Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 39546, 9 pages
doi:10.1155/2007/39546

Research Article

Speech/Nonspeech Detection Using Minimal Walsh Basis Functions

Moe Pwint and Farook Sattar

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Received 1 November 2005; Revised 30 May 2006; Accepted 12 June 2006

Recommended by Mark Clements

This paper presents a new method to detect speech/nonspeech components of a given noisy signal. Employing the combination of binary Walsh basis functions and an analysis-synthesis scheme, the original noisy speech signal is modified first. From the modified signals, the speech components are distinguished from the nonspeech components by using a simple decision scheme. Minimal number of Walsh basis functions to be applied is determined using singular value decomposition (SVD). The main advantages of the proposed method are low computational complexity, less parameters to be adjusted, and simple implementation. It is observed that the use of Walsh basis functions makes the proposed algorithm efficiently applicable in real-world situations where processing time is crucial. Simulation results indicate that the proposed algorithm achieves high-speech and nonspeech detection rates while maintaining a low error rate for different noisy conditions.

1. INTRODUCTION

Speech/nonspeech detection is simply the task of discriminating noise-only frames of a signal from its noisy speech frames. In the literature, this process is usually known as voice activity detection (VAD) and it becomes an important problem in many areas of speech processing such as real-time noise reduction for speech enhancement, speech recognition, digital hearing aids, and modern telecommunication systems. In multimedia communications, silence compression algorithms are usually applied to reduce the average transmission rate during silence periods of speech. These compression algorithms are also based on speech/silence detection and they allow the speech channel to be shared with other information so that the capacity of channel can be improved. Furthermore, VAD is an essential component in variable rate speech coders to achieve efficient bandwidth reduction without speech quality degradation. Several methods that trade off the accuracy, delay, perceptual quality, and computational complexity have been proposed in the literature to deal with the problem of speech/nonspeech detection.

A silence compression speech communication system with VAD was standardized by ITU-T Recommendation G. 729 [1, Annex B]. It uses a feature vector consisting of four parameters: full-band energy, low-band energy, zero-crossing rate, and a spectral measure for the multiboundary decisions. Based on the difference between each parameter and its respective long-term average, the fourteen boundary decisions are defined. The initial voice activity decision for each frame is set to 1 if one of these multiboundary decisions in the space of the four difference measures is true. Final decision is made by smoothing the initial decision in four stages (i.e., hangover scheme). A voice detection algorithm based on a pattern recognition approach and fuzzy logic was proposed for wireless communications in noisy environments [2]. This algorithm uses the same acoustic parameters adopted by G.729 for feature extraction.

A VAD standardized for the GSM cellular communication system is the ETSI speech coder [3]. Based on the spectral estimation and periodicity detection, this adaptive multirate speech coder (AMR) specifies two options for VAD to be used in DTX (discontinuous transmission) mode. For applications like mobile phones and packet networks, discontinuous transmission (DTX) mode is usually required for lower bit-rate transmission speech coder. In AMR Option 1, the input signal is divided into subbands and the level of signal in each band is calculated. The VAD decision is made by using the outputs from pitch detection, tone detection, complex signal analysis modules, and signal level. A hangover scheme is also added before the final decision is made.

In AMR Option 2 the input signal is first converted into frequency domain using discrete Fourier transform (DFT). Then, based on the channel energy estimator, channel SNR estimator, spectral deviation estimator, background noise estimator, peak-to-average-ratio module, and voice metric calculation module, the VAD decision is made.

Apart from the above voice activity detection methods, most of which are based on the parameters of speech, model-based VADs have been introduced recently. Formulating the problem of speech pause detection into a statistical decision theory, two detectors based on maximum a posteriori probability (MAP) and Neyman-Pearson test were described in [4]. A Gaussian statistical model which assumes that the discrete Fourier transform coefficients of speech and noise are asymptotically independent Gaussian random variables was proposed in [5, 6]. Assuming the distributions of speech and noise signals to be Laplacian and Gaussian models, the authors in [7] developed a soft voice activity detector by decomposing the speech signal into discrete cosine transform (DCT) components.

Noise is a well-known factor which degrades the quality and intelligibility of speech in many applications' areas. To reduce the noise level without affecting the quality of speech signals, a noise reduction algorithm is usually employed. Spectral subtraction is a widely used approach in practical noise suppression schemes. This scheme usually estimates the noise characteristics from the nonspeech intervals of the signal. Therefore, identification of nonspeech periods is an important and sensitive part of existing noise reduction schemes. In this context, accuracy and reliability of a VAD becomes critical in determining the performance of noise reduction algorithm. Most papers reporting on noise reduction refer to speech pause detection when dealing with the problem of noise estimation. Speech pause detectors are very sensitive and often limiting part of the systems for the reduction of noise in speech [8].

A speech pause detection algorithm based on an autocorrelation voicing detector algorithm was developed in [9]. The algorithm was designed for real-time system and implemented on a DSP platform for the application of speech enhancement for hearing aids. An adaptive Karhunen-Loéve transform (KLT) tracking-based algorithm was also proposed for enhancement of speech degraded by additive color noise [10]. An algorithm, which detects the speech pauses by tracking the dynamics of the signal's temporal power envelope, was proposed in [8]. Sometimes, detection algorithms were designed for specific applications such as noise suppression [11] and wideband coding [12]. Voice activity detection algorithms for cellular networks in the presence of babble noise and vehicular noise were presented in [13] by adopting the approach used in European digital mobile cellular standard [14]. Combining the geometrically adaptive energy threshold method (GAET) and least-square periodicity estimator (LSPE), conversational speech is separated from silence [15]. A fuzzy polarity correlation function is also applied to determine speech sections and background noise in the environment of telephone network [16].

In this paper, a method to discriminate the active and inactive periods of speech signals corrupted by unknown type and unknown level of noise is presented. It is assumed that intervals of the inactive segments can be short as well as long (i.e., while some active segments are located very closely, some active segments may be separated by longer periods). Taking the simplicity of binary Walsh transform as an advantage, the proposed speech/nonspeech detection algorithm is developed. First, the signal to be classified is modified employing binary Walsh basis functions. The minimal number of basis functions to be applied is determined by using a technique for the selection of wavelet decomposition at natural scale [17]. Using the statistics of the modified signals, which are highly informative about the characteristics of noisy speech frames as well as noise only frames, classification is performed with a decision scheme.

Unlike other VAD methods, in which the decision is made on a frame-by-frame basis, the proposed method instantaneously obtains the set of consecutive frames as speech and nonspeech segments. The effectiveness of the proposed method is evaluated by conducting the objective performance on different types of noise with varying SNRs using the criteria of error rate, speech/nonspeech detection rates, and false alarm rate. ROC analyses have been shown to compare the standardized algorithms: G.729 and AMR Option 1 and Option 2. Experimental results show that the detection accuracy of the proposed algorithm is high for both speech and nonspeech frames regardless of noise levels.

2. PROPOSED ALGORITHM

The block diagram of the proposed speech/nonspeech detection algorithm based on the binary Walsh basis functions is depicted in Figure 1. First, the signal is represented using FFTs. These representations are then modified by Walsh basis functions before reconstructing. The number of basis functions to be applied is determined using SVD. Finally, speech/nonspeech periods are detected from the modified signals utilizing a decision scheme. Details of the algorithm are explained in the following sections.

2.1. Modification of signal

The noisy input signal is reconstructed as a modified sequence based on an analysis/synthesis scheme described in [18]. Firstly, the input signal $x(n)$ of sampling frequency 8 kHz is multiplied by a Hanning window to yield successive windowed segments of $x_s(n)$. These window segments are transformed into the spectral domain by using FFTs of size 128. In this manner, a time varying spectrum $X_s(n,k) = |X_s(n,k)|e^{j\varphi(n,k)}$ with $n = 0, 1, \ldots, N-1$ and $k = 0, 1, \ldots, N-1$ for each windowed segment is computed. Here, $X_s(n,k)$ denotes the spectral component of the noisy input signal at frequency index k and time index n. Before synthesis, each sth windowed segment is modified as the weighted sum of the magnitude $|X_s(n,k)|$ using binary Walsh basis functions. Using basis functions, the number of parameters to track along the variations between active and inactive regions of

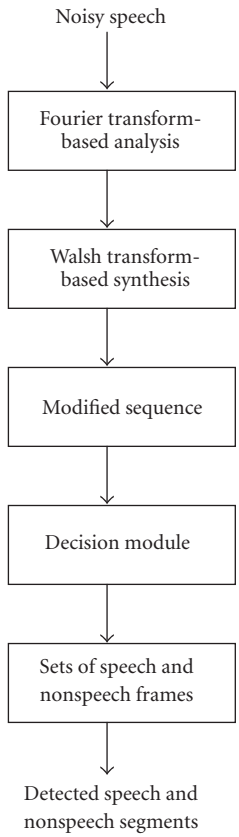

Figure 1: Block diagram of the proposed algorithm.

the noisy signal can be lessened. In this context, SVD is used to determine the minimal number of Walsh basis functions to be applied. The detailed procedure for the identification of the minimal number of Walsh basis functions is described in the next section. Applying the ith basis function ϕ_i, a modified sequence, $y_s(n)$, for each windowed segment can be obtained as

$$y_s(n) = \sum_{k=0}^{N-1} |X_s(n,k)| \cdot \phi_i(k). \tag{1}$$

All the modified segments of S are then concatenated producing an output signal $y(n)$ by showing the time-varying magnitude responses:

$$y_i(n) = \sum_{s=0}^{S-1} y_s(n - sN). \tag{2}$$

2.2. Determination of minimal Walsh basis functions

The Walsh transform is a matrix consisting of a complete orthogonal function set having only two values +1 and −1 over their definition intervals. The motivation for using Walsh transform rather than other transforms is its computational simplicity giving a realistic processing time. The Walsh function of order N can be represented as

$$\phi(x,u) = \frac{1}{N} \prod_{i=0}^{q-1} (-1)^{b_i(x) b_{q-1-i}(u)}, \tag{3}$$

where $u = 0, 1, \ldots, N-1$, $N = 2^q$, and $b_i(x)$ is the ith bit value of x. In this context, the Walsh functions are arranged into sequence order, the number of zero crossings of Walsh function per definition interval, to obtain a set of basis functions. The number of zero crossings increases with the order of basis functions $W = [\phi_0, \phi_1, \ldots, \phi_{N-1}]$.

It is very important to select the proper basis functions so that variations between the dynamics of speech and nonspeech can be captured more precisely. A method to select the global natural scale in discrete wavelet transform [17] is adopted to determine the required number of basis functions. This method adaptively detects the optimal scale using SVD while decomposition is being carried out. Consider an input noisy speech signal x of length $\mathcal{V}$, and $y_d(v)$ being its modified sequence obtained applying the basis functions of order d into (1) and (2).

Modified sequences $\{y_d(v)\}_{d=0}^{D-1}$ can be represented in a matrix P of size $D \times \mathcal{V}$. To determine the order of basis functions with dominant eigenvalues, the SVD of the matrix P is calculated adaptively starting with the first two orders (i.e., ϕ_0 and ϕ_1) while adding the higher orders.

In order to determine the number of basis functions to be applied, we studied the probability distributions of basis function orders as a function of SNRs. In this analysis, speech signals from TIDIGITS database spoken by male and female speakers were used. If there exist long interword silences, they were removed first. Silence segments of different sizes were then introduced to have varying intervals between active regions. To generate the noisy signals, the commonly used white Gaussian noise was artificially added with SNR levels of 20 dB, 10 dB, 5 dB, and 0 dB. Here, SNR is defined as

$$\text{SNR} = 10 \log_{10} \left\{ \frac{\sum_{n=1}^{N_s} s^2(n)}{\sum_{n=1}^{N_v} v^2(n)} \right\}, \tag{4}$$

where s is speech, v is noise, and N_s and N_v are the lengths of speech and noise signals, respectively.

Figure 2 displays the probability of occurrence of a basis function order, termed as coverage, for changing levels of SNR. It is observed that dominant eigenvalue is located only within the first few basis functions. In particular, the minimal order for highly noisy signals of 5 dB and 0 dB is found to be 1. And for the signals at high SNR of 20 dB, 10 dB, and clean, the dominant eigenvalue is found when the order of basis function is 3. Hence, the lower-order basis functions of Walsh transform matrix are highly informative and they should be used in modification process. Moreover, it is found that higher-order coefficients carry less weight in terms of their magnitude and may not be evident to interpret a large Walsh kernel [19].

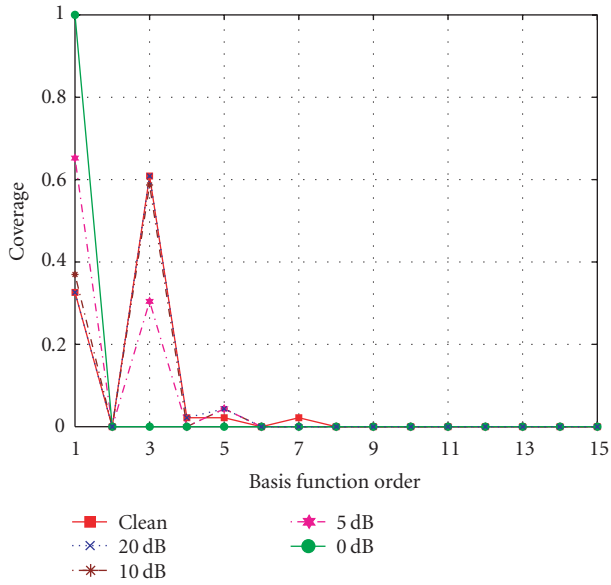

Figure 2: The distribution of the order of basis functions for the signals from clean to 0 dB.

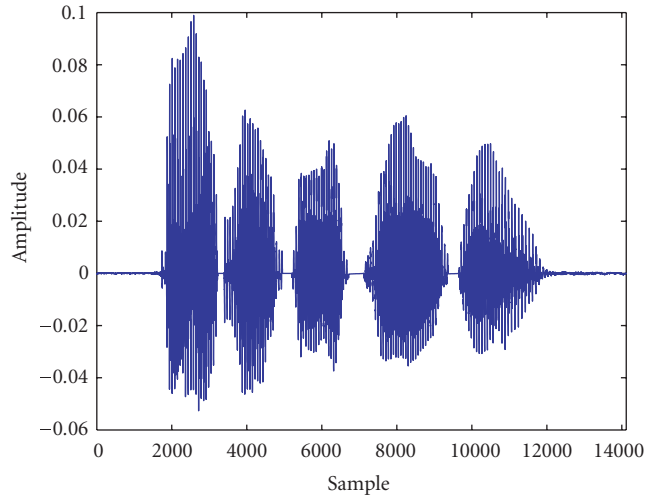

Figure 3: The clean signal.

In practice, it is not possible to obtain any *a priori* information about noise level and noise type. Hence, the proposed algorithm defines the minimal order of basis functions $N_{\min}$ as 3 throughout the experiments. In the original algorithm [17], optimal scale is defined as the average of the details from the first level to the natural scale, the level associated with the dominant eigenvalues. However, this averaging may introduce clipping effect for the signals with low speech level. To avoid this effect, a shifting operator which swaps the right and left halves of the basis function coefficients is applied first. Then a good estimate of the binary Walsh basis function at dominant eigenvalue is defined as

$$\psi = \frac{\phi_0 - \sum_{i=1}^{N_{\min}} \mathrm{CS}\,(\phi_i)}{\max\left\{\left|\phi_0 - \sum_{i=1}^{N_{\min}} \mathrm{CS}\,(\phi_i)\right|\right\}}, \tag{5}$$

where $N_{\min} = 3$ is the largest-order relating the most prominent eigenvalues and $\mathrm{CS}(\cdot)$ is the shifting operator. This new basis function ψ provides sharper representation and higher discriminating features. It is also found that identification between noisy speech periods and noise only components with narrow intervals become more apparent in the modified sequence obtained by using ψ.

For length N, the function ψ consists of 1's for $n = 0,\ldots,N/2-1$ followed by -1's for $n = N/2,\ldots,3N/4-1$ and 1's for $n = 3N/4,\ldots,N-1$, where n is the sample index. Substituting the values of ψ in (1), we find

$$y_s(n) = \sum_{k=0}^{N/2-1} |X_s(n,k)| + \left(\sum_{k=3N/4}^{N-1} |X_s(n,k)| - \sum_{k=N/2}^{3N/4-1} |X_s(n,k)|\right). \tag{6}$$

In order to compare ψ with ϕ_o, we replace ϕ_i with ϕ_o and rewrite (1) as

$$y_s(n) = \sum_{k=0}^{N-1} |X_s(n,k)|. \tag{7}$$

Using (7), the difference between the "short-term area under the magnitude spectrum" for the noisy speech case and the noise only case (specially for white Gaussian noise) will be less due to the sum taken over the whole 0–4 kHz frequency band. Based on the expressions of (6) and (7), we can notice that the discrimination between speech and nonspeech segments will be higher for using ψ compared to ϕ_o.

To demonstrate the effectiveness of the proposed modification presented above, an example is shown in Figures 3–5. A clean signal is shown in Figure 3. The modified version of this signal in white Gaussian noise at 5 dB SNR using 0-order basis function ϕ_0 and estimated basis function ψ is also shown in Figures 4 and 5, respectively. It is observed from Figures 4 and 5 that discriminating ability of the modified signal y_m as obtained using ψ is better for the speech and nonspeech frames due to its deeper and sharper representation.

It seems that the function ψ is more efficient to capture the *intrasegment* variation between the noisy speech segments and noise only segments of narrow interval.

2.3. Decision scheme

First, 0-order basis function, ϕ_0 is used to produce a modified sequence, $y_0(\nu)$, to get the global information of the original noisy signal. This modified sequence is used as a reference or pilot signal as in the area of telecommunication. In telecommunication, a pilot signal is usually transmitted over a communication system for supervisory, control, or reference purposes. Carrying the local characteristics, another modified signal, $y_m(\nu)$, is formed using the new basis function ψ. From

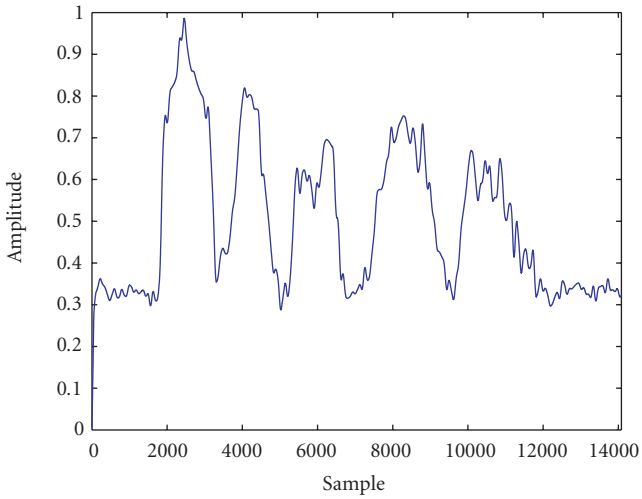

Figure 4: The modified signal using 0-order basis function ϕ_0.

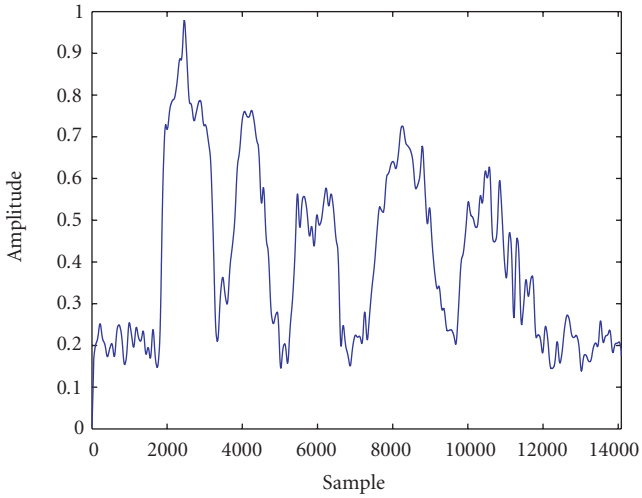

Figure 5: The modified signal using basis function ψ.

this sequence, locations and durations of speech active and inactive periods can be captured more precisely. In this way, the approximate locations of active and inactive frames are first determined from the modified signal, $y_0(\nu)$. Then, the accuracy of these reference decisions are improved by using the second modified signal, $y_m(\nu)$, containing the detailed information. Applying the reconstructed signals y_0 and y_m, the procedure of detection scheme can be described as follows.

(i) Extract two sequences of local minima, $\{\alpha_{0i}\}_{i=1}^{L}$ and $\{\alpha_{mi}\}_{i=1}^{L}$, where L is the number of frames, from every 4 ms frame of $y_0(\nu)$ and $y_m(\nu)$ for which it is assumed that the initial 200 ms consists of noise only period.

(ii) Set thresholds, τ_0 and τ_m, for each minima sequence which are obtained using a simple statistics as $\tau_0 = \mu_0 - \kappa\delta_0$ and $\tau_m = \mu_m - \kappa\delta_m$, where μ_0 and δ_0 are the mean and the standard deviation of the first set of local minima, and μ_m and δ_m are those of the second set of local minima while κ is a positive value. After experimenting with the modified waveforms for a number of clean as well as noisy speech data, κ is set to be 0.75.

(iii) Declare a frame as an inactive frame if either $\alpha_{0i} < \tau_0$ or $\alpha_{mi} < \tau_m$. In this way, the nonactive frame indices are obtained from $y_0(\nu)$ and $y_m(\nu)$ as $\mathcal{R}$ and $\mathcal{T}$:

$$\begin{aligned} \mathcal{R} &= \{r_1, r_2, \ldots, r_P\}, \\ \mathcal{T} &= \{t_1, t_2, \ldots, t_Q\}. \end{aligned} \tag{8}$$

(iv) Combine the two initial boundary decisions as follows:

$$\mathcal{C} = \mathcal{R} \cap \mathcal{T}, \tag{9}$$

where $\mathcal{C}=\{c_1, c_2, \ldots, c_J\}$ is the set of elements common to $\mathcal{R}$ and $\mathcal{T}$. Considering that the members of $\mathcal{C}$ are the indices of the inactive frames, the final decision for detecting speech and nonspeech frames are obtained.

Here, we decide that there exist inactive frames whenever some or all of the prominent local minima obtained from the first modified signal $y_0(\nu)$ would coincide with the local minima found from the second modified signal $y_m(\nu)$. For those detected frames when their corresponding local minima are not obtained from both modified sequences of $y_0(\nu)$ and $y_m(\nu)$ are discarded as outliers.

3. EXPERIMENTAL RESULTS AND COMPARISON

In this section, the results and objective evaluation of the proposed method is presented. The detection result for a noisy speech signal is illustrated in Figure 6, where the signal is at 0 dB SNR and embedded in white Gaussian noise. The results obtained by the proposed detection scheme are shown together with manually determined actual speech and nonspeech detection results. It is seen that the detection accuracy is high for both speech and nonspeech periods. And thus the proposed algorithm achieves a good performance level.

3.1. Evaluation data

To evaluate the efficiency of the proposed method, its performance was compared with G.729 VAD and AMR Options 1 and 2. For the comparison purpose, the speech signals from 11 speakers of TIDIGITS database were extracted. Three signals from each of these male and female speakers were concatenated to generate the signals of 8 s to 11 s long. Silence or pause segments of varying intervals were then inserted between the active segments as described in Section 2.2. Test sequences consist of nearly 70% of active speech components and 30% of inactive speech components. The silence segments of very short as well as long durations are also included in the test sequences. For reference decisions, active and inactive frames of all clean signals were marked manually. Five types of noise, white Gaussian, babble, car, street, and train, were added to the original signals with different SNRs 20 dB, 10 dB, and 0 dB.

Table 1: Comparison of speech detection rates, nonspeech detection rates, and error rates of the proposed method to standard methods (G.729, AMR1, and AMR2) for different levels of SNRs in various noisy environments.

		Speech detection DS(%)				Nonspeech detection DNS(%)				Error rate E(%)			
Noise	SNR	Proposed	G.729	AMR1	AMR2	Proposed	G.729	AMR1	AMR2	Proposed	G.729	AMR1	AMR2
White	20 dB	89.20	96.79	96.26	97.07	95.48	31.51	61.09	48.21	9.81	20.85	12.41	15.56
	10 dB	88.48	90.42	93.03	92.01	95.13	42.21	45.11	52.52	10.53	22.74	18.68	18.12
	0 dB	87.07	67.09	81.32	60.57	81.26	62.37	56.98	77.97	15.97	34.72	24.72	35.49
Car	20 dB	88.76	97.65	97.84	98.06	96.40	19.19	62.61	45.95	9.76	23.55	11.04	15.62
	10 dB	88.01	95.42	96.36	93.64	92.47	17.04	51.21	50.31	11.74	25.59	15.30	17.76
	0 dB	87.37	91.55	81.02	64.46	70.35	16.57	55.53	70.50	18.28	28.67	26.10	34.92
Babble	20 dB	88.34	97.02	98.20	97.82	95.45	19.60	56.84	42.51	10.33	23.84	12.32	17.17
	10 dB	89.11	93.85	98.44	95.28	84.44	18.58	29.09	40.81	13.48	26.91	19.81	19.69
	0 dB	86.19	90.46	90.85	85.87	56.32	14.44	31.02	37.46	22.74	29.99	25.04	27.23
Street	20 dB	88.55	96.41	97.33	98.37	95.20	21.85	66.16	47.36	10.31	23.90	10.45	15.07
	10 dB	89.60	92.49	97.36	93.12	83.51	17.28	45.98	51.95	12.95	27.75	15.85	17.79
	0 dB	84.51	88.81	86.80	69.22	65.61	13.75	46.46	67.87	21.55	31.26	23.89	31.71
Train	20 dB	88.86	97.22	97.20	98.66	95.85	23.47	67.40	50.69	9.84	22.91	10.20	13.85
	10 dB	88.10	93.47	96.44	96.08	92.40	25.50	60.08	54.42	11.50	24.66	12.68	14.81
	0 dB	84.83	90.92	86.10	78.88	82.87	14.22	62.16	70.22	16.75	29.65	19.91	23.89

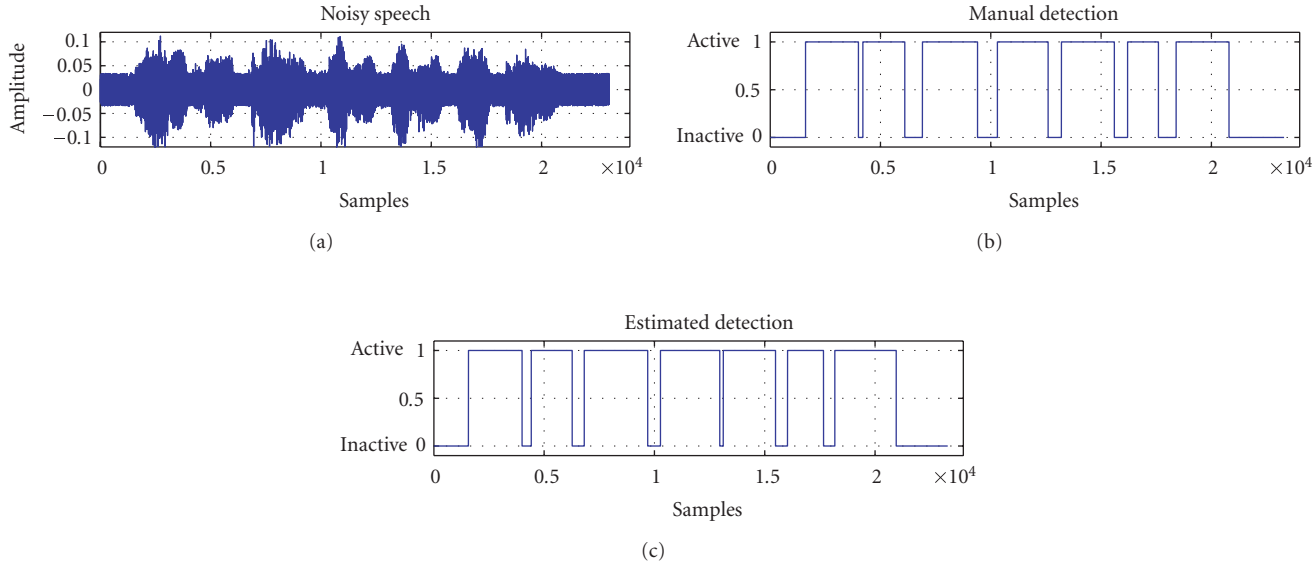

Figure 6: (a) Noisy speech at 0 dB SNR in white Gaussian noise, (b) manual detection, (c) estimated detection.

3.2. Performance evaluation

As performance criteria, the speech detection rate, nonspeech detection rate, and error rate were employed. Speech and nonspeech detection rates are defined as the ratio of the correctly classified speech frames to the total number of speech frames and the ratio of the correctly classified nonspeech frames to the total number of nonspeech frames, respectively. The error rate is defined as the ratio of the incorrectly classified frames to the total number of frames. In Table 1, speech/nonspeech detection rates and error rates of the proposed method are compared to the standardized VADs: G.729, AMR Options 1 and 2 under different noise sources and SNR levels. Speech detection accuracy of ITU G.729, ETSI AMR1, and AMR2 decreases with increasing noise levels in all noise types. Proposed binary Walsh transform based method can consistently detect the speech frames with almost constant rate regardless of noise types and levels. Considering the nonspeech detection rates, G.729 is the worst with an accuracy of less than 20% for most of the time. Although AMR1 and AMR2 yield better detection rate than G.729, the proposed method is found to be the best one in

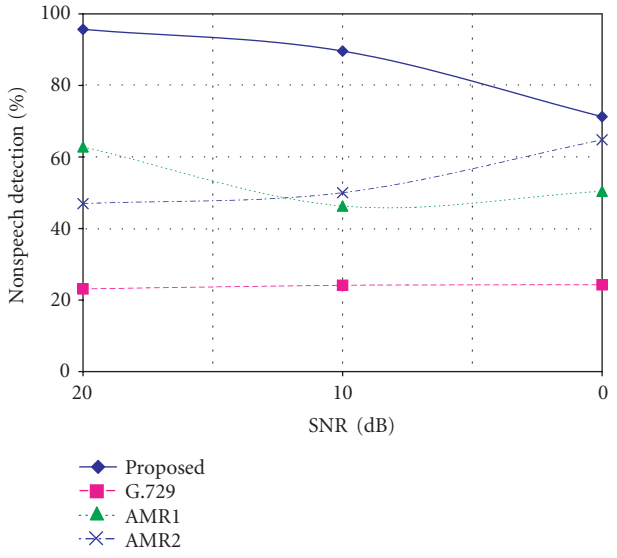

Figure 7: Performance comparison for average nonspeech detection rate of the proposed method and standard VADs (G.729, AMR1, and AMR2) in different backgrounds with varying SNRs.

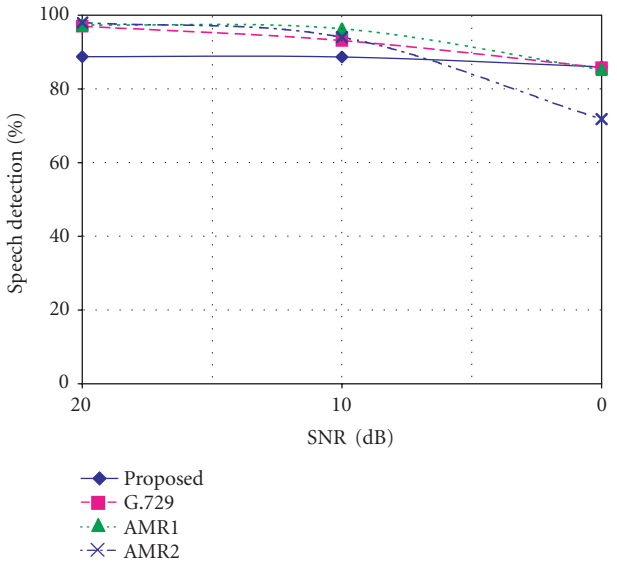

Figure 8: Performance comparison for average speech detection rate of the proposed method and standard VADs (G.729, AMR1, and AMR2) in different backgrounds with varying SNRs.

the problem of nonspeech detection for all noise conditions. Moreover, the proposed method can detect both speech and nonspeech frames with least error probabilities for all levels of SNRs in all environments.

The results of the performance comparisons for average rates of speech detection, nonspeech detection and error of the proposed method to ITU G.729, AMR Options 1 and 2 in five background noise (white, babble, car, street, and train) and SNR ranging from 20 dB to 0 dB are shown in Figures 7, 8, and 9. Average speech detection rates of the proposed

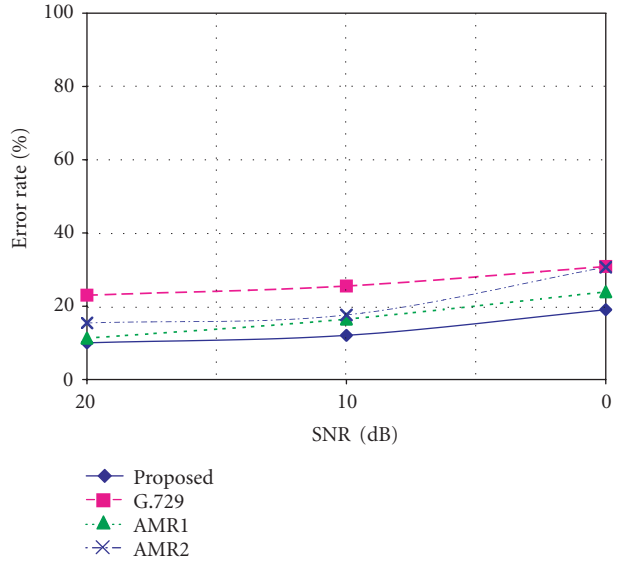

Figure 9: Performance comparison for average error rate of the proposed method and standard VADs (G.729, AMR1, and AMR2) in different backgrounds with varying SNRs.

method is nearly constant for varying SNRs of 20 dB, 10 dB, and 0 dB with their respective values of 88.74%, 88.66%, and 85.99%. Although the speech detection rates of above standardized methods are high in 20 dB, their performance is decreased with decreasing SNRs. In terms of nonspeech detection rates, G.729 yields the lowest rates followed by AMR1. The nonspeech detection rates of the proposed algorithm are the highest although AMR2 achieves improved rates over G.729 and AMR1. The proposed method achieves significantly the lowest error rates (10.01%, 12.04%, and 19.05%) for SNRs of 20 dB down to 0 dB. Error rates of AMR2 are found to be dependent on the noise levels, although it offers moderate nonspeech detection rates over G.729 and AMR1.

3.3. *Computational considerations*

The proposed algorithm is implemented in Matlab whereas the other algorithms are implemented using C. The average execution time of the proposed algorithm, G. 729, AMR I, and AMR II running on Pentium IV (2.4 GHz) with 512 MB RAM are 4.265 s, 2.413 s, 7.353, and 7.316 s, respectively. The minimum processing time of these algorithms are also found as 3.563 s, 2.047 s, 5.594 s, and 5.625 s. The maximum execution time of the proposed algorithm is 5.156 s and that of G.729, AMR I, and AMR II are measured as 2.875 s, 9.734 s, and 9.5 s. It is found that although the proposed algorithm is implemented in Matlab, it takes the least computational time except the G.729 algorithm.

4. RECEIVER OPERATING CHARACTERISTICS ANALYSIS

In this section, the detectability and discriminability of the proposed method is verified in terms of receiver operating

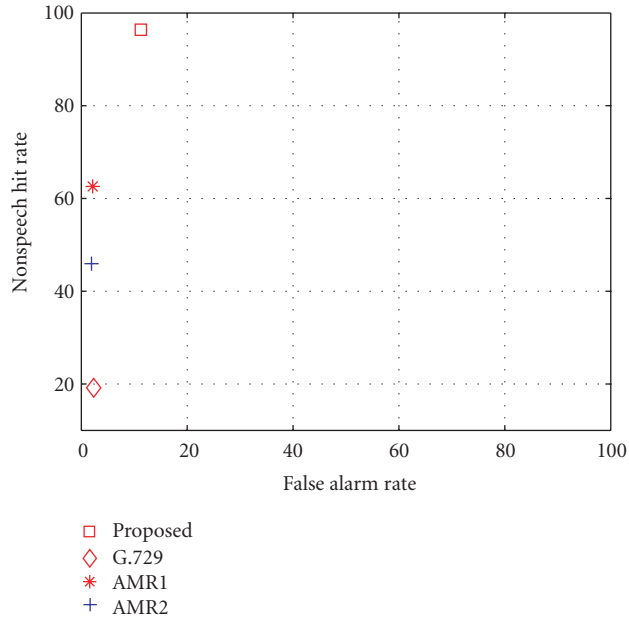

Figure 10: Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1, and AMR2 at 20 dB with car noise.

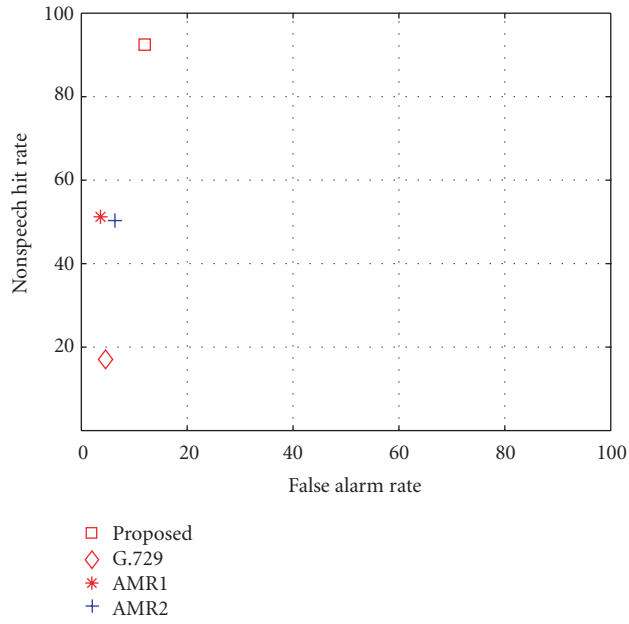

Figure 11: Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1, and AMR2 at 10 dB with car noise.

characteristics (ROC) analysis. In signal detection, the relationship between detection and false alarm probabilities is often characterized by ROC curves. Only the subset of speech database in car noise, as described in Section 3, is used in this ROC analysis. Figures 10, 11, and 12 show the results of ROC analysis at 20 dB, 10 dB, and 0 dB SNRs. For each noise level, nonspeech hit rate (nonspeech detection rate) and false

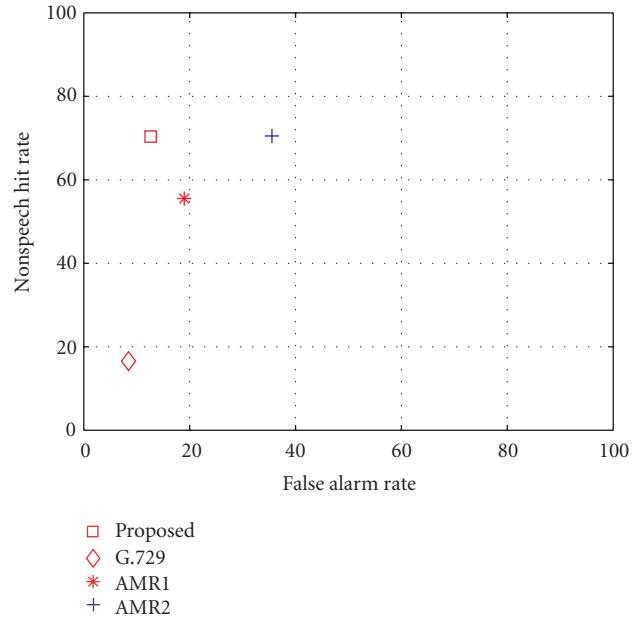

Figure 12: Receiver operating characteristic analysis for proposed method, ITU G.729, AMR1, and AMR2 at 0 dB with car noise.

alarm rate (1-speech detection rate) are determined over the proposed method, G.729, ETSI AMR1, and AMR2. The operating points of G.729, AMR1, and AMR2 shift to the right in ROC plane with decreasing SNRs. However, the operating point of the proposed method can maintain an almost constant false alarm rate.

False alarm rates of AMR2 increases with decreasing SNR although its nonspeech hit rates become higher. Among these standard VADs, G.729 maintains most of the lowest false alarm rates. However, it also has poor nonspeech hit rates for all SNR levels. For more noisy conditions, the nonspeech detectability of AMR2 is better than AMR1. Obviously, the proposed method significantly improves the nonspeech hit rate over the other methods with a nearly constant false alarm rates at changing environments. For a given nonspeech hit rate, the proposed scheme can detect the signal with the lowest false alarm rate. In addition, for a given false alarm rate, the highest nonspeech hit rate can be obtained by our method. From this objective evaluation, it can be concluded that discriminability of the proposed method between speech and noise is found better compared to the standardized methods.

5. CONCLUSION

In this paper, the problem of speech/nonspeech detection in the presence of noise is addressed. A method, which is based on the binary Walsh functions is developed. The basic idea is to reconstruct the noisy speech signal as modified sequences from which speech and nonspeech frames are detected. The main advantage of this method is its very low computational complexity. The Walsh basis functions make the proposed algorithm efficient, simple, fewer parameters to be optimized,

and faster in implementation. Thus the algorithm is applicable in practical situations where processing time is critical. Experimental results indicate that the proposed method can detect speech as well as nonspeech frames with lower error rates across different types of noise with varying SNRs. ROC analysis also shows that the proposed method consistently outperforms G.729, AMR1, and AMR2 in terms of discriminability between speech and noise. Since the computational complexity of the algorithm is relatively low, the algorithm can be applied in the areas such as real time noise cancellation systems and noise reduction for enhancement of speech signals.

ACKNOWLEDGMENTS

The authors would like to thank the Associate Editor and the anonymous reviewers for their useful suggestions that helped to improve this paper. The authors also thank Dr. Anamitra Makur for fruitful discussions.

REFERENCES

[1] ITU-T Recommendation G.729 Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation v.70," 1996.

[2] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1818–1829, 1998.

[3] ETSI, GSM 06.94, "Digital cellular telecommunications system (phase 2+); voice activity detectors (VAD) for adaptive multi-rate (AMR) speech traffic channels; european telecommunications standards institute," 1999.

[4] B. L. McKinley and G. H. Whipple, "Model based speech pause detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 2, pp. 1179–1182, Munich, Germany, April 1997.

[5] J. Sohn, N. S. Kim, and W. Song, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[6] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278, 2001.

[7] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, 2003.

[8] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, 2002.

[9] H. Sheikhzadeh, R. L. Brennan, and H. Sameti, "Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, vol. 1, pp. 808–811, Detroit, Mich, USA, May 1995.

[10] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[11] J. Wei, L. Du, Z. Yan, and H. Zeng, "A new algorithm for voice activity detection," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '03)*, vol. 2, pp. 588–591, Bangkok, Thailand, May 2003.

[12] M. Jelinek and F. Labonté, "Robust signal/noise discrimination for wideband speech and audio coding," in *Proceedings of the IEEE Workshop on Speech Coding*, pp. 151–153, Delavan, Wis, USA, September 2000.

[13] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications*, pp. 85–86, Sainte-Adele, Quebec, Canada, October 1993.

[14] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '89)*, vol. 1, pp. 369–372, Glasgow, Scotland, UK, May 1989.

[15] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, 2000.

[16] Y. Wu and Y. Li, "Robust speech/non-speech detection in adverse conditions using the fuzzy polarity correlation method," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '04)*, vol. 4, pp. 2935–2939, The Hague, The Netherlands, October 2000.

[17] A. Quddus and M. Gabbouj, "Wavelet-based corner detection technique using optimal scale," *Pattern Recognition Letters*, vol. 23, no. 1–3, pp. 215–220, 2002.

[18] D. Arfib, F. Keiler, and U. Zölzer, *DAFX - Digital Audio Effects*, John Wiley & Sons, New York, NY, USA, 2002.

[19] M. Adjouadi, F. Candocia, and J. Riley, "Exploiting Walsh-based attributes to stereo vision," *IEEE Transactions on Signal Processing*, vol. 44, no. 2, pp. 409–420, 1996.