On the Use of Adaptive Fuzzy Wavelet Filter in the Speech Enhancement

Chih-Chia Yao
Department of Computer Science and Information Engineering,
Chaoyang University of Technology, Taiwan
E-mail: ccyao@cyut.edu.tw

Ming-Hsun Tsai and Yuan-Tain Chang
Department of Computer Science and Information Engineering,
Chaoyang University of Technology, Taiwan

***Abstract*—This paper proposes an adaptive fuzzy wavelet filter that is based on a fuzzy inference system for enhancing speech signals and improving the accuracy of speech recognition. In the last two decades, the basic wavelet thresholding algorithm has been extensively used for noise filtering. In the proposed method, adaptive wavelet thresholds are generated and controlled according to the fuzzy rules about the presence of speech in contaminated signals. In this adaptive fuzzy wavelet filter, the relationships between speech and noise are summarized into seven fuzzy rules using four linguistic variables, which are used to determine the state of a signal. A hybrid filter is proposed here, which combines an adaptive fuzzy wavelet filter and the spectral subtraction method to filter contaminated signals. An amplified voice activity detector in the proposed hybrid filter is designed to improve performance when the signal-to-noise ratio (SNR) is lower than 5 dB. The filtering that is performed using the adaptive fuzzy wavelet filter and the spectral subtraction method is controlled by support vector machines. Experimental results demonstrate that the proposed system effectively increases the SNR and the speech recognition rate.**

***Index Terms*—speech enhancement; wavelet thresholding; fuzzy; voice activity detection, spectral subtraction, support vector machines;**

I. INTRODUCTION

Speech enhancement is a continuing challenge in the field of speech and signal processing, particularly in applications such as such as mobile phone systems, speech recognition applications, hearing aid systems, and speech coding [1, 2, 3]. Although speech enhancement algorithms have been studied extensively in the past two decades, enhancement algorithms generate distortions of original speech signals and residual noise in the form of musical tones [4, 5, 6].

The main objective of speech enhancement is to improve the quality and intelligibility of the signal, as perceived by human listeners. The objectives of speech enhancement algorithms vary among applications. The major goal of an automated speech recognition system is to increase the recognition rate whereas that of a communication system is to optimize the signal-to-noise ratio (SNR) of distorted speech [7, 8]. Proposed methods of achieving both goals can be roughly classified as digital signal processing and statistical analysis. Digital signal processing usually removes an estimate of the distortion from contaminated signals such as by spectral subtraction, whereas statistical analysis uses statistical modeling to predict structures and patterns in the signal process [10, 11].

existence of noise are ambiguous [16, 17, 18]. Therefo A spectral subtractive algorithm is based on obtaining the best possible estimates of short term spectra of a speech signal from a given contaminated speech signal. This approach involves estimating the power spectral density of a clean speech signal by subtracting the power spectral density of the noise from that of the contaminated signal. The main appeal of a spectral subtractive algorithm is its simplicity of implementation and its ability to accommodate varying subtraction parameters.

However, spectral subtractive algorithms have various shortcomings, such as imprecision of the estimation of the signal and noise parameters and mismatched probability distribution models of speech and noise. Subtractive denoising methods introduce musical residual noise, arises owing to nonlinear signal processing, leading to a serious deterioration of sound quality. Hence, various methods of suppressing musical noise have been developed [12, 13, 14].

Wavelet-based techniques using coefficient thresholding approaches are extensively utilized to shrink signals and remove noise [15]. Because of its flexible time-frequency resolution, the wavelet transform is an effective tool for analyzing signals that consist of short high-frequency bursts and long quasi-stationary segments. Although an adaptive thresholding algorithm improves the performance, proposed algorithms for generating the adaptive threshold between speech and noise are problematic because the model between speech and noise remains unclear, and the rules for determining the re, this paper proposes an adaptive fuzzy wavelet filter that is based on fuzzy rules and improves speech signal enhancement and the accuracy of automatic speech

doi:10.4304/jcp.9.11.2501-2513

recognition. In this system, relationships between speech and noise are described by seven fuzzy rules which determine adaptive thresholds. Moreover, the design of the filter is optimized by particle swarm optimization (PSO) to maximize the SNR of its output. Membership functions that represent the rules are not required to be obtained in advance.

This paper proposes a novel hybrid filter that is composed of an adaptive fuzzy wavelet filter and a spectral subtraction method for denoising contaminated signals. In this hybrid filter, input signals are classified as either speech or non-speech segments using an amplified voice activity detector, which performs a full wavelet packet transform to decompose the input speech signal into critical sub-band signals [19, 20]. In each critical sub-band signal, a mask is constructed by smoothing the Teager energy operator and the entropy of the corresponding wavelet coefficients. A corresponding adaptive wavelet threshold is then applied to each sub-band signal. After the entropy is incorporated with the Teager energy operator, the amplified voice activity detector enhances the discrimination of signals with an SNR that is lower than 5dB.

The hybrid filter simultaneously filters the contaminated signals using the adaptive fuzzy wavelet filter and the spectral subtraction method. The filtering behavior of the adaptive fuzzy wavelet filter and the spectral subtraction method is controlled using support vector machines (SVMs) [21]. Both the spectral subtraction method and the adaptive fuzzy wavelet filter perform excellently performance in signal de-noising. However, the spectral subtraction method is ineffective when applied to low SNR signals, and the adaptive fuzzy wavelet filter is ineffective when applied to high-frequency signals [22]. The proposed model not only preserves the advantages of the adaptive fuzzy wavelet filter and the spectral subtraction method but it has none of their limitations.

Establishing critical parameters enhances speech detection in noisy environments. In previous studies, speech has been distinguished from background noise by analyzing parameters such as energy, the zero crossing rate, time duration, linear prediction coefficient, linear prediction error energy and pitch information. However, these parameters are difficult to apply to variable-level background noise, even when complex decision strategies are used. In this paper, four parameters, used as linguistic variables were incorporated into fuzzy inference system to detect speech in a noisy environment. The four parameters were energy, zero crossing rate, average residual, and standard deviation. The excellent speech recognition performance of this filter was confirmed by testing it on eight types of noise.

The previous version of this paper has published in 2010 [23]. In the previous version, the adaptive fuzzy wavelet filter is proposed to improve the performance of speech enhancement. However, the adaptive wavelet thresholding method is ineffective for high-frequency signals. For overcoming this shortcoming, spectral subtraction method was introduced to establish a hybrid filter which filtering behavior is controlled using support vector machines. Another hybrid filter has appeared in our previous research in which adaptive wavelet filter and spectral subtraction method were proposed as the pre-filter and microphone array is used as the post-filter [24]. The filtering behavior was controlled by a feed-forward fuzzy neural network. Adaptive wavelet filter and spectral subtraction method were cooperated well on denoising contaminated signals. For providing better performance, in this paper adaptive fuzzy wavelet filter was adapted to cooperate with spectral subtraction method to establish a novel hybrid filter. Morever, there are some improvements on featuer selection and signal mixed controller.

The remainder of this paper is organized as follows. In Section 2 we review the basic concept of wavelet packet transform and decomposition, particle swarm optimization, spectral subtraction method and support vector machines. In Section 3 the adaptive fuzzy wavelet filter is introduced. In Sections 4 we introduce the hybrid wavelet-spectral filter for speech enhancement. In Section 5 a performance evaluation of the proposed system is presented and comparisons with other protocols are made. Our conclusions are made in Section 6.

II. BASIC CONCEEPTS

A. Wavelet Packet Transform and Decomposition

Wavelet transform is intensively used in various fileds of signal processing because processing signals in the frequency domain is often easier to implement [19, 20]. It has the advantage of using variable size time-windows for different frequency bands. This results in a high frequency-resolution in low bands and low frequency-resolution in high bands. The continuous wavelet transform (CWT) of a signal $x(t)$ is given as follows:

$$X_{CWT}(a,\tau)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}x(t)\psi\left(\frac{t-\tau}{a}\right)dt \qquad (1)$$

where τ and a represent the time shift and scale variables, respectively, and $\psi(\cdot)$ is the mother wavelet chosen for the transform.

In the discrete version, the wavelet decomposes the signal with variable frames to perform multi-resolution analysis in a dyadic form known as discrete wavelet transform (DWT). In DWT the scale and translation parameters of the discrete wavelet family are given by $a=2^m$ and $\tau=n2^m$.

The main advantage of wavelet is that they have a varying window size, wide for slow frequencies and narrow for fast, thus leading to an optimal time-frequency resolution in all the frequency ranges. However, slow varying components can only be identified over long time intervals while fast varying components can be identified over short time intervals. A further generalization of the DWT is the wavelet packet transform (WPT) that offers a richer range of possibilities in signal analysis. In WPT the decomposing process is iterated on both high and low frequency components rather than continuing only on low frequency terms as with a standard DWT.

The principle of wavelet packet transform is that, given a signal, a pair of low pass and high pass filters is used to yield two sequences to capture different frequency sub-band features of the original signal. The depth of the wavelet packet tree shown in Fig. 1 can be varied over the available frequency range, resulting in configurable decomposition. The two wavelet orthogonal bases are defined as

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k-2^j n) \tag{2}$$

$$\psi_{j+1}^{2p+1}(k) = \sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k-2^j n) \tag{3}$$

where $h[n]$ and $g[n]$ denote the low-pass and high-pass filters, respectively. $\psi(n)$ is the wavelet function and parameters j and p are the number of decomposition levels and nodes, respectively.

B. Spectral Subtraction Algorithm

Spectral subtraction is a signal processing method in frequency domain that is applied widely [14, 25]. The noisy speech $y(k)$ is assumed to consist of the clean speech $x(k)$ additively degraded by uncorrelated random noise $n(k)$, as follows:

$$\begin{aligned} &\text{Time domain:} && y(k) = x(k) + n(k) \\ &\text{Frequency domain:} && Y(\omega_k) = X(\omega_k) + N(\omega_k) \\ &\text{or} && Y_k e^{j\phi_{y,k}} = X_k e^{j\phi_{x,k}} + N_k e^{j\phi_{n,k}} \end{aligned} \tag{4}$$

where $Y(\omega_k)$, $X(\omega_k)$, and $N(\omega_k)$ are discrete Fourier transforms (DFT's), with amplitudes Y_k, X_k, and N_k, and phases $\phi_{y,k}$, $\phi_{x,k}$, and $\phi_{n,k}$, respectively, at frequency or frequency channel .

The short-time power spectrum of the noisy speech can be approximated by

$$\left|Y_k e^{j\phi_{y,k}}\right|^2 \approx \left|X_k e^{j\phi_{x,k}}\right|^2 + \left|N_k e^{j\phi_{n,k}}\right|^2 \tag{5}$$

The term $\left|N_k e^{j\phi_{n,k}}\right|^2$ can not be obtained directly and is approximated as $E\left\{\left|N_k e^{j\phi_{n,k}}\right|^2\right\}$, where $E\{\cdot\}$ denotes the expectation operator. Typically, $E\left\{\left|N_k e^{j\phi_{n,k}}\right|^2\right\}$ is estimated during non-speech activity, and is denoted by $\left|\widetilde{N}_k e^{j\phi_{n,k}}\right|^2$. Thus, the estimate of the clean speech power spectrum, denoted as $\left|\widetilde{X}_k e^{j\phi_{x,k}}\right|^2$, can be obtained by

$$\left|\widetilde{X}_k e^{j\phi_{x,k}}\right|^2 = \left|Y_k e^{j\phi_{y,k}}\right|^2 - \left|\widetilde{N}_k e^{j\phi_{n,k}}\right|^2 \tag{6}$$

Berouti *et al.* proposed an important variation of spectral subtraction for reduction of residual musical noise [13]. An overestimate of the noise power spectrum is subtracted and the resulted spectrum is limited from going below a preset minimum level. The proposed algorithm could be expressed as

$$\left|\widetilde{X}_k e^{j\phi_{x,k}}\right|^2 = \begin{cases} \left|Y_k e^{j\phi_{y,k}}\right|^2 - \alpha\left|\widetilde{N}_k e^{j\phi_{n,k}}\right|^2, & \text{if } \left|\widetilde{X}_k e^{j\phi_{x,k}}\right|^2 > \beta\left|\widetilde{N}_k e^{j\phi_{n,k}}\right|^2 \\ \beta\left|\widetilde{N}_k e^{j\phi_{n,k}}\right|^2, & \text{otherwise} \end{cases} \tag{7}$$

where α is the subtraction factor and β is the spectral parameter.

The enhanced speech spectrum is obtained using the magnitude estimate $\widetilde{X}_{k,\alpha}^{\alpha}$ of the enhanced speech and the noisy phase $\phi_{y,k}$:

$$\widetilde{X}_k e^{j\phi_{x,k}} = \left|\widetilde{X}_k e^{j\phi_{x,k}}\right| e^{j\phi_{y,k}} \tag{8}$$

To reduce the speech distortion caused by large values of α, its value is adapted from frame to frame. The basic idea is to take into account that the subtraction process must depend on the segmental noisy signal to noise ratio (NSNR) of the frame, in order to apply less subtraction with high NSNRs and vice versa.

The segmental noisy signal-to-noise ratio NSNR is calculated for every frame is obtained as

$$NSNR = 10\log_{10}\frac{\sum_{k=0}^{N-1}\left|Y_k e^{j\phi_{y,k}}\right|^2}{\sum_{k=0}^{N-1}\left|\widetilde{N}_k e^{j\phi_{y,k}}\right|^2} \tag{9}$$

The over-subtraction factor α can be calculated as

$$\alpha_i = \begin{cases} 1, & NSNR_i \geq 20\ dB \\ \alpha_0, & -6\ dB \leq NSNR_i < 20\ dB \\ 4.9, & NSNR_i < -6\ dB \end{cases} \tag{10}$$

where $\alpha_0 = 4$ is the desired value at 0 dB NSNR.

C. Particle Swarm Optimization

In 1995, Kennedy and Eberhart introduced the particle swarm optimization algorithm (PSO) into the field of social and cognitive behavior [26]. Traditionally the main problem in designing a neural fuzzy system is training the parameters. Backpropagation training is commonly adopted to solve this problem. However the steepest descent approach, commonly used in backpropagation training to minimize the error function, may reach the local minima very quickly and never find the global solution. Accordingly, a new optimization algorithm, called particle swarm optimization (PSO), appears to provide better performance than the backpropagation algorithm.

Like other population-based optimization approaches PSO is initialized with a swarm of random solutions, each swarm consists of many particles. Each particle is characterized by its current position $\vec{x}_i = [x_i^1, x_i^2, \cdots, x_i^D]$ and current velocity $\vec{v}_i = [v_i^1, v_i^2, \cdots, v_i^D]$, where D stands for the dimensions of the solution space. In the PSO the trajectory of each particle in the search space is adjusted by dynamically altering the velocity of each particle. Then the particles rapidly search the solution space using the moving velocity of each particle. Each of these particle positions is scored to obtain a fitness value based on how to define the solution of the problem. During the evolutionary process the velocity and position of particle i are updated as

$$\begin{aligned} \vec{v}_i = \omega\times\vec{v}_i &+ \varphi_1\times rand_1()\times\left(NBest_i - \vec{x}_i\right) \\ &+ \varphi_2\times rand_2()\times\left(GBest_i - \vec{x}_i\right) \end{aligned} \tag{11}$$

$$\vec{x}_i = \vec{x}_i + \vec{v}_i \tag{12}$$

where ω is the inertia weight, φ_1 and φ_2 are the acceleration coefficients, respectively. The second term in Eq. (11), called the cognitive component, reflects the experience of a particle since it is dependent on the best position of the respective particle. The third term is referred to as the social component and contains the information of a social group due to the dependence on the neighborhood best position. The random numbers $rand_1()$ and $rand_2$ are chosen from the interval $U(0,1)$. In Eq. (11), *GBest* is the position with the best fitness found so far for the *i*th particle, and *NBest* is the best position in the neighborhood. The term $\vec{v}_i$ is limited to the range $\pm\vec{v}_{\max}$. If the velocity violates this limit, then it is set to the actual limit. Changing the velocity enables each particle to search around its individual best position and global best position. After initialization of the positions and velocities of the particles update equations are applied to every particle in each iteration until a stopping criterion is fulfilled.

D. Support Vector Machines

Consider the training samples $\{(x_i, d_i)\}_{i=1}^N$, where x_i is the input pattern for the *i*th sample and d_i is the corresponding desired response; $x_i \in R^m$ and $d_i \in \{-1,1\}$. The objective is to define a separating hyperplane which divide the set of examples such that all the points with the same class are on the same sides of the hyperplane.

Let w_o and b_o denote the optimum values of the weight vector and bias, respectively. Correspondingly, the optimal separating hyperplane, representing a multidimensional linear decision surface in the input space, is given by

$$w_o^T x + b_o = 0 \tag{13}$$

The set of vectors is said to be optimally separated by the hyperplane if it is separated without error and the margin of separation is maximal. Then, the separating hyperplane $w^T x + b = 0$ must satisfy the following constraints:

$$d_i(w^T x_i + b) \geq 1, \quad i = 1,2,\cdots,N \tag{14}$$

Extending to the non-separable case requires a slack variable ξ_i, to be introduced, to measure the deviation of a data point from an ideal value which would yield pattern separability. Hence the constraint of Eq. (14) is modified to,

$$\begin{aligned} &d_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1,2,\cdots,N \\ &\xi_i \geq 0 \end{aligned} \tag{15}$$

According to Eq. (14), the optimal separating hyperplane is the maximal margin hyperplane with the geometric margin $2/\|w\|$. Hence the optimal separating hyperplane is the one that satisfies Eq. (15) and minimizes the cost function,

$$\Phi(w) = \frac{1}{2} w^T w + C\sum_{i=1}^N \xi_i \tag{16}$$

Since the cost function is a convex function, a Lagrange function can be used to minimize the constrained optimization problem and the optimal weight vector is given by,

$$w_o = \sum_{i=1}^N \alpha_i d_i x_i \tag{17}$$

Classical Lagrangian duality enables the primal problem to be transformed to its dual problem. The dual problem of Eq. (16) is reformulated as

$$Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j d_i d_j x_i^T x_j \tag{18}$$

with constraints,

$$\begin{aligned} &\sum_{i=1}^N \alpha_i d_i = 0 \\ &0 \leq \alpha_i \leq C, \ i = 1,2,\cdots,N \end{aligned} \tag{19}$$

III. ADAPTIVE FUZZY WAVELET FILTER

This section proposes an adaptive fuzzy wavelet filter based on a fuzzy inference system. In the filter, the adaptive threshold for each sub-band signal was generated using the fuzzy inference system, and the noisy components were removed by thresholding the wavelet coefficients.

Let $s(t)$ be clean speech with a finite length and $n(t)$ be noise. Contaminated speech $y(t)$ can then be expressed as

$$y(t) = s(t) + n(t) \tag{20}$$

If W denotes the wavelet transform matrix, then Eq. (20) can be written in the wavelet domain as

$$Y(t) = S(t) + N(t) \tag{21}$$

where $Y(t) = W \cdot y(t)$, $S(t) = W \cdot s(t)$, and $N(t) = W \cdot n(t)$. The estimated speech signal $\hat{S}(t)$ can be obtained by using the thresholding function

$$\hat{S}(t) = F_T(Y, T) \tag{22}$$

where $F_T(Y, T)$ denotes the thresholding function and T is the threshold. The standard thresholding function includes the soft thresholding function, which is defined as

$$F_{T_S}(Y, T) = \begin{cases} sign(Y)(|Y| - T), & |Y| \geq T \\ 0, & |Y| < T \end{cases} \tag{23}$$

In the past two decades, many studies have applied the thresholding function for speech enhancement. Because of the difficulty of modeling speech signals, heuristic algorithms and learning machines, such as neural network models and support vector machines, are often used to determine the threshold [27]. Therefore, ambiguities arise in which the behavior and function of the learning machine are difficult to distinguish. To overcome this problem, this study proposes an adaptive threshold decision module based on a fuzzy inference system.

The adaptive threshold decision module based on a fuzzy inference system produces the adaptive threshold $T(k+1)$ at the $(k+1)$th frame as follows:

$$T(k+1) = (1 - \alpha(k+1)) \times T(k) + \alpha(k+1) THS(k+1) \tag{24}$$

In Eq. (24), $THS(k+1)$ is the threshold used in the $(k+1)$th frame and is generated from fuzzy wavelet filter. Besides,

$\alpha(k+1)$ is the step length for the threshold variation. In most case, dramatic change is seldom appeared in voice signal. Hence, $\alpha(k+1)$ denotes the membership function indicating to what extent a dramatic change is considered to be happened at (k+1)th frame. Since it is difficult to judge whether a dramatic change is happened or not, α (k+1) should take a continuous value from 0 to 1 to cope with ambiguous cases.

A. Linguistic Variable for Fuzzy System

In the fuzzy inference system, four linguistic variables are used to detect signal state. They are: energy, the zero crossing rate, standard deviation of formants and average residual. The reasons for adopting those four linguistic variables are listed as follows.

-Energy:

Energy is an effective factor on measuring the degree of noise when the SNR is bigger than 0dB. The short time energy of speech signals reflects the amplitude variation and is defined as

$$En = \sum_{m=-\infty}^{\infty} S^2(m) \cdot h(n-m) \tag{25}$$

In order for E_n to reflect the amplitude variations in time (for this a short window is necessary), and considering the need for a low pass filter to provide smoothing, $h(n)$ was chosen to be a hamming window powered by 2.

Studies have shown that energy provides excellent performance on voice activity detection. However, rapid variation of energy in the speech model causes implementation difficulties.

-Zero crossing rate:

The zero crossing rate (ZCR) is the number of zero crossing of a waveform within a given frame [28]. The ZCR of both unvoiced sounds and environment noise generally exceed that of voiced sounds. A Zero Crossing Rate can be calculated by the mathematical formula:

$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1} \left| Sgn[S(n)] - Sgn[S(n-1)] \right| \tag{26}$$

where $Sgn[S(n)] = 1$ if $S(n) \geq 0$; otherwise $Sgn[S(n)] = -1$.

ZCR is often used in conjunction with volume for end-point detection. It is hard to distinguish unvoiced sounds from environment noise by using ZCR alone since they have similar ZCR values.

-Standard deviation of formants:

Standard deviation is a measure of how wide any given numbers are spread. It is useful in comparing sets of data, which may have the same mean but different range. In a given frame standard deviation of formants is helpful to distinguish the signal's mode. The standard deviation of formants of a stable noise is different to voice sound. The observation of standard deviation of formants is useful in the threshold decision level.

-Average residual:

The statistical property of average residual is similar to standard deviation. In this paper standard deviation is used to detect the variation within a frame but average residual is used to compare the difference between frames. The formula for Energy is

$$Ar = \sum_{n=0}^{N-1} \left| S(n) - \bar{S} \right| \tag{27}$$

where $\bar{S}$ is the mean value.

By calculating the average residual across several frames the signal's mode can be distinguished, whether the variation is temporal or not, in a very short time interval. In this study the average residual is calculated based on the standard deviations of previous frame, current frame and next frame.

$\alpha(k)$ and $THS(k)$ are decided by the four parameters. The threshold range is first determined according to energy and the zero crossing rate. Previous studies have indicated that energy is an accurate measure of noise exceeding 0 dB. However, a threshold based only on energy obtains irrational results when energy rapidly varies. Additionally, because most mechanisms use an 8 k sample rate to record the speech signal, energy is an ineffective measure when the frequency of the noise exceeds the sample rate. Overcoming the zero crossing rate problem requires measuring energy when determining the threshold.

As noted previously, the ZCR of unvoiced sounds and environmental noise exceed those of voiced sounds. Many studies have used a zero crossing rate to detect voice activity because it easily distinguishes signal modes. For high-frequency noise and low-frequency speech, the zero crossing rate can be considered as an auxiliary parameter for determining the threshold. When the frequency of noise is low, the threshold depends mainly on energy. As the frequency of noise increases, the zero crossing rate must be considered. In short, the importance of the zero crossing rate to the threshold decision is proportional to the frequency of noise.

Here, the standard deviation of formants and average residual are introduced to control threshold variation. The variation in the model of signals between two adjacent frames is reflected in the standard deviation of formants. As the standard deviation of formants increases, the threshold variation increases. However, the correlation differs when speech is contaminated by an instantaneous impulse signal, which, in a given frame, usually increases the standard deviation of formants and violently vibrates the threshold. To avoid this problem, the average residual of any adjacent three frames is calculated to smooth the threshold variation. The average residual is calculated according to the standard deviations of formants between three adjacent frames and can effectively reduce the impact of instantaneous impulse signals. An increase in the value of the average residual is a strong indication that it is neither temporal nor instantaneous.

B. Fuzzy Inference Rules

Let $Y(k)$ denotes the input speech signal, where k is the frame index and $1 \leqq k \leqq N$. $En(k)$, $ZCR(k)$, $Sd\_f(k)$, $Ar(k)$ denote the speech features that the input signal $Y(k)$ is processed by the energy function, zero crossing rate function, standard deviation of formants and average residual function, respectively.

The threshold can be set by the local characteristics of the input signals but, as noted previously, constructing the noise model based on the contaminated signal is difficult in ambiguous cases. The proposed speech enhancement system involves adopting a fuzzy inference system to address these problems and generates an adaptive threshold to suppress noise. The following discussion summarizes the various models of noise and the relationships between noise and speech signals. The fuzzy sets of linguistic variables and their corresponding linguistic terms are then defined according to the analysis. Table 1 shows the relationships between linguistic variables and their fuzzy sets.

After eliminating cases in which noise is indistinguishable from speech signals, seven fuzzy rules are proposed, listed as follows.

Rule 1: If En(*k*) *is* ***High*** *and ZCR*(*k*) *is* ***Low****, then THS*(*k*) *is* ***Medium.***

Energy is an effective measure of noise when SNR exceeds 0dB. In most cases, speech signals have higher energy compared to noise. To retain speech while suppressing noise, the threshold can be set lower than the speech signal. This rule is applicable in places like airports or streets with heavy traffic. In this model the zero crossing rate is classified to low when the frequency is between 50Hz and 5 KHz.

Rule 2: If En(*k*) *is* ***High*** *and ZCR*(*k*) *is* ***High****, then THS*(*k*) *is* ***High.***

As the zero crossing rate increases, the probability of noise increases. Since the probability of noise is higher under rule 2 than under rule 1, the threshold can be set higher under rule 2. This rule is applicable when the energy is large and the frequency is high such as noise generated by high frequency electric products or machines. In this model the zero crossing rate is classified to high when the frequency is bigger than 5K Hz.

Rule 3: If En(*k*) *is* ***Low*** *and ZCR*(*k*) *is* ***Low****, then THS*(*k*) *is* ***Low.***

A low zero crossing rate usually indicates that the probability of noise is not high. In most cases, the energy of the speech signal is higher than the energy of noise. Therefore, a low threshold is set when the energy of a contaminated speech signal is low. This type of noise is generated by, for example, low-frequency electric products (e.g., indoor fans).

Rule 4: If En(*k*) *is* ***Low and*** *ZCR*(*k*) *is* ***High****, then THS*(*k*) *is* ***Medium.***

On the basis of rule 4, high zero crossing rate denotes a high probability of noise. The threshold should be bigger than in rule 3 but not exceed the energy of speech signal. This rule is applicable on the environments such as the electromagnetic wave.

Rule 5: If Sd_f(*k*) *is Low and Ar*(*k*) *is* ***Low****, then* $\alpha(k)$ *is* ***Low.***

This rule is applicable when the noise model is stable, whether within a short interval or within a long interval. This noise is characterized by a small standard deviation of formants and a small average residual. The average residual is calculated based on the standard deviation of formants of the previous frame, current frame, and next frame. Because the standard deviation of formants is small, the threshold must be low. The waveform in this situation is shown in Fig. 2, in which the vertical axis represents amplitude, and the horizontal axis represents time. In addition, each frame in Fig. 2 is bounded by a rectangle. The standard deviation of formants is small because the waveform within each frame is stable. The average residual of the standard deviation of formants is also small because standard deviations of formants in each frame are all small.

Rule 6: If Sd_f(*k*) *is* ***High and*** *Ar*(*k*) *is* ***Low****, then* $\alpha(k)$ *is* ***Medium.***

This rule is applicable when noise is stable varies within a very short time interval. This noise is characterized by a high standard deviation of formants, but the average residual is small. Compare to the standard deviation of formants under rule 5, a high standard deviation of formants under rule 6 indicates the amplitudes of noise are changed rapidly. Hence, the variation of threshold should be set higher than that in rule 5. In cases of multiple sources of noise, this rule can bring the function into full play. The waveform under this situation is shown in Fig. 3. In Fig. 3, a large variation in the waveform within the middle frame causes a large standard deviation. The average residual of the standard deviation of formants is small since the standard deviations of formants are large.

Rule 7: If Sd_f(*k*) *is* ***High*** *and Ar*(*k*) *is* ***High****, then* $\alpha(k)$ *is* ***High.***

This rule is applicable when noise is unstable, no matter in a very short time interval or in a long time interval. In rule 7, the standard deviation of formants for the current frame is high and is higher than that of neighboring frames, so that the average residual calculated from the standard deviation of formants for the current and neighboring frame is high. The waveform under this situation is shown in Fig. 4.

After the fuzzy If-Then inference rules are applied, the fuzzy wavelet filter should be defuzzified. One of the commonly used defuzzification method is the centroid defuzzification method:

$$O = \frac{\sum_i \mu_A(y_i) y_i}{\sum_i \mu_A(y_i)} \tag{28}$$

where $\mu_A(y_i)$ is the membership function.

C. Setting the Membership Function

The next step is to design this function *THS*(*k*) and $\alpha(k)$ for the energy, zero crossing rate, standard deviation and average residual. In most fuzzy systems, the membership function is obtained by using fuzzy approximate reasoning and the membership function of the *Low* and the *High* of energy, zero crossing rate, standard deviation of formants, average residual, and its relationship to *THS*(*k*) and $\alpha(k)$. In this study, the Gaussian type (Eq. 29) and sigmoid function (Eq. 30) were used to approximate the membership function and achieve the benefit of efficient calculation.

$$f(x) = e^{-\frac{(x-\mu)^2}{2c^2}} \tag{29}$$

$$f(x) = \frac{1}{1+e^{-\alpha x}} \tag{30}$$

In supervised learning, the output is as close to the desired signal as possible; thus some optimization is required in designing this filter. The PSO algorithm is applied to optimize the parameters in these nonlinear functions. Based on research, the PSO algorithm provides higher performance in determining the global solution while optimizing the overall structure. In the training phase an iterative training algorithm proposed by this paper includes the following six steps:

Step 1: Initially set *k*=1.

Step 2: For each input speech signal *Y*(*k*), calculating *En*(*k*)、*ZCR*(*k*)、*Sd_f*(*k*) with *Ar*(*k*).

Step 3: Using rules 1-4 to generate *THS*(*k*) and rules 5-7 to generate $\alpha(k)$.

Step 4: Calculating adaptive threshold *T*(*k*+1) by using Eq. (24).

Step 5: Let *O*(*k*) be the output of adaptive wavelet filter at the *k*th frame. Using Eq. (28) to get the value of *O*(*k*), where $\mu_A(y_i)$ is the mathematical form of Eq. (29) or Eq. (30).

Step 6: Setting $\min \sum_{k=1}^{N} |S(k) - O(k)|^2$ be the cost function. Using PSO algorithm the find the optimal solution and adjusting the parameters of membership functions. Repeat step 2 to step 6 until optimal solution is reached.

IV. HYBRID WAVELET-SPECTRAL FILTER

This section proposes a new framework for speech enhancement called the "hybrid wavelet-spectral filter". Figure 5 shows the schematic diagram of the hybrid wavelet-spectral filter. The structure of a hybrid wavelet-spectral filter comprises an adaptive fuzzy wavelet filter, a spectral subtractive module, a voice activity detection module, an SVM controller, and a feature selection module.

In the hybrid wavelet-spectral filter, an amplified voice activity detection module determines the presence or absence of speech. The amplified voice activity detection module outperforms existing methods regarding low SNR. Signals are then simultaneously filtered by the adaptive fuzzy wavelet filter and spectral subtraction filter. The filtering behavior between the adaptive fuzzy wavelet filter and the spectral subtraction method is controlled using SVMs.

A. Definition of Hybrid Wavelet-Spectral Filter

Additive noise can be classified as stationary or non-stationary. The spectral subtraction method and wavelet filter are known to perform effectively for stationary noise and non-stationary noise, respectively. However, the spectral subtraction method is ineffective for low SNR signals and may produce residual noises. The adaptive wavelet thresholding method is ineffective for high-frequency signals because it can cause distortion. In proposed filters, contaminated signals are simultaneously filtered by both the adaptive fuzzy wavelet filter and spectral subtractive method. Ideally, the filtering algorithm should vary systematically by frame according to local information. However, setting the conditions under which a certain filter should be selected is extremely difficult, if not impossible, because the local conditions can be evaluated only vaguely in some portions of contaminated signals. An SVM trained with a set of input signals and desired signals can function as a desired classifier.

Definition 1: The output of the Hybrid Wavelet-Spectral Filter is defined by

$$y_H(n) = \frac{\mu+1}{2} y_W(n) + \frac{1-\mu}{2} y_S(n) \tag{31}$$

where *n* is the frame index. $y_H(n)$, $y_W(n)$ and $y_S(n)$ are the output signal of Hybrid Wavelet-Spectral filter, spectral subtraction filter and adaptive fuzzy wavelet filter, respectively.

In Eq. (31), the parameter μ is generated by the support vector machine, $\mu \in \{-1,1\}$. If $\mu = -1$, the input signal is considered to be filtered by spectral subtractive filter. If $\mu = 1$, then the input signal is considered to be filtered by adaptive fuzzy wavelet filter.

B. Amplified Voice Activity Detection

Voice activity detection (VAD) is used to distinguish speech from contaminated speech signals and is required in various speech communication systems [16]. To distinguish noise from speech, previous studies have adopted a Teager energy operator (TEO), which has been proven to provide excellent performance in both additive noisy and real noisy environments [29]. The discrete form of the TEO is given by

$$T[y(n)] = y^2(n) - y(n+1)y(n-1) \tag{32}$$

where *T*[*y*(*n*)] is called the TEO coefficient of *y*(*n*). However, a TEO is insensitive when the SNR is low. For example, when the SNR is lower than 0 dB, the difference between noisy energy and speech energy is not obvious, and the performance of the TEO is not satisfactory. Figure 6 shows an example of using TEO to distinguish speech from noise signals contaminated by -5 dB of Gaussian white noise. To overcome this problem, in our proposed module, the TEO is combined with entropy to improve the ability of distinction [30]. The formula of entropy is shown as follows:

$$Ey(S) = \sum_i p(s_i) \log p(s_i) \tag{33}$$

Entropy represents the degree of variation. When the noise model is stationary or slightly non-stationary, the corresponding entropy is kept stable or is slightly changed. Figure 6 illustrates the difference between the output of TEO VAD and entropy VAD. In Fig. 6 entropy VAD provide better solution in time slot 310, 420, and 720. In these time slots speech signals are detected when the threshold of entropy is set to 1000. However, the state of TEO VAD shows there are only noise signals. In Fig. 6, the VAS (voice activity shape) operator is applied to each TEO signals to discriminate the speech and non speech regions. Figure 7 shows the comparison of using the TEO and entropy on VAD to distinguish signals contaminated by -5 dB of Gaussian white noise. Obviously, in Fig. 7 entropy VAD provides more precise voice/unvoice segment than TEO VAD, such as in time slot 3000, 5300, 7500 and the interval between 8000 and 8200. Although the energy of speech signals is reduced gradually and is closed to the energy of noise signals in these time slots, the entropy VAD can distinguish the speech and non speech regions precisely.

The proposed VAD algorithm computes signals $w_m(x_i)$ that have been produced by the wavelet package transform on each input frame to produce 2^J sub-band wavelet packets, where J is the number of levels for the wavelet packet decomposition tree and $1 \le m \le 2^J$. A set of $T[w_m(x_i)]$ can then be derived from Eq. (32). The scheme of the voice activity detection is designed as follows:

$$V_m(x_i) = \begin{cases} 0, & \text{if} \quad \text{var}(T[w_m(x_i)]) < \lambda_J \text{ or} \\ & \text{var}(Ey[w_m(x_i)]) < \varsigma \\ x_i, & \text{otherwise} \end{cases} \tag{34}$$

where ς is user-defined and $\lambda_J = \sigma_J \sqrt{2\log(N_J)}$, as proposed by Johnston and Silverman [31], and $\text{var}(\cdot)$ denotes the variance. In our experiments, ς is set to be 0.458.

C. Feature Selection and Support Vector Machine

In the hybrid wavelet-spectral filter, the filtering behavior between the adaptive fuzzy wavelet filter and the spectral subtraction filter is controlled using an SVM. SVMs have been shown to provide higher performance than traditional learning machines. The advantage of SVMs involves minimizing the risk of misclassifying not only examples in the training set, but also the unseen examples of the test set. Based on research, choosing critical features plays a vital role in the performance of classification. In this part, energy, the zero crossing rate, entropy, and Mel frequency cepstral coefficients were adopted to analyze the presence of speech in noisy environments. Mel frequency cepstral coefficients are described as follows:

■ Mel Frequency Cepstral Coefficient

Mel-frequency cepstral coefficients (MFCCs) are derived from a type of cepstral representation of the audio clip. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. Energy value of each band was calculated by

$$Y(m) = \log\{ \sum_{k=f_{m-1}}^{f_{m+1}} |X(k)|^2 B_m(k)\} \tag{35}$$

where $B_m(k)$ was the triangular bandpass filter of m th band. Taking the discrete cosine transform on the derived energy value from Eq. (35), a Mel frequency cepstral coefficient was then obtained, which can be expressed as follows:

$$MFCC(n) = \frac{1}{M} \sum_{m-1}^{M} Y(m) \cos(\frac{\pi n(m-\frac{1}{2})}{M}) \tag{36}$$

where M was the number of bands.

In training the SVM, supervised class labels of training patterns were decided according to the performance of the adaptive fuzzy wavelet filter and spectral subtraction filter. In the supervised mode, the performance of these two filters can be easily determined by comparing the filtering signal with the desired signal. The class labels were decided systematically by frame. Let $d_i = 1$ if the SNR of the adaptive fuzzy wavelet filter is larger than that of the spectral subtraction filter. Otherwise, $d_i = -1$. By training the SVM with a set of input signals and desired signals, it acquires the function of a desired classifier.

Let $Y(k)$ denotes the input speech signal, where k is the frame index and $1 \leqq k \leqq N$. The training pattern for the support vector machines is defined as $\{(x(k), d(k)\}_{k=1}^{N}$, where $x(k) = [En(k), ZCR(k), Ey(k), MCFF_1(k), MCFF_2(k), \cdots, MCFF_n(k)]^T$ and $d(k) \in \{1, -1\}$. In the definition of $x(k)$, $En(k)$, $ZCR(k)$, $Ey(k)$ represents the energy, the zero crossing rate, entropy, respectively. Moreover, $MCFF_i(k)$ represents the ith Mel Frequency Cepstral Coefficient on the kth frame. In the training phase an iterative training algorithm proposed by this paper includes the following four steps:

Step 1: Initially set k=1.

Step 2: Calculating each training pattern $\{(x(k), d(k))\}_{k=1}^{N}$ for each frame. $d(k)$ = 1 if the SNR of output of adaptive fuzzy wavelet filter is higher than the output of spectral subtraction filter. Otherwise, $d(k)$ = -1.

Step 3: Using smooth support vector machines [32] to transfer the constrained cost function Eqs. (15), (16) into unconstrained cost function and spplying conjugate gradient algorithm to modify the parameters. Repeat step 2 to step 4 until the optimal solution is reached.

V. Experimental Results

This section demonstrates the effectiveness of our proposed system. The experimental results that pertain to the proposed speech enhancement system were compared

to those obtained by the method based on spectral subtraction and the signal subspace approach.

The proposed system is tested to determine its effectiveness in the enhancement and recognition of noisy speech using the Aurora 2 database, which is one of the currently standard databases that is used in the enhancement and recognition of noisy speech [33]. The speech files in the database include recordings of the ten English digits and the 26 English letters, as spoken by males and females. The Aurora 2 database not only includes 4004 uncontaminated spoken sentences; it also includes 48048 spoken sentences that are contaminated with eight types of noise (Subway, Babble, Car, Exhibition, Restaurant, Street, Airport and Station) at -5, 0, 5, 10, 15 and 20dB. In this experiment, the Daubechies wavelet filter with a length of four is adopted [34].

First, 640 test speech signals, the probabilities of detection P_d and the false-alarm P_f are used to evaluate the performance of the proposed VAD algorithm. P_d is calculated as the percentage of test cases in which the hand-marked speech regions are correctly detected by the VAD algorithm while P_f is the percentage of test cases in which hand-marked noise regions are erroneously identified as speech. For a variety of noise sources and SNRs, the P_d and P_f of the proposed algorithm are compared with those of Robust VAD and ARM VAD [28, 35]. Table 2 presents the experimental results, which reveal that the proposed VAD algorithm performs well in terms of a low SNR.

Second, to make consistent comparisons, the HTK speech recognition system is adopted as a classifier [36]. The HTK speech recognition system was developed by the Speech Vision and Robotics Group of Cambridge University. This system can be used to establish an HMM model, a language model, and a training model. In the experiment, HTK has been applied on the data of an Aurora 2 database.

In the experiments, tenfold cross-validation is performed on the data of an Aurora 2 database to evaluate how well each algorithm generalizes to future data [37]. The method of tenfold cross-validation involves extracting a certain proportion, typically 10%, of the training set as the tuning set, which is a surrogate of the testing set. All parameters in the proposed algorithm are set to optimize the performance when applied to the tuning set.

The proposed system is evaluated in two steps. First, the segmental SNRs before and after signal enhancement are evaluated. Next, the HTK classifier is applied to the enhanced speech signals to compare the recognition rates of the adaptive fuzzy wavelet filter, the hybrid wavelet-spectral filter, and other methods. The experimental results are as follows.

Figure 8 plots the waveform before and after signal denoising by the adaptive fuzzy wavelet filter. The signal was mixed with 10 dB of car noise. Figure 9 shows the corresponding time-frequency diagrams.

Comparisons of segmental SNRs confirm the effectiveness of the proposed system. Tables 3, 4 and 5 show the segmental SNRs before and after signal denoising using the adaptive fuzzy wavelet filter and the hybrid wavelet-spectral filter. The experimental results show that the proposed system effectively removes noise.

Finally, Tables 6, 7, and 8 compare the signal recognition rates after denoising by adaptive fuzzy wavelet filter, hybrid wavelet-spectral filter, the signal subspace approach, spectral subtraction method, Avci's and Ghanbari's wavelet method [20, 38, 39, 40]. Again, the experimental results confirm that the proposed method outperforms other methods.

VI. Discussion

The proposed novel wavelet filter controlled by fuzzy rules removes additive noises in contaminated speech signals. The rules for setting system parameters in this adaptive fuzzy wavelet filter are based on the local characteristics of the signals. The system was designed to be optimized using the PSO algorithm to minimize the mean square error of the system outputs. The system uses four linguistic variables and seven fuzzy rules to determine the adaptive threshold. Different forms of noises can be effectively represented and distinguished using the four linguistic variables, and an adaptive threshold is determined for each type of noise.

A hybrid wavelet-spectral filter for denoising contaminated signals is proposed. The hybrid wavelet-spectral filter preserves the advantages of the adaptive fuzzy wavelet filter and the spectral subtraction filter but with none of their limitations. The experimental results herein confirm that the proposed method outperforms other de-noising methods.

Further research should seek additional critical parameters to help to identify the type of noise and to determine more rules to add to this system. Although some studies have demonstrated algorithms that can automatically generate inference rules, the inference time of these algorithms increases with the number of fuzzy rules. The balance between the precision and time complexity of such algorithms remains to be determined.

Acknowledgements

This work is supported by National Science Council of the Republic of China under Grant NSC102-2221-E-324-042.

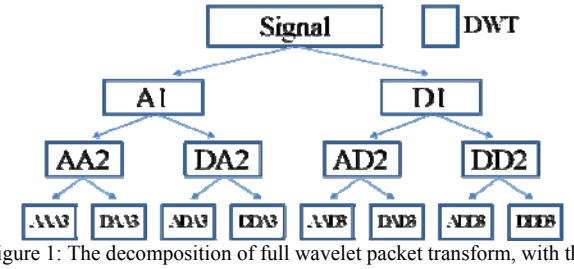


Figure 1: The decomposition of full wavelet packet transform, with the left and right branches at each node representing a matched pair of low-pass and high-pass wavelet filters followed by downsampling.

TABLE 1:
LINGUISTIC VARIABLES - FUZZY SET RELATIONAL TABLE.

Linguistic Variable	Fuzzy Sets		
Energy	Low	High	
Zero crossing rate	Low	High	
Standard deviation of formants	Low	High	
Average residual	Low	High	
Threshold	Low	Median	High
Step length	Low	Median	High

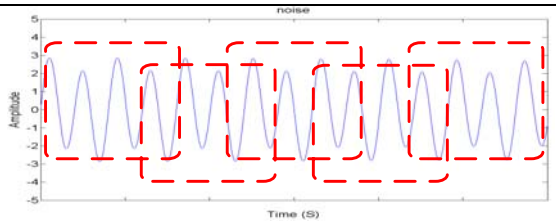


Figure 2: The property of the noise is low standard deviation of formants and low average residual.

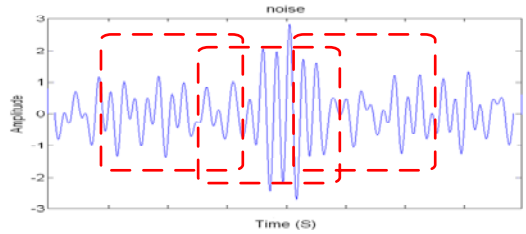


Figure 3: The property of the noise is high standard deviation of formants and low average residual.

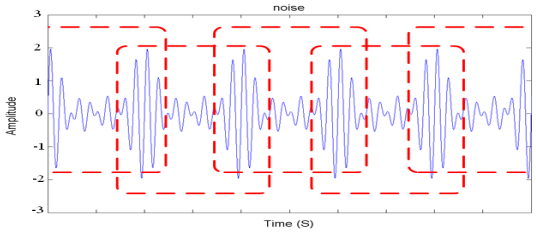


Figure 4: The property of the noise is high standard deviation of formants and high average residual.

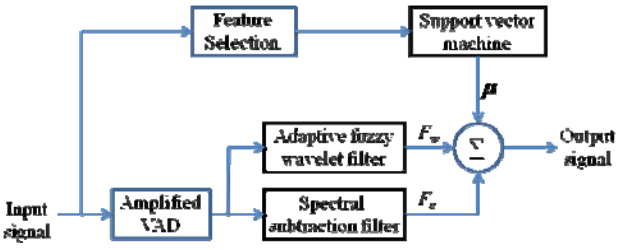


Figure 5: The schematic diagram of hybrid wavelet-spectral filter.

TABLE 2.
PROBABILITY OF DETECTION P_D OF THE PROPOSED VAD WITH OTHER METHODS FOR VARIOUS NOISE CONDITIONS

Environment	Noise	Method					
	SNR (dB)	Proposed VAD		Robust VAD		AMR VAD	
		P_d %	P_f %	P_d %	P_f %	P_d %	P_f %
Station	15	96.3	9.8	96.5	9.6	96.1	19.3
	10	96.1	10.2	96.1	9.9	95.4	28.2
	5	94.5	10.9	94.5	10.5	93.8	37.5
	0	92.8	11.1	92.3	11.3	90.2	43.4
	-5	89.0	13.6	86.2	15.2	85.3	51.5
Street	15	96.5	9.7	96.8	9.1	95.8	25.1
	10	95.8	9.9	96.1	9.3	95.1	29.6
	5	95.1	10.2	95.1	9.8	94.6	38.5
	0	93.7	10.2	93.1	10.2	94.1	45.5
	-5	88.8	13.0	85.8	14.5	90.0	50.3
Car	15	96.5	8.5	96.5	8.2	96.1	21.6
	10	95.2	8.8	95.7	8.9	96.0	29.9
	5	93.0	9.3	93.1	9.6	95.2	40.3
	0	88.2	10.1	86.5	10.8	94.8	47.5
	-5	85.5	13.4	80.5	15.1	91.2	55.6

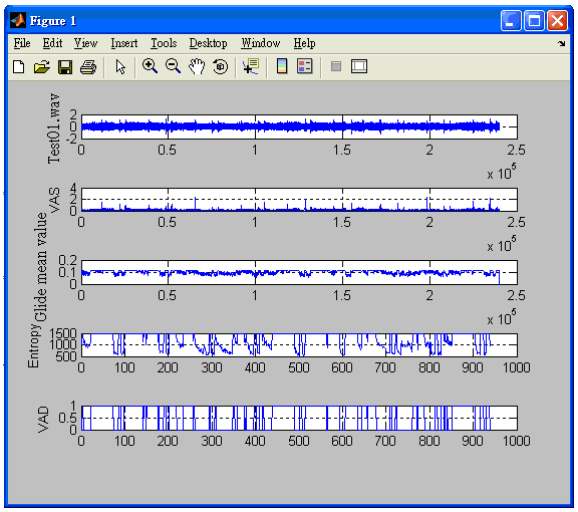


Figure 6: The property of Gaussian white noise with -5dB.

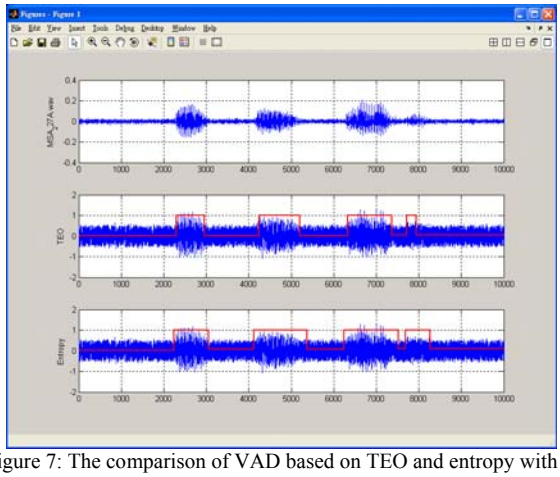
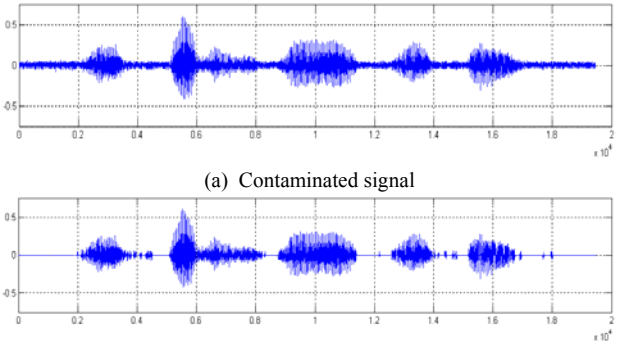


Figure 7: The comparison of VAD based on TEO and entropy with -5dB Gaussian white noise.



(a) Contaminated signal

(b) Enhancement signal

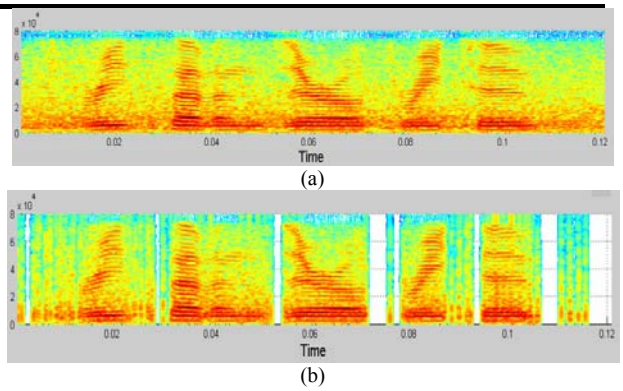
Figure 8 (a) Speech signal contaminated by 10dB car noise. (b) Enhancement signals denoised by adaptive fuzzy wavelet filter.

TABLE 3.
THE SEGMENTAL SNR OF THE SIGNALS IN AURORA 2 DATABASE.

SNR (without enhancing)									
	Sub.	Bab.	Car	Exh.	Res.	Street	Air.	Sta.	AVE
20dB	18.5	18.3	18.4	18.6	18.38	18.32	18.45	18.38	18.41
15dB	13.5	13.6	13.4	13.6	13.45	14.07	14.28	13.55	12.37
10dB	8.85	8.9	8.95	8.92	9.31	9.07	9.28	8.93	7.42
5dB	4.95	4.9	4.98	4.96	4.97	4.95	4.92	4.94	4.94
0dB	0.84	0.62	0.08	0.12	0.11	0.3	0.26	0.09	0.30
-5dB	-4.81	-4.26	-4.33	-4.13	-5.01	-4.86	-4.97	-5.06	-4.67

TABLE 4. THE SEGMENTAL SNR OF THE SIGNALS AFTER ENHANCING BY ADAPTIVE FUZZY WAVELET FILTER.

SNR (after enhancing)									
	Sub.	Bab.	Car	Exh.	Res.	Street	Air.	Sta.	AVE.
20dB	20.1	18.9	20.3	19.7	18.8	19.2	19.4	18.7	19.38
15dB	15.5	14.8	13.4	14.9	13.9	14.98	15.04	14.4	14.33
10dB	11.4	10.3	11.6	11.3	11.88	11.4	11.67	10.94	10.64
5dB	7.4	6.1	8.1	6.6	5.7	6.7	6.4	6.5	6.68
0dB	4.45	3.68	5.3	4.3	3.16	3.77	3.63	3.53	3.97
-5dB	1.33	-1.78	1.64	1.21	0.88	1.16	0.86	1.18	0.81



(a)

(b)

Figure 9 (a) The spectrogram of speech signal contaminated by 10dB car noise. (b) The spectrogram of enhancement signals denoised by adaptive fuzzy wavelet filter.

TABLE 5.
THE SEGMENTAL SNR OF THE SIGNALS AFTER ENHANCING BY HYBRID WAVELET-SPECTRAL FILTER.

SNR (after enhancing)									
	Sub.	Bab.	Car	Exh.	Res.	Street	Air.	Sta.	AVE.
20dB	23.4	20.83	21.83	21.73	21.56	23.49	21.01	22.58	22.05
15dB	19.8	17.74	17.4	16.41	16.84	18.51	17.66	18.26	17.83
10dB	16.16	12.97	13.64	15.63	14.78	13.92	14.49	14.12	14.46
5dB	9.95	7.88	9.78	9.67	8.39	10.62	8.43	9.01	9.21
0dB	4.53	3.68	6.66	6.28	5.04	4.98	5.03	3.53	4.97
-5dB	2.03	1.1	1.76	2.24	2.05	2.13	2.02	1.18	1.81

TABLE 6.
THE COMPARISON OF SIGNAL SUBSPACE WITH ADAPTIVE FUZZY WAVELET FILTER AND HYBRID WAVELET-SPECTRAL FILTER

Recognition rate %									
	method	Sub.	Bab.	Car	Exh.	Res.	Street	Air.	Sta.	AVE.
20dB	subspace	99.3	97.4	98.1	97.8	98.0	97.7	97.6	97.6	97.8
	adaptive	98.1	98.0	98.6	97.6	98.6	97.9	98.3	98.8	98.2
	Hybrid	98.9	98.4	98.7	97.9	98.8	98.7	98.5	99.0	98.6
15dB	subspace	95.5	92.8	96.7	94.5	94.3	93.2	94.2	94.2	94.4
	adaptive	95.7	94.2	96.1	95.4	95.2	94.9	95.6	95.9	95.4
	Hybrid	96.5	95.1	96.7	95.8	95.9	95.6	96.1	96.5	96.0
10dB	subspace	87.7	81.9	90.3	85.3	84.1	83.9	84.5	84.5	85.3
	adaptive	87.9	83.6	82.9	87.8	88.3	85.0	84.9	86.7	85.9
	Hybrid	88.9	84.7	84.7	88.4	88.5	85.6	85.2	87.8	86.7
5dB	subspace	74.7	64.1	74.6	62.9	65.2	67.0	67.5	67.5	67.9
	adaptive	71.2	65.4	72.3	67.1	66.8	64.3	69.8	63.9	67.6
	Hybrid	73.2	66.0	73.8	69.8	68.7	67.3	70.2	67.2	69.5
0dB	subspace	52.2	39.8	45.5	38.3	41.2	39.0	43.7	43.7	42.9
	adaptive	53.3	40.2	44.3	39.1	42.3	41.4	45.1	43.9	43.7
	Hybrid	53.5	40.2	45.8	42.8	43.7	42.8	46.1	43.9	44.9
-5dB	subspace	28.3	21.5	20.3	16.1	17.9	16.8	21.5	21.5	20.5
	adaptive	30.2	24.6	22.3	16.9	18.3	17.8	20.6	19.6	21.3
	Hybrid	30.5	27.9	22.4	17.3	18.8	18.2	21.4	19.6	22.0

TABLE 7.
THE COMPARISON OF SPECTRAL SUBTRACTION WITH HYBRID WAVELET-SPECTRAL FILTER

Recognition rate %									
	method	Sub.	Bab.	Car	Exh.	Res.	Street	Air.	Sta.	AVE.
20dB	spectral	84.8	90.0	89.5	80.6	91.2	91.8	91.4	97.6	93.2
	Hybrid	98.9	98.4	98.7	97.9	98.8	98.7	98.5	99.0	98.6
15dB	spectral	77.2	79.1	81.8	67.8	79.3	86.0	84.8	94.2	90.6
	Hybrid	96.5	95.1	96.7	95.8	95.9	95.6	96.1	96.51	96.0
10dB	spectral	65.7	55.0	66.9	47.6	57.0	67.8	64.9	84.5	77.9
	Hybrid	88.9	84.7	84.7	88.4	88.5	85.6	85.2	87.8	86.7
5dB	spectral	45.5	27.3	38.7	25.1	29.5	43.5	36.1	67.5	47.5
	Hybrid	73.2	66.0	73.8	69.8	68.7	67.3	70.2	67.2	69.5

TABLE 8.
THE COMPARISON OF HYBRID WAVELET-SPECTRAL FILTER WITH OTHER WAVELET DENOISING METHODS

	Recognition rate %								
	method	Sub.	Bab.	Car	Exh.	Res.	Street	Air.	Sta.	AVE.
15dB	Avci's	96.3	95.0	95.2	95.0	95.8	95.5	94.9	95.0	95.4
	Ghanbari's	94.0	92.8	94.0	94.6	93.0	94.1	94.4	94.9	94.
	Hybrid	96.5	95.1	96.7	95.8	95.9	95.6	96.1	96.5	96.0
10dB	Avci's	88.9	83.8	82.2	88.0	86.0	84.0	83.9	84.3	85.1
	Ghanbari's	84.2	81.8	82.6	85.3	84.9	83.7	82.9	85.7	83.9
	Hybrid	88.9	84.7	84.7	88.4	88.5	85.6	85.2	87.8	86.7
5dB	Avci's	71.8	66.1	74.5	71.0	69.5	66.1	68.4	65.9	69.2
	Ghanbari's	68.5	60.3	65.8	66.1	62.5	61.1	64.5	64.1	64.1
	Hybrid	73.2	66.0	73.8	69.8	68.7	67.3	70.2	67.2	69.5
0dB	Avci's	51.3	38.6	40.4	41.6	41.8	38.9	41.0	41.8	41.9
	Ghanbari's	46.9	37.9	40.0	40.8	39.5	39.1	41.2	38.1	40.4
	Hybrid	53.5	40.2	45.8	42.8	43.7	42.8	46.1	44.0	44.9
-5dB	Avci's	29.9	28.0	21.3	17.1	17.3	18.6	20.7	19.0	21.5
	Ghanbari's	28.0	27.5	20.8	18.0	16.7	16.4	19.3	18.9	20.7
	Hybrid	30.5	27.9	22.4	17.3	18.8	18.2	21.4	19.6	22.0

REFERENCES

[1] X. Li and X. Li , "Speech emotion recognition using novel HHT-TEO based features," *Journal of Computers*, vol. 6, no. 5, pp. 989-998, 2011.

[2] J. Bai, J. Wang and X. Zhang, "A parameters optimization method of v-support vector machine and its application in speech recognition," *Journal of Computers*, vol. 8, no. 1, pp. 113–120, 2013.

[3] H. Taşmaz and E. Erçelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE–STSA estimation in various noise environments," *Digital Signal Processing*, vol. 18, no. 5, pp. 797–812, 2008.

[4] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 650–664, 2009.

[5] C. Plapous, C. Marro and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2098-2108, 2006.

[6] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," IEEE *Trans. on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2348-2359, 2007.

[7] N. Ma, M. Bouchard and R. A. Goubran, "Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 19–32, 2006.

[8] Y. Shao and C.-H. Chang, "A generalized time–frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," *IEEE Trans. on Systems, Man and Cybernetic—Part B: Cybernetics*, vol. 37, no. 4, pp. 877–889, 2007.

[9] K. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1579-1591, 2007.

[10] J. Hao, H. Attias, S. Nagarajan, T.-W. Lee and T.J. Sejnowski, "Speech enhancement, gain, and noise spectrum adaptation using approximate Bayesian estimation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 24-37, 2009.

[11] J.-H. Chang, Q.-H. Jo, D.-K. Kim and N.-S. Kim, "Global soft decision employing support vector machine for speech enhancement," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 57-60, 2007.

[12] M.K. Hasan, S. Salahuddin and M.R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Processing Letters*, vol. 11, no. 4, pp. 450-453, 2004.

[13] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE* ICASSP, vol. 4, pp. 208-211, 1979.

[14] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, 1979.

[15] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator, " *IEEE Signal Processing Letters*, vol. 8, pp. 10–12, 2001.

[16] Z. Lin, R. A. Goubran and R. M. Dansereau, "Noise estimation using speech/non-speech frame decision and subband spectral tracking, " *Speech Communication*, vol. 49, pp. 542–557, 2007.

[17] R. Tahmasbi and S. Rezaei, "A soft voice activity detection using GARCH filter and variance *gamma* distribution," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1129–1134, 2007.

[18] A. Davis, S. Nordholm and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[19] M.T. Johnson, X. Yuan and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Communication*, vol. 49, pp. 123-133, 2007.

[20] Y. Ghanbari and M.R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, pp. 927-940, 2006.

[21] V. N. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[22] S. Kamath and P.C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE ICASSP*, vol. 4, p. IV-4164, 2002.

[23] C.-C. Yao and M.-H. Tsai, "Adaptive fuzzy filter for speech enhancement," *Lecture Notes in Computer Science*, vol. 6018, pp. 511-525, 2010.

[24] C.-C. Yao and R.-W. Hung, "A hybrid microphone array filter for speech enhancement," *International Review on Computers and Software*, vol. 5, no. 6, pp. 640-651, 2011.

[25] R. M. Udrea, N. D. Vizireanu and S. Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter," *Digital Signal Processing*, vol. 18, pp. 581-587, 2008.

[26] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Network*, pp. 1942-1948, 1995.

[27] J.-H. Chang, Q-H. Jo, D.-K. Kim and N.-S. Kim, "Global Soft Decision Employing Support Vector Machine For Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 57-60, 2009.

[28] W. Done and R.L. Kirlin, "Speech zero-crossing rate compression for bandwidth compression," *IEEE Trans. on*

Acoustics, Speech and Signal Processing, vol 23, no 5, pp. 433-438, 1975.

[29] S.-H. Chen, H.-T. Wu, Y. Chang, and T.K. Truong, "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator," *Pattern Recognition Letters*, vol. 28, pp. 1327-1332, 2007.

[30] H.-K. Jeff Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 873-881, 2006.

[31] I. M. Johnson and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist. Soc.*, Ser. B 59, pp. 319–351, 1997.

[32] Y.-J. Lee, and O. L. Mangasarian, "SSVM: a smooth support vector machine for classification, " *Computational Optimization and Application*, vol. **20**, pp. 5-22, 2001.

[33] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR'2000: Challenges for the New Millenium*, Paris, France, September.

[34] I. Daubechies, *Ten Lectures on Wavelets*, CBMS, SIAM publish, 1992.

[35] ETSI EN 301 708 V7.1.1, "Voice Activity Detector (VAD) for Adaptive Multi-Rate," 1999.

[36] Speech Vision and Robotics Group, http://htk.eng.cam.ac.uk.

[37] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, vol. 36, pp. 111-147, 1974.

[38] A.W.C. Tan, M.V.C. Rao and B.S.D. Sagar, "A signal subspace approach for speech modelling and classification," *Signal Processing*, vol. 87, no.3, pp. 500-508, 2007.

[39] Y. Lu and P.C. Loizou, "A geometric approach to spectral subtraction, " *Speech Communication*, vol. 50, no. 6, pp. 453-466, 2008.

[40] E. Avci and Z.H. Akpolat, "Speech recognition using a wavelet packet adaptive network based fuzzy inference system", *Expert Systems with Applications*, vol. 31, pp. 495-503, 2006.