Progress In Electromagnetics Research, Vol. 130, 17–40, 2012

A NEW KIND OF NON-ACOUSTIC SPEECH ACQUISITION METHOD BASED ON MILLIMETER WAVE RADAR

S. Li, Y. Tian, G. Lu, Y. Zhang, H. Xue, J. Wang*, and X. Jing †

College of Biomedical Engineering, The Fourth Military Medical University, Xi'an 710032, China

Abstract—Air is not the only medium that can spread and can be used to detect speech. In our previous paper, another valuable medium — millimeter wave (MMW) was introduced to develop a new kind of speech acquisition technique [6]. Because of the special features of the MMW radar, this speech acquisition method may provide some exciting possibilities for a wide range of applications. In the proposed study, we have designed a new kind of speech acquisition radar system. The super-heterodyne receiver was used in the new system so that to mitigate the severe DC offset problem and the associated $1/f$ noise at baseband. Furthermore, in order to decrease the harmonic noise, electro-circuit noise, and ambient noise which were combined in the MMW detected speech, an adaptive wavelet packet entropy algorithm is also proposed in this study, which incorporates the wavelet packet entropy based voice/unvoiced radar speech adaptive detection method and the human ear perception properties in a wavelet packet time-scale adaptation speech enhancement process. The performance of the proposed method is evaluated objectively by signal-to-noise ratio and subjectively by mean-opinion-score. The results confirm that the proposed method offers improved effects over other traditional speech enhancement methods for MMW radar speech.

1. INTRODUCTION

It is well known that speech, which is produced by the speech organ of human beings [1–3], has significant effects on the communication

Received 22 May 2012, Accepted 6 July 2012, Scheduled 13 July 2012

* Corresponding author: Jian-Qi Wang (Sheng@mail.xjtu.edu.cn).

† The first three authors contributed equally to this work and should be regarded as co-first authors.

and the information exchange among human beings. Acoustic speech signals have many other applications such as those involving conversion to text or coding for transmission. However, thus far, the popular method for speech signal acquisition is almost limited to the air-conducted speech; this method is based on the theory that speech can be conducted by *air* in free space and can easily be heard and recorded when conducted by air. However, this method has some serious shortcomings: (1) the acquisition distance of a traditional microphone or acoustic sensor is quite limited; therefore, people have to carry the microphone during their lectures, news reports, telephone calls, or theatrical performances. (2) The directional sensitivity of traditional speech/acoustic transducers (including microphones) is quite weak; as a result, the ability of traditional transducers to set off other acoustic disturbance may be poor. Therefore, it is not possible to acquire speech (or hear a particular sound) in a background with considerable noise, such as in the cockpit of a tank or a plane, or in any other rumbustious environment. (3) The frequency bandwidth of a traditional acoustic transducer is narrow; hence, the traditional acoustic transducer cannot be used for wide-spectrum acoustic signal acquisition. (4) The sensitivity of a traditional acoustic transducer is poor; therefore, it is not possible for the traditional acoustic transducer to detect a tiny acoustic or vibratory signal.

Another speech acquisition method, which does not depend on conduction by air or can overcome the shortcomings of the traditional speech acquisition method, is required. Previous studies have proposed some methods. For example, voice content can be transmitted by way of bone vibrations. These vibrations can be picked up at the top of the skull using bone-conduction sensors. Strong voicing can be facilitated using this method [4]. Other media such as infrared rays, light waves, and lasers can also be used to acquire the non-air-spread speech or acoustical vibrations. However, their application is limited since the constraint of their application conditions or their materials in detail are usually difficult to obtain [5].

A novel non-air conducted speech acquisition method has been developed in our laboratory [6]. This method uses a different medium — the millimeter wave — to detect and exactly identify the speech (or acoustic) signals generated by a person that exist in free space. Radar has some special features, such as low-range attenuation, good sense of direction, wide frequency bandwidth, and high sensitivity [7–13], which the traditional speech acquisition method does not have. Therefore, this special microphone, which we call "*radar-microphone*," may extend the speech and acoustic signal acquisition method to a considerable extent. Moreover, the method that involves the use of the MMW radar has the same attributes as the traditional speech acquisition method, such as noninvasiveness, safety, speed, portability, and cheapness [14]. Therefore, this new speech acquisition method may offer exciting opportunities for the following novel applications: (1) a hands-free, long-distance ($> 10\,\mathrm{m}$), directional speech/acoustic signal acquisition system, which can be used in both common and complex/rumbustious acoustic environments (e.g., a sharp whistle blows at the left-hand side of the radar when we are conducting the experiment in the playground; however, there is no whistle sound in the recorded speech); (2) the tiny, wide frequency bandwidth acoustic or vibratory signal acquisition,which cannot be detected by a traditional microphone; (3) MMW radar that can also be used for assisting clinical diagnosis or for measuring speech articulator motions [15].

However, there have been only a few reports on the MMW non-air conducted speech. A similar experiment had been carried out more than ten years ago [5], and a further research report has not been found. Other researches with respect to the radar speech concentrated on the non-acoustic sensors [14, 16, 17] and the measurement of speech articulator motions, such as vocal tract measurements and glottal excitation [15], but not on the MMW radar speech itself. Therefore, there is a need to explore this new speech acquisition method (as well as the corresponding speech enhancement algorithm) to extend the existing speech acquisition method.

Although the MMW radar offers exciting possibilities in the field of speech (or other acoustic signal) acquisition, the MMW radar speech itself has several serious shortcomings, including artificial quality, reduced intelligibility, and poor audibility. This is because that the theories governing the acquisition of MMW radar speech and traditional air-conducted speech are different. Therefore, some combined harmonics of the MMW and electro-circuit noise are present in the detected speech. Furthermore, channel noise and some ambient noise also exist in the MMW radar speech [18–20]. Among these noises, the harmonic noise and electro-circuit noise is quite larger and more complex than traditional air-conducted speech, and they degrade the MMW radar speech, this is especially true for the low-frequency components (see Figure 4). This is the biggest problem that must be resolved for the application of the MMW radar speech. Therefore, speech enhancement is a challenging topic for MMW radar speech research.

The special characters of the radar speech noise suggest that a special speech enhancement method should be developed and applied to the MMW radar speech. However, very little research has been carried out on the MMW radar speech enhancement.

Li et al. [6] proposed a multi-band spectral subtraction approach that takes into account the fact that colored noise affects the speech spectrum differently at various frequencies. Although the speech quality was improved by this algorithm, it suffered from an annoying artifact called "musical noise," [21, 22] which is caused by narrow band tonal components appearing somewhat periodically in each frame and occurring at random frequencies in voice or silence regions. They also explored other methods focused on masking the musical noise using psychoacoustic models [6]; results obtained by using these algorithms show that there is a need for further improvement in the radar speech enhancement algorithm, especially at a very low SNR condition (SNR $< 10\,\mathrm{dB}$). Furthermore, these algorithms are based on the spectral subtraction method, which is in general effective in reducing the noise but not in improving intelligibility. Therefore, it is necessary to find a new way to improve intelligibility and reduce speech distortion when reducing noise.

The wavelet transforms (WT), which can be easily obtained by filtering a signal with multi-resolution filter banks [23, 24], has been applied to various research areas, including signal and image denoising, compression, detection, and pattern recognition [25–30]. Recently, WT have been applied in denoising signals on the basis of the threshold of the wavelet coefficients, where the wavelet threshold (shrinking) introduced by D. L. Donoho et al. [31] is a simple but powerful denoising technique based on the threshold of the wavelet coefficients. Previous studies have also reported the application of wavelet shrinking for speech enhancement [32, 33]; however, it is not possible to separate the signal from noises by a simple threshold because applying a uniform threshold to all wavelet coefficients would remove some speech components while suppressing additional noise, especially for the colored noise corrupted signal and some deteriorated speech conditions [34].

In order to overcome the limit of uniform threshold, many previous researches combined the wavelet transforms successfully with other denoising algorithms, such as Wiener filtering in the wavelet domain [35], wavelet filter bank for spectral subtraction [36], or coherence function [37]. The results of these methods suggest that they can improve the performance of speech enhancement methods; however, these wavelet-based methods generally need an estimation of noise.

Therefore, an algorithm that is based on an adaptive time-scale threshold of wavelet packet coefficients [38] without the requirement of any knowledge of the noise level is used in this study. To improve the performance of this algorithm for MMW radar speech, this study extends their wavelet filter-banks to nonlinear Bark-scaled frequency spacing because the human ear sensibility is a nonlinear function of frequency. The proposed method attempts to find the best tradeoff between speech distortion and noise reduction that is based on properties closely related to human perception.

Another issue to be resolved in most of the speech enhancement algorithms is the decision regarding the sectioning of voice/unvoiced speech. Bahoura's algorithm [38] discriminated speech from silence by experimentally determining a discriminatory value of 0.35. However, this value is fixed and cannot be changed from frame to frame. This limitation is worse for the enhancement of speech, particularly important for MMW radar speech, where the combined noise decreases the SNR, thereby making it quite difficult to detect voiced/unvoiced speech sections. In order to effectively resolve this issue, the present study presents a novel approach to the segmentation of voice/unvoiced speech sections that is based on wavelet packet analysis and entropy.

Entropy is defined as a measure of uncertainty of information in a statistical description of a system [39], and the spectral entropy is a measure of how concentrated or widespread the Fourier power spectrum of a signal is. In this study, a time-frequency description of MMW radar speech, as described by the wavelet packet coefficient, is used to calculate the entropy, which forms the wavelet packet spectral entropy. By its very definition, wavelet packet entropy is considerably sensitive both to the time-frequency distribution and to the uncertainty of information; therefore, this novel tool may have very useful characteristics with regard to speech section detection.

Therefore, compare to the first generation of the radar speech acquisition method [6], this research develops a new kind of Doppler radar system by using a super-heterodyne receiver, in order to (1) increase the system stability; (2) decrease the hardware system noise by reducing the DC offset and $1/f$ noise which has degradation effects on system signal-to-noise ratio and detection accuracy. Also, in order to enhance the detected radar speech, a speech enhancement algorithm is also proposed, which is on the basis of time-scale adaptation of wavelet packet coefficient thresholds by incorporating the human ear perception and wavelet packet entropy. The steps for radar speech enhancement and its effectiveness evaluation are as follows: (1) to adopt the wavelet packet analysis to decompose speech into nonlinear critical sub-bands and to compute the wavelet packet entropy using these wavelet packet coefficients so as to detect the voiced/unvoiced speech segment; (2) to apply the time-scale adaptation of wavelet packet thresholds to the speech enhancement algorithm, incorporating the human ear perception and wavelet packet entropy approach for improving MMW radar speech; and (3) to evaluate the quality of the enhanced MMW speech in comparison to speech enhanced by some other representative algorithm.

2. METHOD

2.1. Doppler Radar Detection Principle

For Doppler radar detection of throat vibration, the un-modulated signal sent from transmitting antenna is:

$$T_r\,(t) = A\cos(2\pi ft + \varphi_0) \tag{1}$$

where f is carrier frequency (or transmitting frequency), t is the elapsed time, and φ is the residual phase. If this signal is reflected (by a vital target) with a phase shift $\varphi(t)$ produced by the electromagnetic propagation between the transmitting antenna and the vital target, the received signal can be approximated as:

$$R_e\,(t) = k_1 T_r\,(t-\tau) = k_1 A\cos\left[2\pi f(t-\tau) + \varphi_0 + \varphi(t)\right] \tag{2}$$

where k_1 is attenuation coefficient, $\tau = 2R/c$, here R is the target distance, and c is the velocity of light. The radar receiver down-converts the received signal $R_e\,(t)$ into baseband signal in the mixer:

$$B\,(t) = A\cos(2\pi ft + \varphi_0)\cdot k_1 A\cos\left[2\pi f(t-\tau) + \varphi_0 + \varphi(t)\right] \tag{3}$$

The high frequency and DC components can be filtered by proper digital signal processing (DSP), then:

$$B\,(t) \approx K\sin\varphi(t) \tag{4}$$

where K is the gain of the filter and mixer, if the displacement of throat vibration is small compare to the wavelength of the transmitting signal, the phase shift $\varphi(t)$ is approximate to $\sin\varphi(t)$, then Eq. (4) can be rewritten as [40]:

$$B\,(t) \approx K\varphi(t) \tag{5}$$

Equation (5) suggest that the phase shift of the radar received signal almost has a linear relation to the baseband signal, suggest that the displacements (or vibration) of the human throat or chest can be detected by using Doppler radar.

2.2. Description of the MMW Speech Acquisition System

The schematic diagram of this nonacoustic speech acquisition system is shown in Figure 1. A phase-locked oscillator generates a very stable MMW at 34.5 GHz with an output power of 100 mW. This output is fed into both the transmitting circuit and the receiving circuit. In

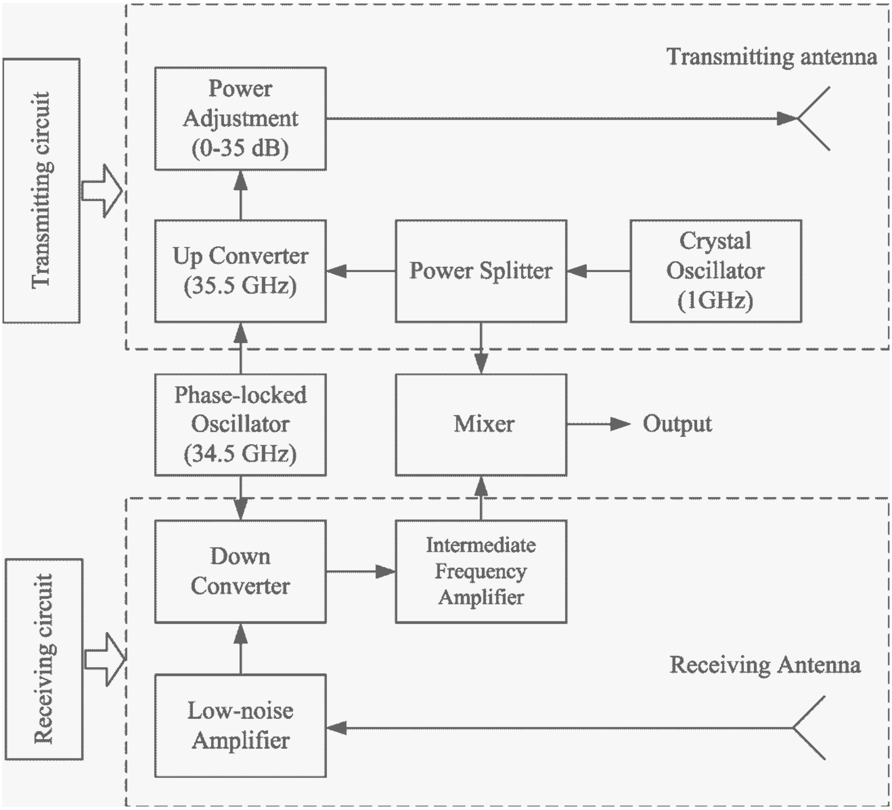


Figure 1. Schematic diagram of the speech-acquisition system.

the transmitting circuit, the MMW is up-converted to 35.5 GHz by mixed with a 1 GHz crystal oscillator, this wave is fed through a power attenuator before reaching the transmitting antenna. By using the variable power attenuator, the power level of the microwave signal to be radiated by the antenna is controlled, and the adjusting range is $0 \sim 35$ dB.

For the receiving circuit, the reflected wave is amplified by a low-noise amplifier (Noise figure is 4 dB, the Gain is 18 dB) after received by the receiving antenna. The transmitting and receiving antennas are both parabolic antennas with a diameter of 300 mm, the estimated beam width is $9^\circ \times 9^\circ$, and the maximum antenna gain is 38.5 dB at 35.5 GHz. The amplified wave is down converted with the 34.5 GHz phase-locked oscillator frequency, and then mixed with

1 GHz crystal oscillator frequency after amplified by an intermediate frequency amplifier. A power splitter is used to divide the power of the crystal oscillator, with half of the power fed to the up-converter (transmitting circuit) and the other half to the mixer. The mixer output provides the speech signal from the body, which is amplified by a signal processor and is then passed through an A/D converter before reaching a computer for further processing. All the signals were sampled at a frequency of 1000 Hz.

As shown in Figure 1, two dashed boxes form the transmitting circuit and receiving circuit separately, the advantage of this kind of radar component layout is that it employs two-step indirect-conversion transceiver, so that to mitigate the severe DC offset problem and the associated $1/f$ noise at baseband, that occurs normally in the direct-conversion receivers. Compared to single antenna which we used before [6], the antenna array has higher directive gain, which can both increase the detection distance and reduce interference from other directions. When the detection was performed, a 16-channel Power Lab data acquisition system (AD Instruments) displayed and recorded the radar baseband signal, this signal was further processes by using a MATLAB program (R2007b).

2.3. Wavelet Packet Noise Reduction Algorithm

2.3.1. Bark (Critical) Band

It is well known that the sensibility of the human ear varies nonlinearly in the frequency spectrum, which denotes the fact that the perception of the auditory system of a signal at a particular frequency is influenced by the energy of a perturbation signal in a critical band around this frequency. The bandwidth of this critical band, furthermore, varies with frequency. A commonly used scale for signifying the critical bands is the Bark-band, which divides the audible frequency range of $0 \sim 16$ KHz into 24 abutting bands. An approximate analytical expression to describe the relationship between linear frequency and critical band number B (in Bark) is [41]:

$$B(f) = 13 \arctan(0.7f) + 3.5 \left[\left(\frac{f}{7.5} \right)^2 \right] \tag{6}$$

In this paper, the linear frequency of the radar speech is $0 \sim 5000$, therefore the Bark-band number is set to 6. Figure 2 illustrates the relationship between the frequency in hertz and the critical-band rate in Bark.

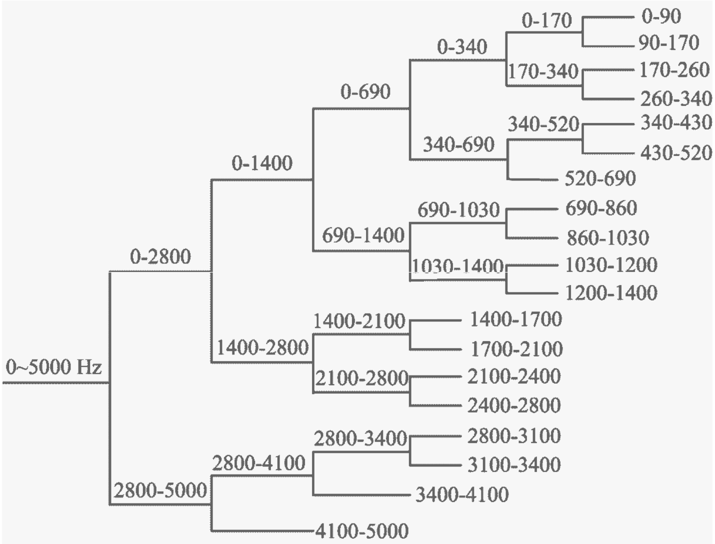


Figure 2. Nineteen bands of a wavelet packet tree that closely mimic the critical bands.

2.3.2. Wavelet Packet Analysis

The wavelet packet analysis (WPA), which is based on the wavelet transform, can offer a large range of possibilities for signal analysis [42]. If $y(n)$, the noisy speech, consists of the clean speech signal $s(n)$ and the uncorrelated additive noise signal $d(n)$, then:

$$y(n) = s(n) + d(n) \tag{7}$$

For a given level, the wavelet packets transform (WPT) decomposes the noisy signal $y(n)$ into 2^i subbands, with the corresponding wavelet coefficient sets as $w^i_{j,m}$:

$$w^i_{j,m} = WP\left\{s(n), i\right\} \quad n = 1, \dots N. \tag{8}$$

where $w^i_{j,m}$ denotes the mth coefficient of the jth subband for the ith level, and $m = 1, \dots, N/2^i$, $k = 1, \dots, 2^i$. In this study, i is set to 6 in the Bark-band.

The enhanced speech is synthesized with the inverse transformation of the processed wavelet packet coefficients:

$$\hat{s}(n) = WP^{-1}\left\{\hat{w}^i_{j,m}, i\right\} \tag{9}$$

where $\hat{s}(n)$ is the enhanced radar speech, and $\hat{w}^i_{j,m}$ is the updated wavelet packet coefficient which is calculated by the algorithm stated below.

2.3.3. Wavelet Packet Entropy

The subband wavelet packet entropy is defined in terms of the relative wavelet energy of the wavelet coefficients [43]. The energy for each subband j and level i can be calculated as:

$$E_j^i = \sum_m \left|w_{j,m}^i\right|^2 \tag{10}$$

The total energy of the wavelet packet coefficients will then be given by:

$$E_{\text{total}}^i = \sum_j \left|w_j^i\right|^2 \tag{11}$$

and the probability distribution for each level can be defined as:

$$p_j^i = \frac{E_j^i}{E_{\text{total}}^i} \tag{12}$$

Since, following the definition of entropy given by Shannon (1948) [44], the subband wavelet packet entropy is defined by using the probability distribution associated with scale level i (for further details see [43] and [45]), we have:

$$H(i) = -\sum_j p_j^i \log p_j^i \tag{13}$$

Two adaptive wavelet packet entropy thresholds are selected to detect the onset and offset of MMW radar speech. The speech onset threshold is T_s and the offset threshold is T_n. T_s is defined by adding a fixed value E_s to a past mean wavelet packet entropy value T_m. T_m is calculated over the previous t ms (five frames). The speech offset (noise) threshold T_n is calculated by adding another fixed value E_n to T_m. When $H(i)$ (in Eq. (13)) exceeds T_s, speech onset is detected and speech offset is detected when $H(i)$ drops below T_n. Therefore, the wavelet packet entropy thresholds can be dynamically adjusted. In this study for MMW radar speech, E_s and E_n are set at the constant values of 1.7 and 1.3, respectively.

2.3.4. Speech Enhancement Based on Time-scale Adaptation

The proposed enhancement scheme is presented in Figure 3. First, we performed wavelet packet decomposition by using a nonlinear Bark-band, and then, we carried out voiced/unvoiced speech section detection by using adaptive wavelet packet entropy thresholds. Then, we calculated the time-scale adapted threshold on the basis of the Teager energy operator, masks construction, and thresholding process.

Finally, we synthesized the enhanced speech with a wavelet packet inverse transform of the processed wavelet coefficients.

a. **Teager energy operator**: In order to carry out the time-adapting approach, the Teagerenergy operator (TEO) [46] was used to create a mask [38], which can be calculated by the resulting wavelet coefficients $w^i_{j,m}$ of each subband j:

$$t^i_{j,m} = \left[w^i_{j,m}\right]^2 - w^i_{j,m-1} w^i_{j,m+1} \tag{14}$$

b. **Mask processing for the time-adapting threshold**: An initial mask for each subband j is constructed by smoothing the corresponding TEO coefficients and normalizing, which is determined as following [46]:

$$M^i_{j,m} = \frac{t^i_{j,m} * h_j(m)}{\max\left(\left|t^i_{j,m} * h_j(m)\right|\right)} \tag{15}$$

where $h_j(m)$ is an IIR low-pass filter (second order) and the max is the maximum of the smoothed TEO coefficients in the considered sub-band. For an unvoiced speech section, the mask is directly set to 0. For a voiced speech section, the mask is normalized before applying a root power function of 1/8, in order to implement a compromise between noise removal and speech distortion [38]:

$$M'^i_{j,\,m} = \left[\frac{\left|M^i_{j,m}\right| - S^i_j}{\max\left(\left|M^i_{j,m}\right| - S^i_j\right)}\right]^{\frac{1}{8}} \tag{16}$$

where S^i_j is given by the abscissa of the maximum of the amplitude distribution of the corresponding mask $M^i_{j,m}$.

c. **Thresholding process**: A scale-adapted wavelet threshold, which is derived from the level-dependent threshold [38, 47] is used in this study. For a given subband j, the corresponding threshold λ_j can be defined as:

$$\begin{cases} \lambda_j = \sigma_j \sqrt{2\log(N)} \\ \sigma_j = MAD_j/0.6745 \end{cases} \tag{17}$$

where N is the length of the noisy speech for each subband, MAD_j is the median of the absolute value estimated on the subband j. Therefore, the time-scale adapted threshold is obtained by adapting the corresponding threshold in the time domain:

$$\lambda_{j,m} = \lambda_j \left(1 - \alpha M'^i_{j,m}\right) \tag{18}$$

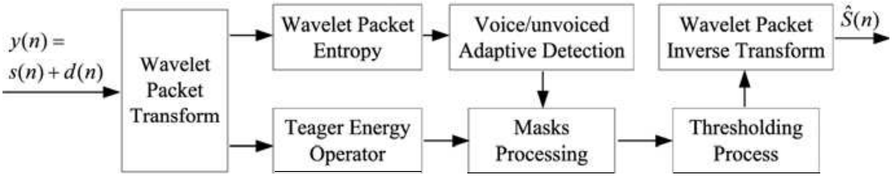


Figure 3. The proposed adaptive wavelet packet entropy speech enhancement scheme.

where α is an adjustment parameter ($\alpha = 1$).

The soft thresholding, which is defined by Donoho and Johnstone [48, 49], is then applied to the wavelet packet coefficients:

$$\begin{aligned}\hat{w}^{i}_{j,m} &= Ts\left(\lambda_{j,m}, w_j\right)\\ &= \begin{cases} \operatorname{sgn}\left(w_j\right)\left(\left|w_j\right| - \lambda_{j,m}\right) & \text{if} \quad \left|w_j\right| > \lambda_{j,m} \\ 0 & \text{if} \quad \left|w_j\right| \le \lambda_{j,m} \end{cases}\end{aligned} \tag{19}$$

The enhanced signal, therefore, can be synthesized with the inverse transformation WP^{-1} of the processed wavelet coefficients (Eq. (9)).

3. EXPERIMENTS

3.1. Subjects

Ten healthy volunteer speakers, 6 males and 4 females, participated in the radar speech experiment. All the subjects were native speakers of Mandarin Chinese. Their ages varied from 20 to 35, with a mean age of 28.1 (SD = 12.05). All the experiments were conducted in accordance with the terms of the Declaration of Helsinki (BMJ 1991; 302: 1194), and appropriate consent forms were signed by the volunteers.

The distance between the radar antenna and the human subjects ranged from 2 m to 30 m. Ten sentences of Mandarin Chinese were used as the speech material for acoustic analysis and acceptability evaluation. The lengths of the sentences varied from 6 words (5.6 s) to 30 words (15 s). The sentences were spoken by each participant in a quiet experimental environment. The speakers were instructed to read the speech material at normal loudness and speaking rates.

3.2. Additive Noise

In order to test the effectiveness of the proposed method, two different types of background noise, namely, white Gaussian noise and speech babble noise, were added to the enhanced MMW radar speech; both noises were taken from the Noise X-92 database. These two representative noises have a greater similarity to actual talking conditions than the other noises. Noises with varying SNRs of −10, −5, 0, +5, and +10 dB were added to the original MMW radar speech signal. SNR is defined as:

$$\mathrm{SNR} = 10 \times \log_{10}\left(\frac{\sum_{n=1}^{N} y^2(n)}{\sum_{n=1}^{N}\left|y(n) - \hat{s}(n)\right|^2}\right) \tag{20}$$

where $\hat{s}(n)$ is the enhanced speech, and N is the number of samples in the clean and enhanced speeches.

3.3. Perceptual Evaluation

For the perceptual experiment, eight listeners were selected to evaluate the acceptability of each sentence based on the criteria of the mean opinion score (MOS), which is a five-point scale (1: bad; 2: poor; 3: common; 4: good; 5: excellent). All the listeners were native speakers of Mandarin Chinese, had no reported history of hearing problems, and were unfamiliar with MMW radar speech. Their ages varied from 22 to 36, with a mean age of 26.37 (SD = 4.63). The listening tasks took place in a soundproof room, and the speech samples were presented to the listeners at a comfortable loudness level (60 dB sound pressure level (SPL)) via a high quality headphone. A 4-s pause was inserted before each citation word, and the order in which the speech samples were presented was randomized, to allow the listeners to respond and to avoid rehearsal effects.

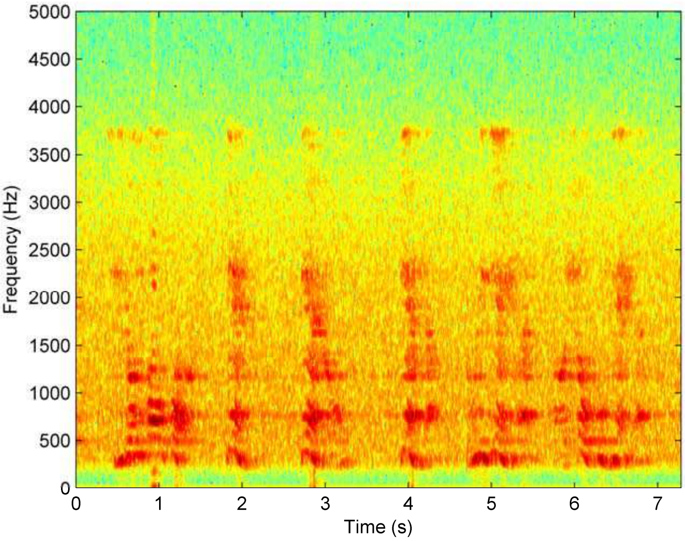
4. RESULTS AND DISCUSSIONS

The performance of the proposed algorithm is evaluated and compared to that of other algorithms. The other algorithms include the noise estimation algorithm [50], traditional wavelet transform denoising methods [49], and the time-scale adaptation algorithm [38]. For evaluation purposes, 100 sentences, which are spoken by 6 male and 4 female volunteer speakers, are used.

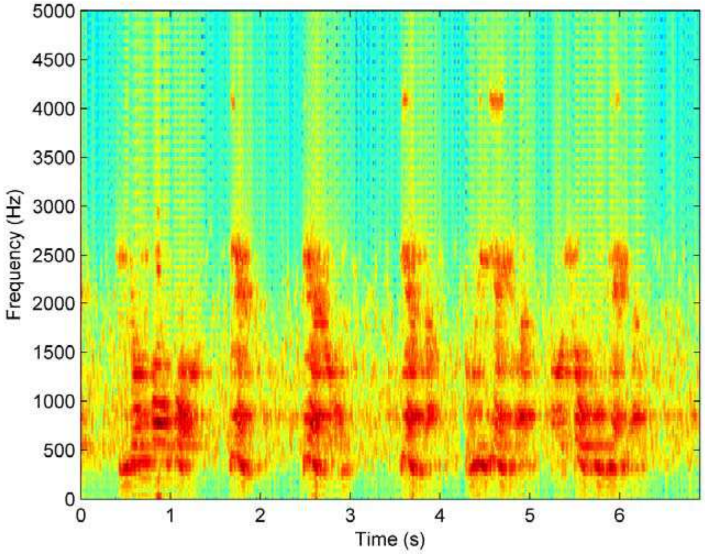
Generally, a speech enhancement system produces two main undesirable effects: residual noise and speech distortion. However, these effects are difficult to quantify with the help of traditional objective measures. Therefore, speech spectrograms were used in this study since they have been identified as a well-suited tool for observing both the residual noise and the speech distortion. In addition, the results were evaluated objectively by signal-to-noise ratio (SNR) and subjectively by mean opinion score (MOS) under conditions of different additive white Gaussian noise as well as bobble noise (for MOS) for the algorithm evaluation.

Figure 4 shows the spectrograms of the original MMW radar speech (a), the enhanced speech using the noise estimation algorithm (b), the enhanced speech using the traditional wavelet transform denoising algorithm (c), the enhanced speech using the time-scale adaptation algorithm (d), and the proposed adaptive wavelet packet entropy algorithm (e).

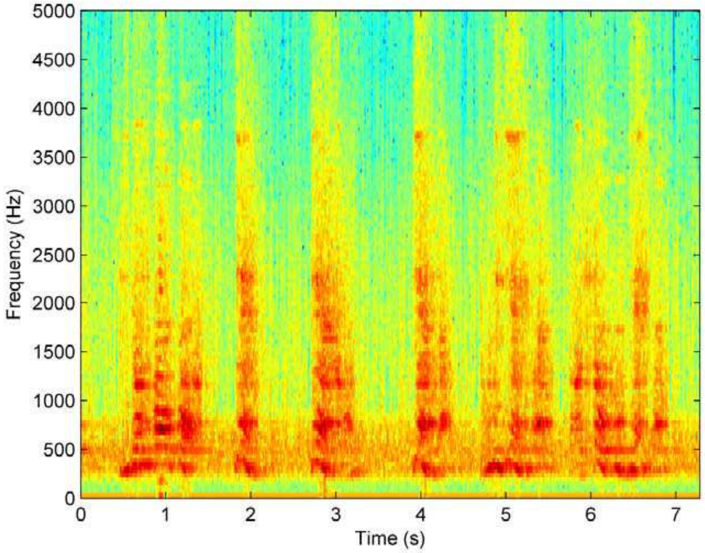
As stated earlier, the combined noises are introduced into the original MMW radar speech. These noises can be clearly seen in Figure 4(a), especially in the speech-pause region. It can also be seen from the figure that the noises are mainly concentrated in the low-frequency components, roughly below 3 kHz. Figures 4(b) and (c) show that the noise estimation algorithm and the traditional wavelet transform denoising methods are effective in reducing the combined radar noises, both in the speech and the non-speech sections. However, there is still too much remnant noise in the enhanced speech, especially in the frequency section in which the noise is concentrated, suggesting that the noise reduction is not satisfactory. It can be seen from Figure 4(d) that the time-scale adaptation algorithm has great effects



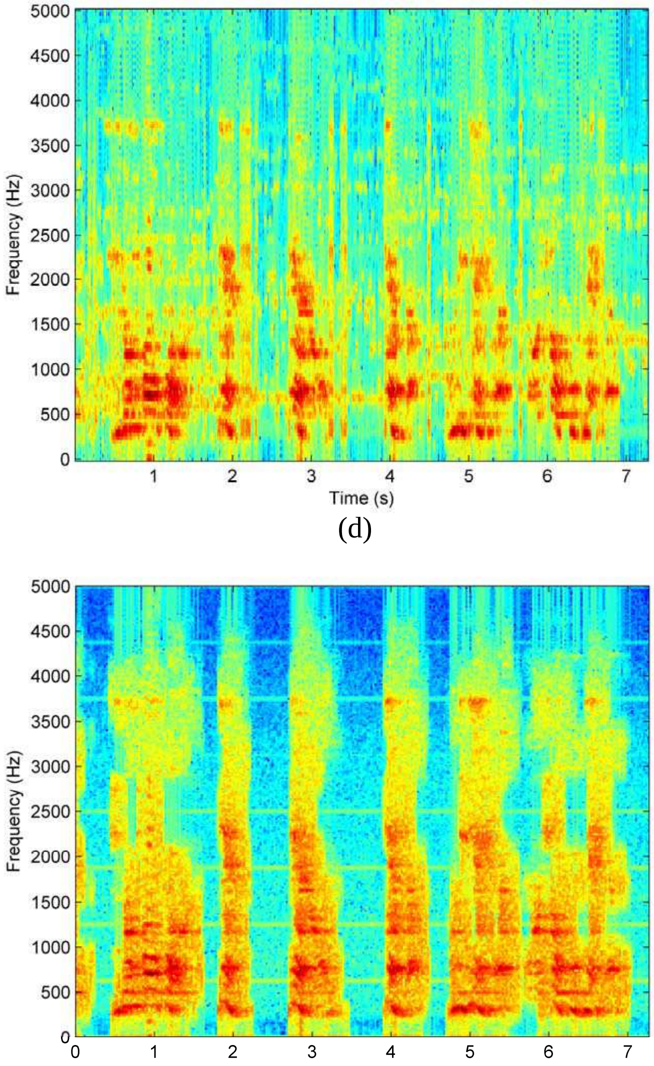
(a)


Frequency (Hz)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Time (s)
0
1
2
3
4
5
6

(b)


Frequency (Hz)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Time (s)
0
1
2
3
4
5
6
7

(c)



(d)

(e)

Figure 4. The Spectrogram of the radar speech: (a) The original MMW radar speech; (b) enhanced speech by the noise estimation algorithm; (c) by the traditional wavelet transform denoising methods; (d) by the time-scale adaptation algorithm; (e) by the proposed adaptive wavelet packets entropy algorithm.

on enhancing MMW radar speech, but does not remove entirely the noise. Figure 4(e) shows that the proposed adaptive wavelet packet entropy threshold algorithm can not only greatly reduce the low-frequency noise, in which the combined radar noise is concentrated, but it also completely eliminates the high-frequency noise. It can be seen from the figure that in the speech-pause regions the residual noise is almost eliminated. Moreover, it is clear that the residual noise is greatly reduced and has lost its structure. These results suggest that the proposed algorithm achieves a better reduction of the whole-frequency noise than other methods.

The mean results of the SNR measurements in terms of an objective measure for 100 MMW radar sentences are shown in Figure 5; the values for each sentence were corrupted by white noise at -10, -5, 0, $+5$, and $+10$ dB SNR levels. Methods compared included the noise estimation algorithm (noise estimation), traditional wavelet transform denoising methods (wavelet transform), time-scale adaptation algorithm (time-scale adaptation), and the proposed adaptive wavelet packet entropy algorithm (wavelet packets entropy). As shown in Figure 5, the proposed method has the best performance, followed by the time-scale adaptation algorithm and the noise estimation method. It also can be seen from the figure that the proposed method has a nearly 2 dB better performance than any of the other above mentioned methods in the -10 dB noise case; further, this difference decreases with an increase in the SNR values, suggesting

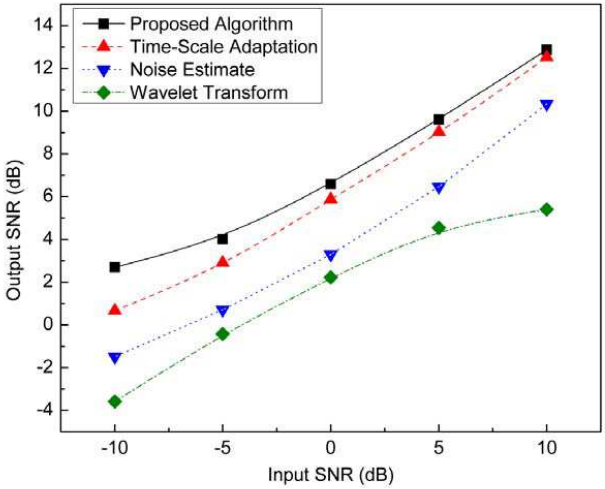


Figure 5. SNR results for white noise case at -10, -5, 0, $+5$, and $+10$ dB SNR levels.

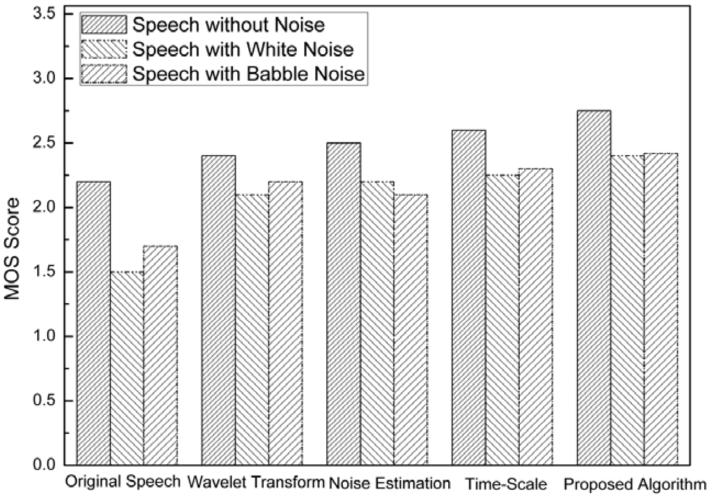


Figure 6. Acceptability scores of the original and enhanced MMW radar speech. The noisy speech in the case of additive noise has an input SNR of 0 dB.

that the proposed method has considerably better performance than any of the other above mentioned methods, especially in the low SNR noise cases.

The perceptual analyses score (subjective results) obtained using MOS for these same conditions are shown in Figure 6. Eight listeners were asked to rate the sentences for quality as stated before. MOSs were used for 100 original MMW radar sentences produced by ten volunteer speakers and for the noisy sentences for white and bubble noise at 0 dB SNR levels. The score of the enhanced speech obtained by using the proposed adaptive wavelet packet entropy algorithm is the highest, followed by that from the time-scale adaptation algorithm and the noise estimation algorithm. This is true for both the original speech and the noisy speech. Informal listening tests also indicated that the speech enhanced with the proposed algorithm is more pleasant, the residual noise is much reduced, and has minimal, if any, speech distortion. This is because the time and scale wavelet packet thresholds can be adaptively adjusted in each Bark-band, the Bark-band also takes into account the frequency-domain masking properties of the human auditory system, thus prevents quality deterioration in the speech during the threshold process.

These results indicate that the proposed adaptive wavelet packet entropy algorithm is better suited for MMW radar speech enhancement than the other above mentioned methods, especially in the case of additive noise. Because the thresholds, which are used to determine the voiced/unvoiced speech section, are fixed and cannot be changed from frame to frame, the time-scale adaptation method cannot reduce the noise effectively. This limitation will be worse for the enhancement of MMW radar speech, especially for low SNRs (see Figure 5). With regard to the wavelet packet entropy thresholds in the proposed algorithm, voiced/unvoiced speech sections can be determined adaptively. Furthermore, based on the frame-by-frame adaptations of time-scale wavelet packet thresholds in each Bark-band, the algorithm can realize a good tradeoff between reducing noise, increasing intelligibility, and keeping the distortion acceptable to a human listener. Moreover, the time-scale threshold of wavelet packet coefficients can be adequately and adaptively adjusted; this makes it possible to get better speech quality via speech enhancement in some rigorous speech environments.

The adaptive wavelet packet entropy algorithm is also effective. Although wavelet packet entropy analysis increases the computational load, a great benefit of the proposed algorithm is that the explicit estimation of the noise level or of the a priori knowledge of the SNR is not necessary, which can avoid a great computational load. Considering its better effects on speech enhancement, the proposed algorithm is quite efficient.

As a single channel wavelet-type speech enhancement method, the adaptive wavelet packet entropy algorithm proposed in this paper can be applied for the enhancement of MMW radar speech in a practical situation. For example, a MMW speech enhancing system, into which this algorithm is embedded, can be developed. With the help of digital signal processing (DSP) technology, we can realize the speech enhancement function with a microprocessor and implanted into a *radar-telephone*, *radar-microphone*, or other electronic equipment. Different enhancement algorithms, suitable for different noise conditions, can be selected by a switch. With the development of efficient enhancement methods, the quality of MMW speech will be vastly improved and will provide better perception.

As a novel speech acquisition method, i.e., MMW radar speech acquisition method can not only be used as a substitute for the existing speech acquisition method but also compensate for several serious shortcomings of the traditional microphone speech, such as acquisition distance and directional sensitivity. Therefore, the MMW radar speech acquisition method can be combined with traditional speech acquisition equipment in order to improve the performance of the speech acquisition method and to extend the application fields of the speech acquisition.

5. CONCLUSION

By means of super-heterodyne millimeter wave radar, a new kind of non-air conducted speech acquisition method (radar system) is introduced in this study. Because of the special features of the millimeter wave radar, this method can provide some exciting possibilities for a wide range of applications. However, radar speech is substantially degraded by additive combined noises that include radar harmonic noise, electro-circuit noise, and ambient noise. This study proposes an adaptive wavelet packet entropy algorithm that incorporates the human ear perception and the time-scale adaptation. Results from both the objective and the subjective measures/evaluations suggest that this method can not only greatly reduce the whole-frequency noise but also prevent speech from quality deterioration, especially in the low SNR noise cases. Furthermore, the proposed algorithm is effective because an explicit estimation of the noise level is not required.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (NSFC, No. 60571046), and the National postdoctoral Science Foundation of China (No. 20070411131). We also want to thank the participants from the E.N.T. Department, the Xi Jing Hospital, the Fourth Military Medical University, for helping with data acquisition and analysis.

REFERENCES

1. Titze, I. R., *Principles of Voice Production*, Prentice Hall, 1994.
2. Li, S., R. C. Scherer, M. Wan, S. Wang, and H. Wu, “The effect of glottal angle on intraglottal pressure,” *J. Acoust. Soc. Am*, Vol. 119, No. 1, 539–548, 2006.
3. Li, S., R. C. Scherer, W. Minxi, S. Wang, and H. Wu, “Numerical study of the effects of inferior and superior vocal fold surface angles on vocal fold pressure distributions,” *J. Acoust. Soc. Am*, Vol. 119, No. 5, 3003–3010, 2006.
4. Yanagisawa, T. and K. Furihata, “Pickup of speech signal utilization of vibration transducer under high ambient noise,” *J. Acoust. Soc. Jpn.*, Vol. 31, No. 3, 213–220, 1975.
5. Li, Z.-W., “Millimeter wave radar for detecting the speech signal

applications," *International Journal of Infrared and Millimeter Waves*, Vol. 17, No. 12, 2175–2183, 1996.

6. Li, S., J. Wang, M. Niu, and X. Jing, "The enhancement of millimeter wave conduct speech based on perceptual weighting," *Progress In Electromagnetics Research B*, Vol. 9, 199–214, 2008.
7. Park, J.-I. and K.-T. Kim, "A comparative study on isar imaging algorithms for radar target identification," *Progress In Electromagnetics Research*, Vol. 108, 155–175, 2010.
8. Lazaro, A., D. Girbau, and R. Villarino, "Analysis of vital signs monitoring using an ir-UWB radar," *Progress In Electromagnetics Research*, Vol. 100, 265–284, 2010.
9. Lee, K.-C., J.-S. Ou, and M.-C. Fang, "Application of svd noise-reduction technique to PCA based radar target recognition," *Progress In Electromagnetics Research*, Vol. 81, 447–459, 2008.
10. Byrne, D., M. O'halloran, M. Glavin, and E. Jones, "Data independent radar beamforming algorithms for breast cancer detection," *Progress In Electromagnetics Research*, Vol. 107, 331–348, 2010.
11. Conceição, R. C., M. O'halloran, E. Jones, and M. Glavin, "Investigation of classifiers for early-stage breast cancer based on radar target signatures," *Progress In Electromagnetics Research*, Vol. 105, 295–311, 2010.
12. Lazaro, A., D. Girbau, and R. Villarino, "Wavelet-based breast tumor localization technique using a UWB radar," *Progress In Electromagnetics Research*, Vol. 98, 75–95, 2009.
13. Hasar, U. C., "Procedure for accurate and stable constitutive parameters extraction of materials at microwave frequencies," *Progress In Electromagnetics Research*, Vol. 109, 107–121, 2010.
14. Holzrichter, J. F., G. C. Burnett, and L. C. Ng, "Speech articulator measurements using low power EM-wave sensors," *J. Acoust. Soc. Am*, Vol. 103, No. 1, 622–625, 1998.
15. Hu, R. and B. Raj, "A robust voice activity detector using an acoustic doppler radar," *IEEE Workshop on Automatic Speech Recognition and Understanding*, Vol. 27, 319–324, 2005.
16. Quatieri, T. F., K. Brady, D. Messing, and J. P. Campbell, "Exploiting nonacoustic sensors for speech encoding," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 2, 533–544, 2006.
17. Jiao, M., G. Lu, X. Jing, S. Li, Y. Li, and J. Wang, "A novel radar sensor for the non-contact detection of speech signals," *Sensors*, Vol. 10, No. 5, 4622–4633, 2010.

18. Bellomo, L., S. Pioch, M. Saillard, and E. Spano, “Time reversal experiments in the microwave range: Description of the radar and results,” *Progress In Electromagnetics Research*, Vol. 104, 427–448, 2010.
19. Polivka, J., P. Fiala, and J. Machac, “Microwave noise field behaves like white light,” *Progress In Electromagnetics Research*, Vol. 111, 311–330, 2011.
20. Guo, B. and G. Wen, “Periodic time-varying noise in current-commutating cmos mìxers,” *Progress In Electromagnetics Research*, Vol. 117, 283–298, 2011.
21. Boll, S. F., “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics Speech and Signal Processing*, Vol. 27, No. 2, 113–120, 1979.
22. Berouti, M., R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, Vol. 4, 208–211, 1979.
23. Mallat, S., *A Wavelet Tour of Signal Processing*, A Harcourt Science and Technology, Academic-Press, 1999.
24. Strang, G. and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, 1996.
25. Chong, N. R., I. S. Burnett, and J. F. Chicharo, “A new waveform interpolation coding scheme based on pitch synchronous wavelet transform decomposition,” *IEEE Trans. Speech Audio Process*, Vol. 8, No. 3, 345–348, 2000.
26. Srinivasan, P. and L. Jamieson, “High-quality audio compress using an adaptive wavelet packet decomposition and psychoacoustic modeling,” *IEEE Trans. Signal Procession*, Vol. 46, 1085–1093, 1998.
27. Deng, H. and H. Ling, “Clutter reduction for synthetic aperture radar imagery based on adaptive wavelet packet transform,” *Progress In Electromagnetics Research*, Vol. 29, 1–23, 2000.
28. Alyt, O. A. M., A. S. Omar, and A. Z. Elsherbeni, “Detection and localization of RF radar pulses in noise environments using wavelet packet transform and higher order statistics,” *Progress In Electromagnetics Research*, Vol. 58, 301–317, 2006.
29. Hatamzadeh-Varmazyar, S., M. Naser-Moghadasi, E. Babolian, and Z. Masouri, “Calculating the radar cross section of the resistive targets using the haar wavelets,” *Progress In Electromagnetics Research*, Vol. 83, 55–80, 2008.
30. Tsai, H.-C., “Investigation into time- and frequency-domain emi-induced noise in bistable multivibrator,” *Progress In*

Electromagnetics Research, Vol. 100, 327–349, 2010.

31. Dl, D., "Denoising by soft thresholding," *IEEE Trans. Inform Theory*, Vol. 41, No. 3, 613–627, 1995.
32. Fu, Q. and E. Wan, "Perceptual wavelet adaptive denoising of speech," *Eurospeech*, 578–80, 2003.
33. Ayat, S., M. T. Manzuri-Shalmani, and R. Dianat, "An improved wavelet-based speech enhancement by using speech signal features," *Computers and Electrical Engineering*, Vol. 32, 411–425, 2006.
34. Sheikhzadeh, H. and H. Abutalebi, "An improved wavelet-based speech enhancement system," *Eurospeech*, 1855–1858, 2001.
35. Mahmoudi, D., "A microphone array for speech enhancement using multiresolution wavelet transform," *Proc. of Eurospeech*, 339–342, Rhodes, Greece, 1997.
36. Gulzow, T., A. Engelsberg, and U. Heute, "Comparison of a discrete wavelet transformation and nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, Vol. 64, 5–19, 1998.
37. Sika, J. and V. Davidek, "Multi-channel noise reduction using wavelet filter bank," *Eurospeech*, 2595–2598, Rhodes, Greece, 1997.
38. Bahoura, M. and J. Rouat, "Wavelet speech enhancement based on time-scale adaptation," *Speech Communication*, Vol. 48, 1620–1637, 2006.
39. Gray, R. M., *Entropy and Information Theory*, Springer, 1990.
40. Wang, J., C. Zheng, X. Jin, G. Lu, H. Wang, and A. Ni, "Study on a non-contact life parameter detection system using millimeter wave," *Space Medicine & Medical Engineering,* Vol. 17, No. 3, 157–161, 2004.
41. Udrea, R. M., S. Ciochina, and D. N. Vizireanu, "Multi-band bark scale spectral over-subtraction for colored noise reduction," *International Symposium on Signals, Circuits and Systems*, Vol. 1, 311–314, 2005.
42. Ghanbari, Y., M. Reza, and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, Vol. 48, 927–940, 2006.
43. Blanco, S., A. Figliola, R. Q. Quiroga, O. A. Rosso, and E. Serrano, "Time-frequency analysis of electroencephalogram series. III. Wavelet packets and information cost function," *Phys. Rev.*, Vol. 57, No. 1, 932–940, 1998.

44. Shannon, C. E., "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27, Nos. 379–423 and 623–656, 1948.
45. Rosso, O. A., S. Blanco, J. Yordanova, V. Kolev, A. Figliola, M. Schürmann, and E. Basar, "Wavelet entropy: A new tool for analysis of short duration brain electrical signals," *Journal of Neuroscience Methods*, Vol. 105, No. 1, 65–75, 2001.
46. Bahoura, M. and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Process. Lett.*, Vol. 8, 10–12, 2001.
47. Johnstone, I. and B. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist. Soc.*, Vol. 59, No. 2, 319–351, 1997.
48. Donoho, D. and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, Vol. 81, 425–455, 1994.
49. Donoho, D. and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, Vol. 90, No. 432, 1200–1224, 1995.
50. Rangachari, S. and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, Vol. 48, 220–231, 2006.