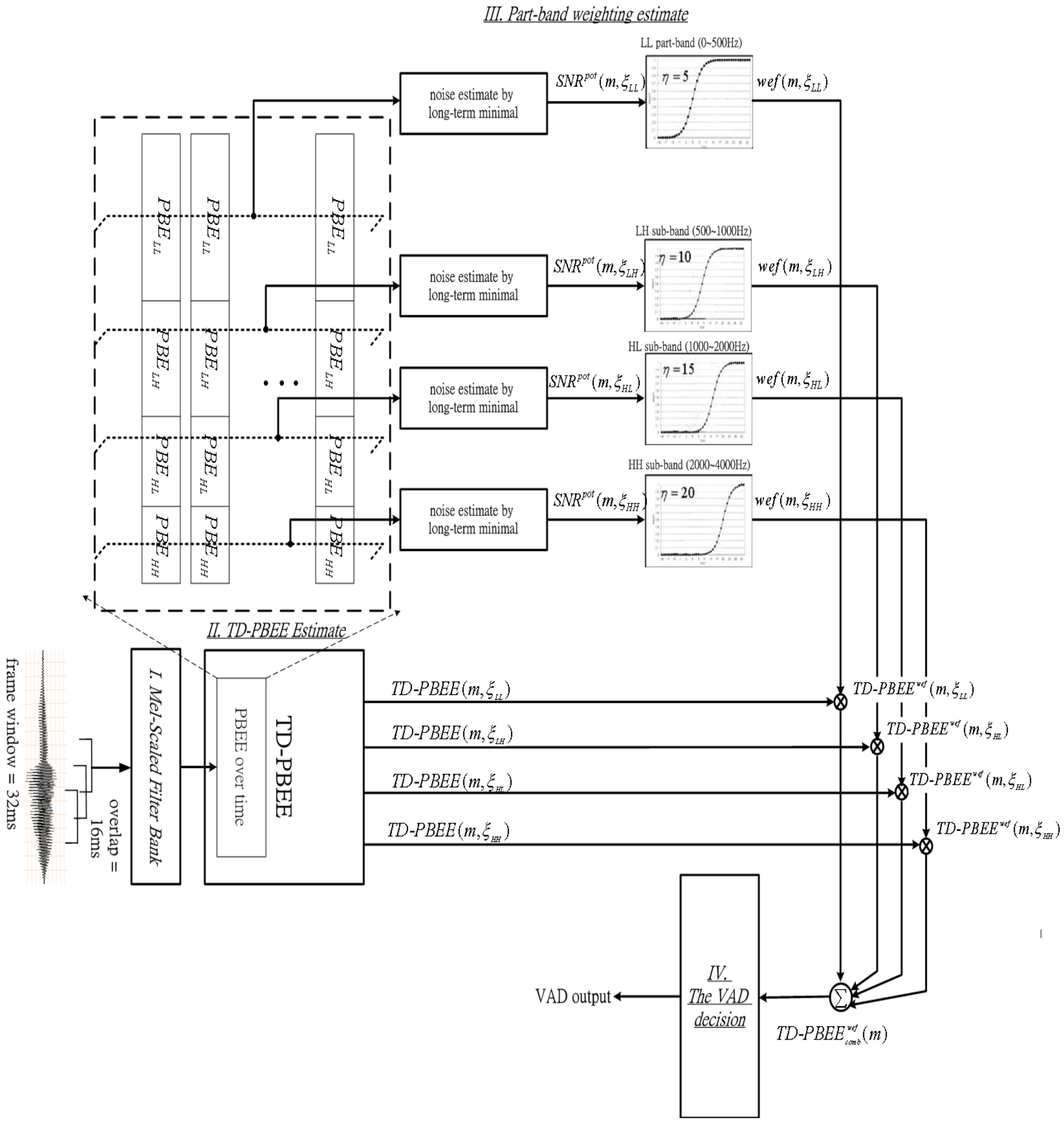
$$TD\text{-}PBEE(m,\xi_{HH})=\sum_{n=m-R_{HH}}^{m}PBEE(n,\xi_{HH})\Big/R_{HH} \tag{12}$$

From the above equation can be found that each TD-PBEE is averaged over the long-term window size. Figure 4 clearly shows the block diagram of four TD-PBEE values determined from four PBEEs over time derived from different long-term window sizes: $R_{LL}$, $R_{LH}$, $R_{HL}$ and $R_{HH}$.

## 3. The Proposed VAD Based on TD-PBEE Measure

In this section we propose the TD-PBEE based VAD algorithm as shown in Figure 5. The proposed TD-PBEE VAD method consists of four components: (1) Mel-scaled filter bank; (2) TD-PBEE estimate; (3) part-band weighting estimation; and (4) the VAD decision. TD-PBEE estimate has been introduced in Section 2. The remainder will be introduced in details as follows: first, the PBEE vector is applied to determine the part-band weighting estimate for suppressing voiceprints corrupted by noise. Secondly, we can use a part-band weighting estimate to adjust a robust TD-PBEE parameter. Finally, the VAD decision can adaptively judge whether the current frame is a noise-dominated frame or speech-dominated frame through a decision rules.

**Figure 5.** The block diagram of TD-PBEE based VAD algorithm.


III. Part-band weighting estimate
LL part-band (0~500Hz)
LH sub-band (500~1000Hz)
HL sub-band (1000~2000Hz)
HH sub-band (2000~4000Hz)
noise estimate by long-term minimal
$SNR^{pot}(m,\xi_{LL})$
$SNR^{pot}(m,\xi_{LH})$
$SNR^{pot}(m,\xi_{HL})$
$SNR^{pot}(m,\xi_{HH})$
$\eta = 5$
$\eta = 10$
$\eta = 15$
$\eta = 20$
$wef(m,\xi_{LL})$
$wef(m,\xi_{LH})$
$wef(m,\xi_{HL})$
$wef(m,\xi_{HH})$
$PBE_{LL}$
$PBE_{LH}$
$PBE_{HL}$
$PBE_{HH}$
II. TD-PBEE Estimate
TD-PBEE
PBEE over time
I. Mel-Scaled Filter Bank
frame window = 32ms
overlap = 16ms
$TD\text{-}PBEE(m,\xi_{LL})$
$TD\text{-}PBEE(m,\xi_{LH})$
$TD\text{-}PBEE(m,\xi_{HL})$
$TD\text{-}PBEE(m,\xi_{HH})$
$TD\text{-}PBEE^{wef}(m,\xi_{LL})$
$TD\text{-}PBEE^{wef}(m,\xi_{HL})$
$TD\text{-}PBEE^{wef}(m,\xi_{HL})$
$TD\text{-}PBEE^{wef}(m,\xi_{HH})$
$TD\text{-}PBEE^{wef}_{comb}(m)$
IV. The VAD decision
VAD output

### *3.1. The Normalization of Mel-Scale Filter Bank*

Figure 6 shows in detail the process of including the Mel-scale bank for getting the normalized energy. The Mel-scale first suggested by Stevens and Volkman in 1937 [20] is a perceptually motivated scale. The scale was devised through human perception experiments where subjects were were asked to adjust a stimulus tone to perceptually half the pitch of a reference tone. Equation (1) is the Hz to Mel warping used in the experiments [21]:

$$Mel = 2595 \cdot \log(1 + f/700) \tag{13}$$

where $Mel$ is the Mel-frequency scale and $f$ is in Hertz. The filter banks 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 0–4 KHz. The energy of each frequency band for each time frame of a speech signal can be calculated through the Mel-scale frequency bank:

The spectrum, $x_{freq}(m,k)$, of this signal is first calculated by a Discrete Fourier Transform (256-point DFT), while considering a given time-domain noisy speech signal, $x_{time}(m,n)$, representing the magnitude of the $n$th point of the $m$th frame:

$$x_{freq}(m,k) = \sum_{n=0}^{N-1} x_{time}(m,n) \cdot \exp(-j2\pi/N)^{kn}, \\ 0 \le k \le N-1;\ 0 \le m \le M-1 \tag{14}$$

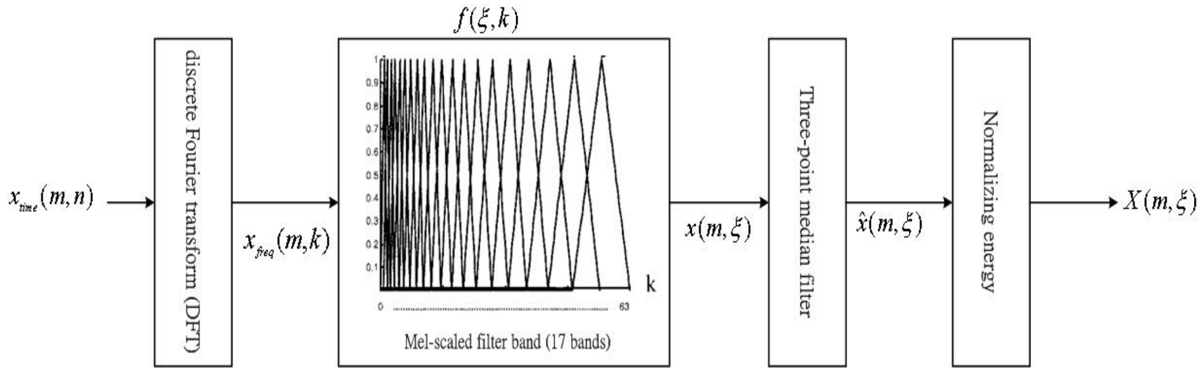
where $x_{freq}(m,k)$ means the magnitude of the $k$th point of the spectrum of the $m$th frame, and $M$ is the number of total frames of the speech signal for analysis.

First, the spectrum $x_{freq}(m,k)$ is then multiplied by the weighting factors $f(\xi,k)$ on the Mel-scale frequency bank. Then, 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:

$$x(m,\xi) = \sum_{k=0}^{N-1} \left| x_{freq}(m,k) \right| \cdot f(\xi,k) \\ 0 \le m \le M;\ 1 \le \xi \le 17 \tag{15}$$

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

**Figure 6.** The processing of the Mel-scale filter bank.



Some undesired impulse noise is resulted from our experiments that the energy $x(m,\xi)$ obtained in Equation (15). Hence, a three-point median filter is further used to get the smoothed energy, $\hat{x}(m,\xi)$:

$$\hat{x}(m,\xi)=\left[x(m-1,\xi)+x(m,\xi)+x(m+1,\xi)\right]/3. \tag{16}$$

In fact, the noise can focus the same as speech. Based on these finds, we can remove the frequency energy of the beginning interval from the smoothed energy, $\hat{x}(m,\xi)$, to get the pure energy, $X(m,\xi)$:

$$X(m,\xi)=\hat{x}(m,\xi)-\sum_{j=0}^{4}\hat{x}(j,\xi)\Big/5 \tag{17}$$

where $\sum_{j=0}^{4}\hat{x}(j,\xi)\Big/5$ means the frequency energy of the beginning interval estimated by averaging the frequency energy of the first five frames of the recording.

### *3.2. Part-Band Weighting Estimation*

We need a parameter will help us know how much the current part-band is corrupted by noise due to the influence of noise upon the detection performance. A posterior part-band SNR, $SNR^{pot}(m,\xi_p)$ is required in order to determine the part-band utility rate on $\xi_p$ part for $m$th frame, and it is formulated as:

$$SNR^{pot}(m,\xi_p)=10\cdot\log_{10}\left[PBE_{N+S}(m,\xi_p)/PBE_N(m,\xi_p)\right] \tag{18}$$

where $PBE_{N+S}(m,\xi_p)$ means the part-band energy (PBE) range from on $\xi_p$th part for $m$th frame for the observed noisy speech signal. $PBE_N(m,\xi_p)$ is the estimated noise part-band energy.

According to Equation (18), we know that the estimated noise part-band energy, $PBE_N(m,\xi_p)$, has to be estimated while determining the value of a posterior SNR, $SNR^{pot}(m,\xi_p)$. In order to estimate the noise-level quickly and accurately, the method tracking the minimum of the noisy speech power spectrum energy over a fixed search window length was proposed [22]. We use an efficient method [23] to speed up the determination of local minimum of noisy speech spectrum over a search window size, which is not constrained by any window length to update noise spectrum estimate, and it is calculated as below:

$$\begin{aligned}
&\text{If}\quad PBE_{\min}(m-1,\xi_p)<PBE_{N+S}(m,\xi_p),\\
&\text{then}\quad PBE_{\min}(m,\xi_p)=\gamma\cdot PBE_{\min}(m-1,\xi_p)+\frac{1-\gamma}{1-\beta}\left[PBE_{N+S}(m,\xi_p)-\beta\cdot PBE_{N+S}(m-1,\xi_p)\right],\\
&\text{else}\quad PBE_{\min}(m,\xi_p)=PBE_{N+S}(m,\xi_p).
\end{aligned} \tag{19}$$

where $P_{\min}(m,\xi_p)$ denotes the local minimum of power energy of the noisy speech, and it stands for noise part-band energy. $\beta$ and $\gamma$ are experimentally determined constants.

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

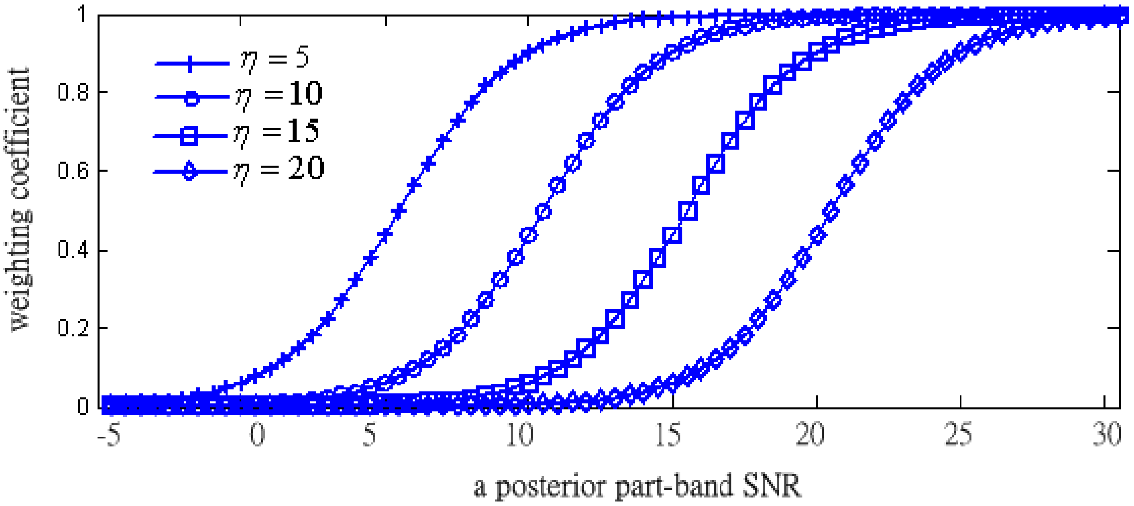
$$wef(m,\xi_p)=\frac{1}{1+\exp\left[-0.5\cdot\left(SNR^{pot}(m,\xi_p)-\eta(m,\xi_p)\right)\right]} \tag{20}$$

where $\eta(m,\xi_p)$ is a center-offset of the sigmoid function, and it is depended on part-band index.

Observing Equation (20), we will use this information to weight each part-band if the *a posteriori* SNR and a center-offset of the sigmoid function are known.

Figure 7 shows the plots of the weighting coefficients from Equation (20) depending on $\eta$. Under the fixed value of a posterior SNR, the weighting coefficients decrease towards zero when $\eta$ is increasing. In addition, the values of all the parameters are determined by experimental tests. According to the fact that the speech components are almost focused in the lower frequency band, we let the sigmoid function with largest $\eta$ (such as $\eta = 20$) locate to the highest frequency band (such as the HH*th* frequency part). On the contrary, we let the sigmoid function with the smallest $\eta$ (such as $\eta = 5$) locate to the lowest frequency band (such as LL*th* frequency part).

**Figure 7.** The plots of weight coefficients against a posterior part-band SNR under variable $\eta$.



Consequently, the TD-PBEE parameter can be further weighted and be shown as below:

$$TD\text{-}PBEE^{wef}(m,\xi_p) = TD\text{-}PBEE(m,\xi_p) \times wef(m,\xi_p) \tag{21}$$

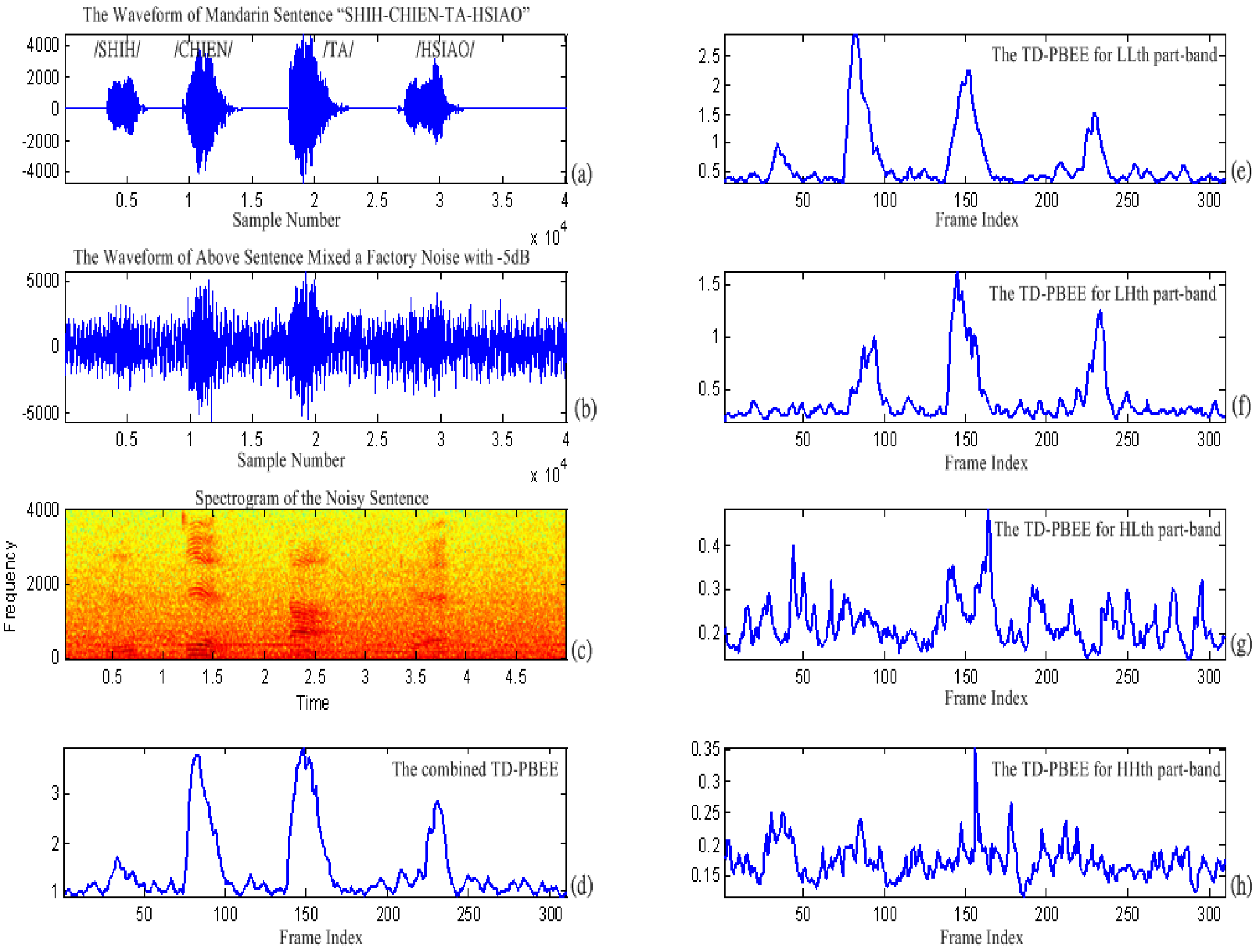
where $TD\text{-}PBEE^{wef}(m,\xi_p)$ denotes the weighted TD-PBEE parameter.

Thus, summing the four TD-PBEEs from each part-band as a combined TD-PBEE, the combined TD-PBEE is expressed as below:

$$TD\text{-}PBEE^{wef}_{comb}(m) = \sum_{\xi_p = LL}^{HH} TD\text{-}PBEE^{wef}(m,\xi_p) \tag{22}$$

Figure 8 shows the results of the combined TD-PBEE compared with TD-PBEE on each part-band. The pronunciation of the Mandarin sentence “SHIH-CHIEN-TA-HSIAO” is diagrammatically shown in Figure 8a. In detail, the waveform of the sentence under factory noise conditions is displayed in Figure 8b. The corresponding spectrogram is also shown in Figure 8c. We find that each TD-PBEE parameter accurately indicates the boundary of voice activity under 5 dB factory noise in Figures 8d–h. We also observe that the combined weighted TD-PBEE summing up the four TD-PBEEs can more accurately extract the voice-activity under 5 dB factory noise conditions than each weighted TD-PBEE.

**Figure 8.** The development of the combined TD-PBEE feature parameter.



### 3.3. The VAD Decision

Based on the description of the combined TD-PBEE using short-partition band number *N* and long-term window length *R*, the voice activity is determined by the decision rules as shown below:

$$\begin{aligned}&\text{if } (TD\text{-}PBEE(m) > Th_S) \quad VAD(m)=1;\\&\text{else if } (TD\text{-}PBEE(m) < Th_N) \quad VAD(m)=0;\\&\text{else} \quad VAD(m)=VAD(m-1)\end{aligned} \tag{23}$$

where $Th_S$ and $Th_N$ mean the speech thresholds and noise thresholds, respectively.

The two values can be recursively updated by using the mean and variance of the logarithmic combined TD-PBEE to estimate the time-varying noise characteristics [24]. In fact, we assume that the first four frames only contain noise 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:

$$Th_S = \mu_N + \alpha_S \cdot \sigma_N \tag{24}$$

$$Th_N = \mu_N + \beta_N \cdot \sigma_N \tag{25}$$

Similarly, $\mu_N$ and $\sigma_N$ represent the mean and the variance of the logarithmic combined TD-PBEE, respectively. In addition, $\alpha_S$ and $\beta_N$ are the adjustment constants which are used to determine the threshold.

The mean and variance of the logarithmic combined TD-PBEE are updated while the decision rule shows a noise period:

$$\begin{aligned}&\mu_N(m) = \gamma \cdot \mu_N(m-1) + (1-\gamma) \cdot \mathit{TD\text{-}PBEE}\\&\left[\mathit{TD\text{-}PBEE}^2\right]_{mean}(m) = \gamma \cdot \left[\mathit{TD\text{-}PBEE}^2\right]_{mean}(m-1) + (1-\gamma) \cdot \mathit{TD\text{-}PBEE}^2\\&\sigma_N(m) = \sqrt{\left[\mathit{TD\text{-}PBEE}^2\right]_{mean}(m) - \left[\mu_N(m)\right]^2}\end{aligned} \tag{26}$$

where $\gamma = 0.5$ is chosen by experiment. We then update the threshold using the updated mean and variance of the logarithmic combined TD-PBEE.

## 4. Evaluation and Results

In order to evaluate the proposed TD-PBEE VAD method, the speech database is first described in this section. In addition, the performances of comparison with state-of-the-art VAD algorithms (such as LSFM [13], BSE [19], G.729B [25], AMR2 [26], LTSD [12] and MTED [27]) will be reported as follows.

### *4.1. Database Description*

We used a set of 12 sentences (about 107 s) from four different speakers: two males and two females from the TIMIT database to evaluate the advantages of the proposed TD-PBEE feature sets for speech detection. 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 ranging between clean and 5 dB (from the NOISEX-92 database). All signals in the database were down-sampled to 8-kHz, mono-channel and 16-bits per sample. In addition, the optimal parameters for the proposed VAD were: $N_{LL} = 8$, $N_{LH} = 4$, $N_{HL} = 3$ and $N_{HH} = 2$; $R_{LL} = 5$, $R_{LH} = 10$, $R_{HL} = 15$ and $R_{HH} = 20$; $\eta_{HH} = 5$, $\eta_{HL} = 10$, $\eta_{HL} = 15$, and $\eta_{HH} = 20$, while the filter bank decomposed the signal into four part-band from Mel-scaled partition.

### *4.2. The Performance of Comparison with Sate-of-The-Art VAD Algorithms*

In order to clearly description the performance of VAD algorithms, the speech/non-speech hit rate (HR1/HR0) as a function of the SNR has been presented in this section. The average speech/non-speech hit rate (HR1/HR0) for each type of noise is employed for comparison between each one and calculated as below:

$$HR0 = \frac{\text{number of non-speech frames correctly classified}}{\text{number of real non-speech frames}} \times 100\% \tag{27}$$

$$HR1 = \frac{\text{number of speech frames correctly classified}}{\text{number of real speech frames}} \times 100\% \tag{28}$$

The speech/non-speech hit rate (HR1/HR0) as a function of the SNR for the proposed TD-LTE, G.729, AMR2, LTSD, MTED, BSE and LSFM VAD algorithms are shown in Figure 9 and Figure 10.

In these two Figures, we provide the results of non-speech hit rate (HR0) and speech hit rate (HR1), respectively. The results compare the proposed TD-PBEE VAD algorithm to G.729, AMR2, LTSD, MTED, BSE, and LSFM VADs from clean to 5 dB under the four types of noise conditions. We observe that the LSFM VAD is comparable to the proposed TD-PBEE VAD in term of HR0 analysis under lower SNR level. The standard G.729 VAD gives the worst performance among the reference VAD algorithms while performing HR0 analysis. Similarly, we also observe that the LTSD VAD is comparable to the proposed TD-PBEE VAD in terms of HR1 analysis under lower SNR level conditions. In addition, the standard AMR2 VAD has the worst performance among the reference VAD algorithms while performing HR1 analysis at lower SNR level.

**Figure 9.** Non-speech hit rate (HR0) from clean to 5 dB under the four types of noise.

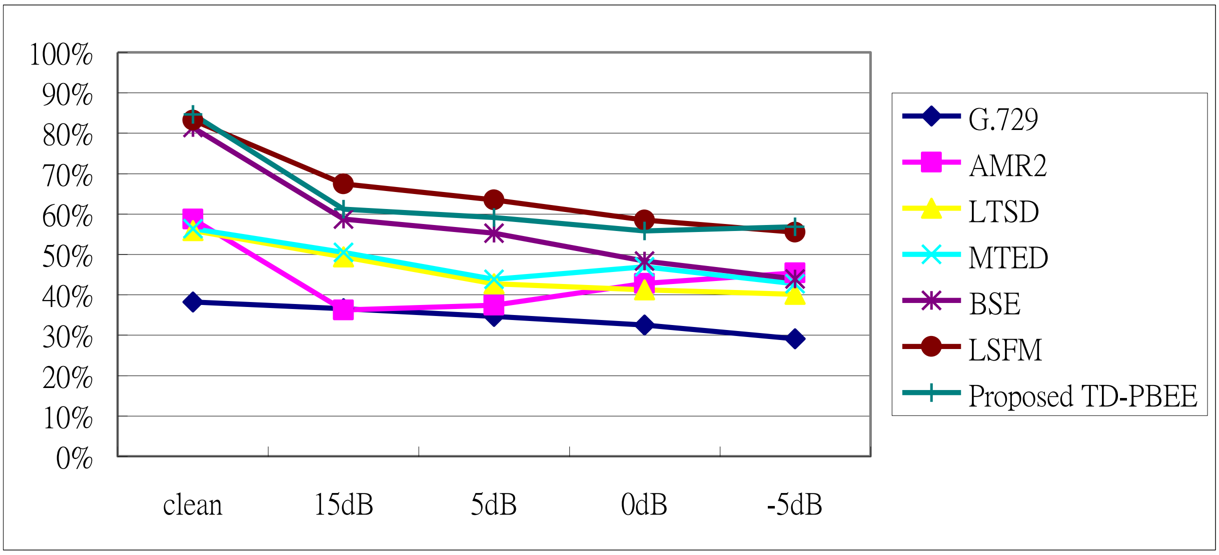
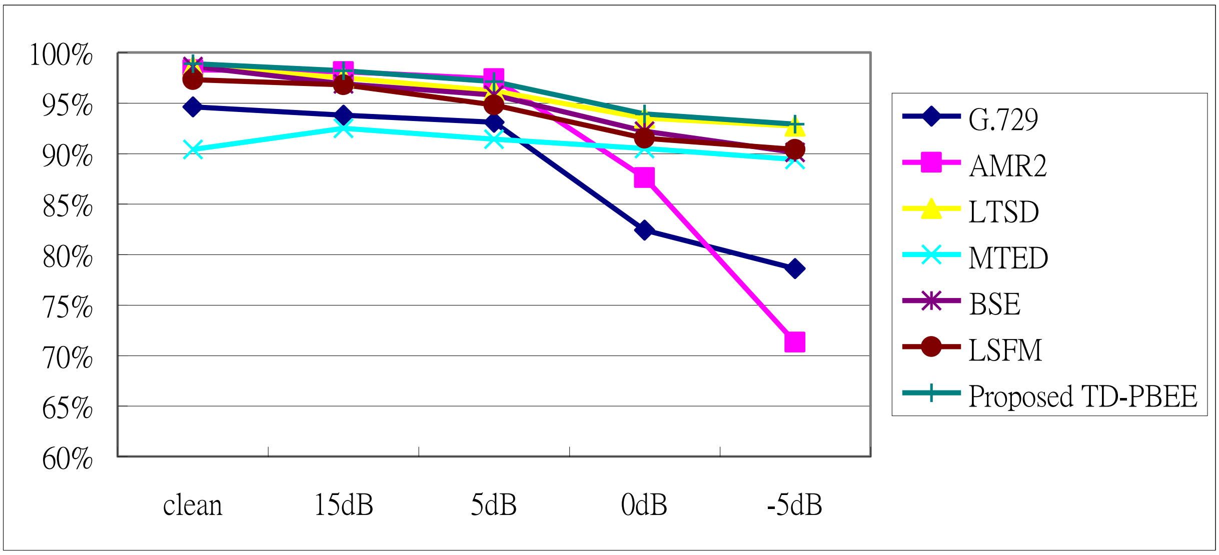


**Figure 10.** Speech hit rate (HR1) from clean to 5 dB under the four types of noise.



In order to further describe the efficiency of VAD for the different types of noises of the NOISEX database, the comparison of performances of VAD algorithms has also been presented in Table 2 and

Table 3. We observe that the average accuracy of LSFM VAD is better than the proposed TD-PBEE VAD in Table 2. In detail, the LSFM VAD is superior to the proposed TD-PBEE VAD while testing in factory noise and car noise. However, the LSFM VAD is worse than the proposed TD-PBEE VAD while testing in babble noise. In Table 3, we also observe that the average accuracy of the proposed TD-PBEE VAD is best among all reference VAD algorithms, especially in babble noise. The LTSD is second accuracy of detecting voice. We summarize that the proposed TD-PBEE VAD attains a 63.55% HR0 average value in non-speech detection. Besides, the proposed TD-PBEE VAD also obtains the best behavior in detecting speech with a 96.2% HR1 average value.

**Table 2.** Non-speech hit rate (HR0) from clean to 5 dB under the four types of noise.

| **NOISEX Database** / **VAD Algorithm** | **White** | **Factory** | **Car** | **Babble** | **Average** |
|---|---|---|---|---|---|
| **G.729** | 71.10% | 23.50% | 21.30% | 20.90% | 34.20% |
| **AMR2** | 58.70% | 37.40% | 37.50% | 42.80% | 44.10% |
| **LTSD** | 54.30% | 43.40% | 45.70% | 40.20% | 45.90% |
| **MTED** | 54.30% | 50.40% | 43.80% | 43.90% | 48.10% |
| **BSE** | 81.20% | 51.70% | 52.20% | 44.90% | 57.50% |
| **LSFM** | 82.90% | 63.20% | 61.50% | 48.20% | 63.95% |
| ***Proposed TD-PBEE*** | *83.10%* | *59.70%* | *58.50%* | *52.90%* | *63.55%* |

**Table 3.** Speech hit rate (HR1) from clean to 5 dB under the four types of noise.

| **NOISEX Database** / **VAD Algorithm** | **White** | **Factory** | **Car** | **Babble** | **Average** |
|---|---|---|---|---|---|
| **G.729** | 89.70% | 90.80% | 91.10% | 82.40% | 88.50% |
| **AMR2** | 92.30% | 91.70% | 92.40% | 85.60% | 90.50% |
| **LTSD** | 98.60% | 95.50% | 96.20% | 92.50% | 95.70% |
| **MTED** | 90.20% | 92.50% | 91.40% | 89.50% | 90.90% |
| **BSE** | 97.80% | 95.00% | 95.10% | 90.90% | 94.70% |
| **LSFM** | 97.50% | 92.30% | 93.60% | 88.00% | 92.85% |
| ***Proposed TD-PBEE*** | ***98.40%*** | ***96.10%*** | ***96.50%*** | ***93.80%*** | ***96.20%*** |

Then, the error norm of false alarm rates, $E_{norm}$, is used to further quantify the speech/non-speech hit rates, and it is defined as:

$$E_{norm} = \sqrt{(1-HR1)^2 + (1-HR0)^2} \tag{27}$$

Table 4 shows an average speech/non-speech hit rates ($HR0$ and $HR1$), and overall false error norm ($E_{norm}$) for SNR level from clean to 5 dB. We found that the proposed TD-PBEE achieved the minimum false alarm error norm with a 36.65% value and was obviously superior to other VAD algorithms.

**Table 4.** Average speech/non-speech hit rates and overall false error norm for SNR level from clean to −5 dB.

| **VAD**<br>**Evaluation** | **G.729** | **AMR2** | **LTSD** | **MTED** | **BSE** | **LSFM** | ***Proposed***<br>***TD-PBEE*** |
|---|---|---|---|---|---|---|---|
| **HR1(%)** | 88.50% | 90.50% | 95.70% | 90.90% | 94.70% | 92.85% | *96.20%* |
| **HR0(%)** | 34.20% | 44.10% | 45.90% | 48.10% | 57.50% | 63.95% | *63.55%* |
| ***Error norm(%)*** | 66.80% | 56.70% | 54.27% | 52.69% | 42.83% | 36.75% | ***36.65%*** |

## 5. Conclusions

In this paper, we present a novel two-dimensional part-band energy entropy (TD-PBEE) based on short-partition band number *N* and long-term window length *R*. The proposed TD-PBEE-based VAD is composed of four components: Mel-scaled filter bank, TD-PBEE feature extraction, part-band weighting estimation, and the VAD decision. We found that the two-dimensional entropy improves one-dimensional entropy according to the experimental results. We also discussed the estimation of the part-band weighting that can help to understand the useful spectral information of each part-band. We also observed that the optimal parameters: *R* and *N* can increase the accuracy of voice detection. We also performed experiments with the VAD decision, the two thresholds for speech and noise can be updated to detect the speech voice. Future research will apply this to voice-command applications in a realistic environment to increase accuracy.

## Acknowledgments

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC 102-2221-E-158-006, and partially sponsored by the research project of Shih-Chien University, Taiwan, under contract number USC-100-05-05007, USC-100-08-03003, USC-101-08-02006, USC-102-08-01005 and USC-102-05-05014.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Freeman, D.K.; Cosier, G.; Southcott, C.B.; Boyd, I. The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK, 23–26 May 1989, pp. 369–372.
2. Srinivasan, K.; Gersho, A. Voice Activity Detection for Cellular Networks. In Proceedings of the Speech Coding for Telecommunications, Sainte-Adele Quebec, Canada, 13–15 October 1993; pp. 85–86.
3. Sasaki, S.; Matsumoto, R. Voice activity detection and transmission error control for digital cordless telephone system. *IEICE Trans. Commun*. **1994**, *7*, 948–955.

4. Rabiner, L.R.; Sambur, M.R. Voiced-Unvoiced-Silence Detection Using the Itakura LPC Distance Measure. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, USA, 9–11 May 1977; pp. 323–326.
5. Haigh, J.A.; Mason, J.S. Robust Voice Activity Detection Using Cepstral Features. In Proceedings of the IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering, Beijing, China, 19–21 October 1993; pp. 321–324.
6. Hoyt, J.D.; Wechsler, H. Detection of Human Speech in Structured Noise. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, SA, USA, 19–22 April 1994; pp. 237–240.
7. Tucker, R. Voice activity detection using a periodicity measure. *IEE Proc. I (Commun. Speech Vis.)* **1992**, *139*, 377–380.
8. Craciun, A.; Gabrea, M. Correlation Coefficient-Based Voice Activity Detector Algorithm. In Proceedings of the Canadian Conference on Electrical and Computer Engineering, Ontario, Canada, 2–5 May 2004; pp. 1789–1792.
9. Lee, Y.C.; Ahn, S.S. Statistical model-based VAD algorithm with wavelet transform. *IEICE Trans. Fundament*. **2006**, *E89-A*, 1594–1600.
10. Shen, J.L.; Hung, J.W.; Lee, L.S. Robust Entropy-Based Endpoint Detection for Speech Recognition in Noisy Environments. In Proceedings of the 5th International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998; pp. 1015–1018.
11. Soleimani, S.A.; Ahadi, S.M. Voice Activity Detection Based on Combination of Multiple Features Using Linear/Kernel Discriminant Analyses. In Proceedings of the 3rd International Conference on Information & Communication Technologies: from Theory to Applications. Damascus, Syria, 7–11 April 2008; pp. 1–5.
12. Ramirez, J.; Segura, J.C.; Benitez, M.C.; Torre, A.D.L.; Rubio, A. Efficient voice activity detection algorithms using long-term speech information. *Speech Commun*. **2004**, *42*, 271–287.
13. Yanna, M.; Akinori, N. Efficient voice activity detection algorithm using long-term spectral flatness measure. *EURASIP J. Audio Speech Music Process.* **2013**, doi:10.1186/ 1687-4722-2013-21.
14. Sohn, J.; Kim, N.S.; Sung, W. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.* **1999**, *6*, 1–3.
15. Cho, Y.D.; Al-Naimi, K.; Kondoz, A. Improved Statistical Voice Activity Detection Based On a Smoothed Statistical Likelihood Ratio. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing Salt Lake City, USA, 7–11 May 2001; pp. 737–740.
16. Renevey, P.; Drygajlo, A. Entropy Based Voice Activity Detection in Very Noisy Conditions. In Proceedings of the EUROSPEECH2001, Aalborg, Denmark, 3–7 September 2001; pp. 1887–1890.
17. Breithaupt, C.; Gerkmann, T.; Martin, R. A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, USA, 30 March–4 April 2008; pp. 4897–4900.

18. Tanyer, S.G.; Ozer, H. Voice activity detection in nonstationary noise. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 478–482.
19. Wu, B.F.; Wang, K.C. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 762–775.
20. Moore, B. *Hearing*; Academic Press, Inc.: San Diego, CA, USA, 1995.
21. O'Shaughnessy, D. *Speech Communication*; Addision-Wesley: Reading, MA, USA, 1987, p. 150.
22. Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 504–512.
23. Doblinger, G. Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proc. Eurospeech* **1995**, *2*, 1513–1516.
24. Gerven, S.V.; Xie, F. A comparative study of speech detection methods. *Eurospeech* **1997**, *97*, 1095–1098.
25. Benyassine, A.; Shlomot, E.; Su, H. 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.* **1997**, *35*, 64–72.
26. Digital Cellular Telecommunications System (Phase 2+), Adaptive Multi Rate (AMR) Speech, ANSI-C Code for AMR Speech Codec; ETSI TS 100 905 V7.0.0 (1999-08); ETSI: Sophia Antipolis, France, 1998.
27. Evangelopoulos, G.; Maragos, P. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 2024–2038.