www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

ARCHIVES OF ACOUSTICS
36, 3, 519–532 (2011)

DOI: 10.2478/v10168-011-0037-5

Speech Enhancement Based on the Multi-Scales and Multi-Thresholds of the Auditory Perception Wavelet Transform

Zhi TAO[(1),(2)], He-Ming ZHAO[(1)], Xiao-Jun ZHANG[(2)], Di WU[(1),(2)]

Soochow University
[(1)] *School of Electronic Information*
[(2)] *School of Physical Science and Technology*
Suzhou 215006, China
e-mail: hmzhao@suda.edu.cn

(*received November 17, 2010; accepted March 7, 2011*)

This paper proposes a speech enhancement method using the multi-scales and multi-thresholds of the auditory perception wavelet transform, which is suitable for a low SNR (signal to noise ratio) environment. This method achieves the goal of noise reduction according to the threshold processing of the human ear's auditory masking effect on the auditory perception wavelet transform parameters of a speech signal. At the same time, in order to prevent high frequency loss during the process of noise suppression, we first make a voicing decision based on the speech signals. Afterwards, we process the unvoiced sound segment and the voiced sound segment according to the different thresholds and different judgments. Lastly, we perform objective and subjective tests on the enhanced speech. The results show that, compared to other spectral subtractions, our method keeps the components of unvoiced sound intact, while it suppresses the residual noise and the background noise. Thus, the enhanced speech has better clarity and intelligibility.

Keywords: speech enhancement, low SNR, auditory perception wavelet transform, unvoiced enhancement, masking effect.

1. Introduction

In all kinds of practical applications in speech processing systems, the decreasing of speech quality due to background noise is a very common phenomenon. For example, we hope that the SNR of a speech signal is as high as possible in mobile communications and speech recognition. Therefore, we try to remove the uncorrelated noise and improve the SNR in the speech signal. Speech enhancement has thus become an important issue for researchers. The common technology for dealing with bandwidth noise is spectral subtraction (Boll, 1979). However,

this will produce an annoying "musical noise". This musical noise is the residual noise that exists in the enhanced speech through spectral subtraction. This is similar to the sound of rhythmic music. The modified form of spectral subtraction (Ephraim, Malah, 1984) aims to change the subtraction parameters in order to reduce the noise and trade off between speech distortion and musical residual noise. It is also limited by hybrid optimization parameters. In order to solve this problem, researchers such as Berouti (1979) (Ephraim, Malah, 1985) (Lockwood, Boudy, 1992) propose one method which solves the problem of residual noise. This method, however, has difficulties in the selection of gate-thresholds. When the SNR is low, there is still much residual noise.

In recent years, algorithm models related to peripheral auditory effects have been proposed. They have made some progress in speech enhancement. The researchers (Virag, 1999) (Tao *et al.*, 2005). N. Virag (1999), for example, make use of the human ear's auditory masking effect to modify the spectral subtraction. This method effectively suppresses the residual noise, while improving both the speech SNR and the auditory quality of speech. But when the SNR is low, the high frequency suffers a great loss. Countering the unvoiced enhancement, many researchers argue that the wavelet transform can solve it. Wavelet theory is a newly developed time-frequency analysis, especially fit for non-stationary time-varying signals. Xu *et al.* (1994) first proposed the idea of removing the noise by making use of different scales in the signal space during the process of wavelet transform decomposition; Mallat, Zhong (1999) states that by finding the maximum point of the wavelet transform coefficients, and then reconstructing the signal, it can better approximate the original signal before the noise pollution, i.e., we are able to suppress the noise when conducting the threshold processing of the polluted wavelet transform parameters of the signal. Donoho (1995) proposes a new wavelet-domain denoising technology based on the notion of threshold processing. When improving the speech SNR, his method can better keep the unvoiced components of the speech signals.

The use of the wavelet transform and the wavelet-packet transform to analyse a speech signal reveals the signal features. As we know, signal features cannot be detected in a number of sub-spaces with different resolutions. This point, however, is unimportant for speech analysis because the auditory perception itself consists of extra information. Moreover, whether it is a binary transform, a wavelet-packet transform or an M-band wavelet transform (Zhu *et al.*, 2003; Seok, Bae, 1997; Shen, Jin, 2000), the division between frequency domains is that of one octave. It does not exactly coincide with the human ear's inherent perceptual characteristics in the frequency domain.

Therefore, this paper proposes a speech enhancement based on the auditory perception wavelet transform (Tao *et al.*, 2008; Tao *et al.*, 2010). Its basic function is the best way to overcome the uncertainty in the time-perception frequency. Furthermore, its expansion and contraction of the analysis scale changes according to the concept of critical band, which makes bandwidth a "frequency group"

in every scale. The wavelet transform obtained by this kind of construction undoubtedly has analysis features that are in good agreement with the auditory system. The enhanced speech has better clarity and a higher intelligibility. However, the effect is not ideal if we directly make use of threshold processing because of the uniqueness of the speech signal. This is because the voiceless section of speech contains relatively large high-frequency components. It will thus be regarded as noise and be removed when doing the threshold processing. Taking this reason into account, we first make a voicing decision before the threshold processing. If the speech is voiced sound, we make use of an auditory masking threshold to process it, and if the speech is unvoiced sound, we make use of a soft threshold function to process it. This will keep the important information in the unvoiced sound while suppressing the noise. The experimental results show that this combination can improve the SNR of a speech signal, suppress background noise, and protect the high frequency components in the speech signal as well.

2. The principles of speech enhancement

Assume that the noisy speech signal $y(n)$ can be expressed as: $y(n) = s(n) + d(n)$, where $s(n)$ is the clean speech signal and $d(n)$ is the additive noise. As the enhancement is carried out according to the frame, the above model can be written in the form of the frame:

$$y(m,n) = s(m,n) + d(m,n), \qquad m = 1,2,\ldots,N, \quad n = 0,1,\ldots,N-1, \tag{1}$$

where m is the number of the frame and N is the length of the frame. Applying the Fourier transform to the formula, we get:

$$Y(m,k) = S(m,k) + D(m,k). \tag{2}$$

Suppose that the estimate energy of $D(m,k)$ is $P_n(m,k)$, and smooth the noisy speech frame by frame

$$\left|\overline{Y(m,k)}\right|^2 = \rho\left|\overline{Y(m-1,k)}\right|^2 + (1-\rho)\left|Y(m,k)\right|^2, \qquad (\rho \le 0.9). \tag{3}$$

The obtained enhanced spectrum amplitude of the speech is then:

$$\left|\widehat{S}(m,k)\right| = \begin{cases} \sqrt{\beta(m,k)\cdot P_n(m,k)}, \\ \qquad\qquad \text{if} \quad |Y(m,k)|\,R(m,k) \le \sqrt{\beta(m,k)\cdot P_n(m,k)}, \\ |Y(m,k)|\cdot R(m,k), \qquad \text{else}, \end{cases} \tag{4}$$

where $R(m,k) = 1 - \sqrt{\alpha(m,k)\dfrac{P_n(m,k)}{|Y(m,k)|^2}}$. $\alpha(m,k)$ and $\beta(m,k)$ are the function of the time and frequency, respectively. $\alpha(m,k)$ $(\alpha(m,k) > 1)$ is the over-reduction factor. The increase of $\alpha(m,k)$ can increase the signal to noise ratio

of the processed speech, but it also increases the musical noise and hearing distortion at the same time. $\beta(m,k)$ is the factor of the background noise. It can easily mask the residual music noise by introducing the appropriate background noise. The formula is:

$$\alpha(m,k) = \frac{T(m,k)}{\Gamma(\lambda/2+1)\left|(\xi(m,k)-\rho(m,k)+1)\,E|D^2_{m,k}|\right|}, \tag{5}$$

$$\beta(m,k) = 1 + \left(\frac{\alpha(m,k)-1}{\alpha(m,k)}\right)\xi^{\lambda/2}(m,k), \tag{6}$$

where $\rho(m,k) = E[Y^2_{m,k}]/E[D^2_{m,k}]$ is the posterior SNR of the speech signal and $\xi(m,k) = E[X^2_{m,k}]/E[D^2_{m,k}]$ is the prior SNR of the speech signal. $T(m,k)$ is the auditory masking threshold.

3. Speech enhancement based on the multi-scales and multi-thresholds of the auditory perception wavelet transform

3.1. Auditory perception wavelet transform

Existing research and experimental results show that the human ear's basement membrane has the role of frequency selection and tuning to external sound signals. In different center frequencies, the signal will cause the basement membrane's to vibrate at different locations that correspond to the critical band. The hearing threshold between 20 Hz and 16000 Hz consists of 24 critical bands. These kinds of critical bands have a stable bandwidth on the Bark scale. This is related to perception effects such as the masking effect. If we regard b as the critical band rate for the reason that the scale is a Bark scale, then its relationship with the frequency can be regarded as (Traunmuller, 1990):

$$b(\text{Bark}) = 13\arctan[0.76f(\text{kHz})] + 3.5\arctan\left[f(\text{kHz})/7.5\right]^2, \tag{7}$$

where b is the Bark frequency and f is the linear frequency.

Figure 1 shows the auditory perception wavelet cluster diagram of the linear frequency domain.

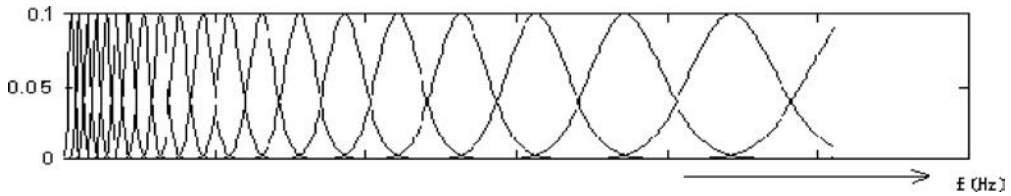

Fig. 1. The auditory perception wavelet cluster diagram of the linear frequency domain.

The required selection form of the wavelet mother-function used to construct the auditory perception wavelet transform wavelet is $W(b) = e^{-c_1 b^2}$, where c_1 is

taken as $4\ln 2$ if the unit bandwidth is defined as 3 dB. The bandwidths of the mother wavelet are all unit bandwidths on the Bark scale, i.e., 1 Bark. At the same time, this mother-function meets the demand that the bandwidth is the smallest found in the Bark domain. Suppose that the linear frequency bandwidth of the speech signal $s(t) \sim S(f)$ to be analyzed is $|f| \in [f_1, f_2]$ and the corresponding Bark frequency bandwidth is $[b_1, b_2]$. We can then define the wavelet function in Bark fields as:

$$W_k(b) = W(b - b_1 - k\Delta b), \qquad k = 0, 1, \ldots, k-1, \tag{8}$$

where Δb is the translation step length of $W_k(b)$. According to the bandwidth features in the Bark domain, it is always the case that $\Delta b = (b_2 - b_1)/K-1$, where K is the scale parameter.

We can transfer $W_k(b)$ to the form of the auditory perception wavelet transform $W_k(f)$ in the linear frequency according to the formula (10), so as to obtain the auditory perception wavelet transform in the definition of the frequency domain.

$$s_k(t) = BW_k(t) = \int_{-\infty}^{\infty} S(f)W_k(f)e^{j2\pi ft}df, \tag{9}$$

where $S(f)$ is the spectrum of the speech signal $s(t)$. Figure 2 shows the system block diagram of the auditory perception wavelet transform.

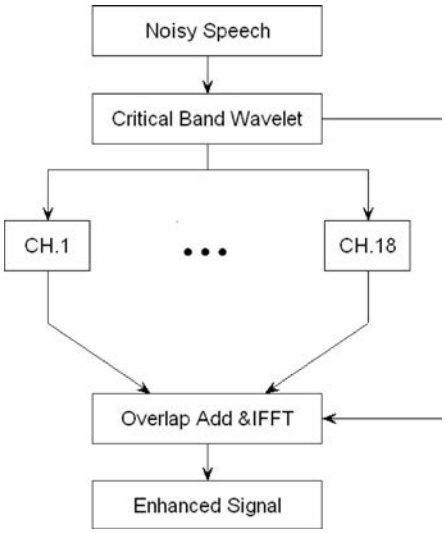

Fig. 2. The system block diagram of the auditory perception wavelet transform.

3.2. Voicing decision

The actual waveform of a speech signal is complex. It contains both the components of voiced sound and unvoiced sound and the "silent zone" of pure

noise. As we know, voiced sound has a large amplitude, which displays an obvious periodicity in the time domain and it also has a resonant peak structure in the frequency domain. Otherwise, unvoiced sound has a small amplitude, which does not show clear characteristics in the time and frequency domains, but it has features similar to random noise. During speech enhancement, we can make use of the periodicity of voiced sound to extract the speech component or suppress the noise by using the comb filter. Thus, unvoiced sound cannot be distinguished from broadband noise.

After applying the auditory perception wavelet transform to the noisy speech signal, it is difficult to separate them during threshold processing, because it contains many high frequency components and the random noises are mixed together. This is what causes the decline in the quality in the speech enhancement, realizing the possibility that the method suppresses the noise while keeping as much of the unvoiced information intact as possible. According to the different distribution of wavelet transform coefficients of unvoiced and voiced signals at different scales, the method of judging voicing is as follows (Hu, 2006):

1. Calculate the average energy band distribution of speech signal. The unvoiced speech has a strong energy in the higher frequency domain, while the voiced speech is mainly distributed in the lower frequency domain. This is an important feature when we make voicing decision. After decomposing the wavelet of each sub-frame of the speech signal, the frequency band can be divided into two parts: the low frequency part [0–2 kHz] and the high frequency part [2–4 kHz]. This will produce two sets of parameters: low frequency coefficients and high frequency coefficients. We can then make use of the two sets of coefficients to calculate the average energy of the low frequency part and of the high frequency part, respectively. This makes it possible to calculate the energy ratio of the low frequency part and of the high frequency part for each frame. The ratio value can be defined as E_i.
2. The high frequency components of the speech signal have a higher zero crossing rate, while the low frequency components have a lower zero crossing rate. Because of this, the zero crossing rate of unvoiced speech is high, while the zero crossing rate of voiced speech is low. We then calculate the short-time zero crossing rate of each frame speech signal. The value is defined as Z_i. Then, we regard the median of all the frames' zero crossing rate in the speech segment as a threshold and define the threshold as T_h. This gives us $T_h = median(Z_i)$.
3. When $E_i \leq 0.3$, we can judge this frame as voiced speech; when $E_i \geq 0.9$, we can regard this frame as unvoiced speech.
4. If E_i does not meet the demands of (3), we then perform a sub-judgment. If $Z_i > T_h$, we can regard it as unvoiced speech, otherwise it is voiced speech.
5. As voiced segment and a non-voiced segment often overlap in the speech signal. Thus, the above algorithm sometimes will produce errors and we need to do some post processing. As we know, the short non-voiced segment cannot appear in consecutive sections. For example, if the decision result is

$\cdots VVVUVV\cdots$, where V represents a voiced segment and U stands for a non-voiced segment, then the result should be modified as $\cdots VVVVVV\cdots$, or *vice versa*.

3.3. The auditory mask thresholding of voiced sound

After the analysis of applying the auditory perception wavelet transform to the speech signals, we first get the power spectrum $E(m,k)$ of the speech signal using FFT and then divide the frequency domain of the speech signal into different critical frequency bands. The speech's energy in each critical frequency band is $M_i = \sum_{k=ml_i}^{mh_i} E(m,k)$, where mh_i and ml_i represent the upper limit and the lower limit of the critical frequency band i, respectively. $i = 1, 2, \ldots, i_{\max}$ is the number of the frequency band. Taking the mutual masking effect of each frequency band into account, we define the following spread function (Johnston, 1998) which is given to estimate the effects of masking across the critical band.

$$SF_{ij} = 15.81 + 7.5(\Delta + 0.474) - 17.5\sqrt{1 + (\Delta + 0.474)^2}, \tag{10}$$

where $\Delta = i - j$, $i, j = 1, 2, \ldots, i_{\max}$, and $|\Delta| \leq i_{\max}$.

Taking the effected energy between the frequency bands into account, we define:

$$C_j = \sum_{i=1}^{i_{\max}} M_i \cdot SF_{ij}, \qquad j = 1, 2, \ldots i_{\max}, \tag{11}$$

where C_j is the power spectral of the expanded Bark domain.

Because of the different masking characteristics between noise and tone, we should first determine whether each frequency band is noise or tone. This can be determined according to the spectral flatness SFM $= \mu_g/\mu_a$, where μ_g and μ_a represent the geometric mean value and the arithmetic mean value of the power spectrum in every frequency band and SFM $\in [0, 1]$. When the value of SFM is 0, it shows that it has the characteristics of pure tone and the masking threshold offset of the pure tone is $(14.5+i)$ dB. When the value of SFM is 1, it shows that it has the characteristics of white noise and the masking threshold offset of the white noise is 5.5 dB. According to the definition of SFM, we define the tone coefficient as $\phi = \min(\text{SFM}_{\text{dB}}/-60.1)$. The relative masking threshold offset is then $O_i = \phi(14.5+i) + 5.5(1-\phi)$ dB. At this point, the masking threshold is $T(m,i) = 10^{\lg(C_j - (O_i/10))}$, where C_j is the power spectral of the expanded Mel domain and O_i is the masking threshold offset. In every Mel frequency band, the speech signal has the same masking features. $T(m,i)$ is then expanded to every frequency band, written as $T'(m,k)$. The final masking threshold is thus $T(m,k) = \max(T'(m,k), T_a(m,k))$, where $T_a(m,k)$ is the absolute hearing threshold.

3.4. The thresholding of the soft threshold function of unvoiced sound

The main information of a speech signal is stored in the unvoiced sound, while most of the unvoiced sound's energy is distributed in the high frequency part. Once the unvoiced sound is damaged, even if most of the speech's energy can be maintained, the intelligibility of a sentence can still be reduced. In order to solve this problem, this paper makes use of a threshold function to process the threshold of the unvoiced sound. There are two schemes to select the threshold function: the hard threshold function and the soft threshold function. They are as follows:

$$\mathrm{THY}(X,T)_h = \begin{cases} X, & |X| > T, \\ 0, & |X| < T, \end{cases} \tag{12}$$

$$\mathrm{THY}(X,T)_S = \begin{cases} \mathrm{sgn}(X)(|X| - T), & |X| > T, \\ 0, & |X| < T, \end{cases} \tag{13}$$

where X represents the noisy speech signal and T represents the removed-noise threshold. The method using the hard threshold function to denoise mainly keeps the low frequency components of the signals. Though it effectively eliminates the noise parts that display the high-frequency signals, it also removes a large number of the unvoiced information from the speech signal. Thus, the selection of the threshold T during the denoising process directly affects the results. Although the unvoiced speech can be retained when the threshold T increases, the residual noises in the reconstruction signal increase. In order to better denoise and keep the unvoiced sound, we make use of a new threshold

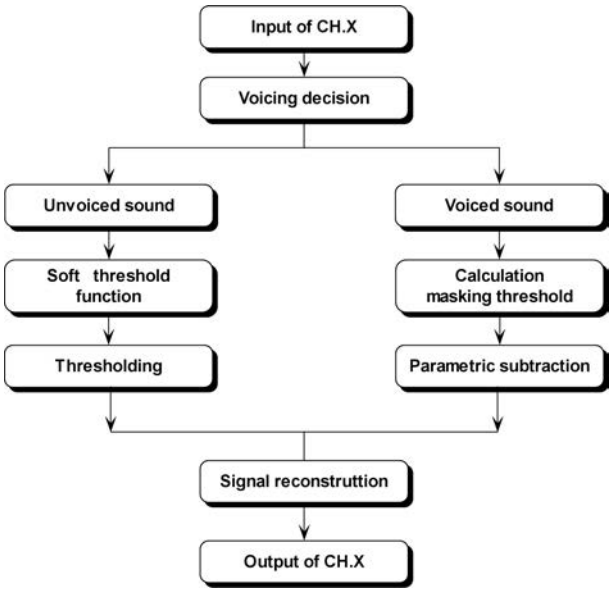

Fig. 3. The flow chart of the speech enhancement system in each critical band.

$T_j = \widehat{\sigma}\sqrt{2\ln(N\log_2(N))}/\ln(j^2{+}1)$ with the changing of scales to denoise the unvoiced sound, where j is the decomposition scale of the auditory perception wavelet. With the increase of j, this threshold decreases according to the non-linear approach. In the high frequency band that contains a lot of unvoiced information, the threshold reduced by the non-linear approach is very beneficial to keeping the unvoiced components of the high-frequency signals. Figure 3 shows the flow chart of the speech enhancement system in each critical band.

4. Results and performance evaluation

We make use of the standard speech database CASIA (http://www.chinese-ldc.org/). The noise data is from the noise database NoiseX-92 (http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html), and the sampling frequency of the signal is 8 kHz, 16 bit. The length of the frame is 256. The window sliding is 128. We use a Hamming window as the data window.

4.1. Voicing decision

In order to show the effectiveness of voicing classification using this paper's algorithm, we can compare the results produced by our proposed algorithm (MMWT), that of the short-time energy method (STE), the short-time zero crossing rate (STZ) and the energy-zero combination decision method (E-Z). As for the algorithm performance, we can adopt the decision accuracy of silent, unvoiced speech, voiced speech (S/U/V) and all kinds of misjudged rates to measure it. When evaluating the performance of an algorithm, two problematic situations may occur. One is that voiced speech is misjudged as unvoiced speech (V/U misjudge). In the other situation, unvoiced speech is misjudged as voiced speech (U/V misjudge). From the experiment, we can see that the algorithm proposed in this paper produces more accurate results.

Table 1. Results of voicing decision.

Methods	S/U/V decision accuracy [%]	V/U misjudge [%]	U/V misjudge [%]	Silence/Speech misjudge [%]
STE	87.21	5.32	4.16	8.21
STZ	78.34	3.11	4.73	11.32
E-Z	90.87	1.78	3.53	7.73
MMWT	94.57	0.45	0.58	5.21

4.2. The improvement of SNR

The experiment adopted the clean speech of the standard speech database CASIA. The background noise was the white noise of NoiseX-92. We made use of

spectral subtraction, modified spectral subtraction, auditory imitating spectral subtraction, and the method presented in this paper as the speech enhancement method. The input SNRs respectively were −6 dB, −3 dB, 0 dB, 3 dB, 6 dB, 9 dB. We also compared the improvement of the output SNRs. We used the common segmented SNR SNR_{seg}, the formula as follows:

$$SNR_{seg} = \frac{1}{N}\sum_{i=0}^{N-1} 10\log\left\{\frac{\sum_{n=0}^{N} x(iN+n)^2}{\sum_{n=0}^{N}\left[x(iN+n)-\widehat{x}(iN+n)\right]^2}\right\}, \tag{14}$$

where $x(n)$ and $\widehat{x}(n)$ are the time domain signals of the clean speech and the denoising speech.

We can see from the experimental results in Fig. 4 that the method in this paper is superior to spectral subtraction, improved spectral subtraction, and auditory imitating spectral subtraction in terms of improving the SNR. When the input speech has a low SNR, the improvement effect of SNR presented in this paper is more obvious.

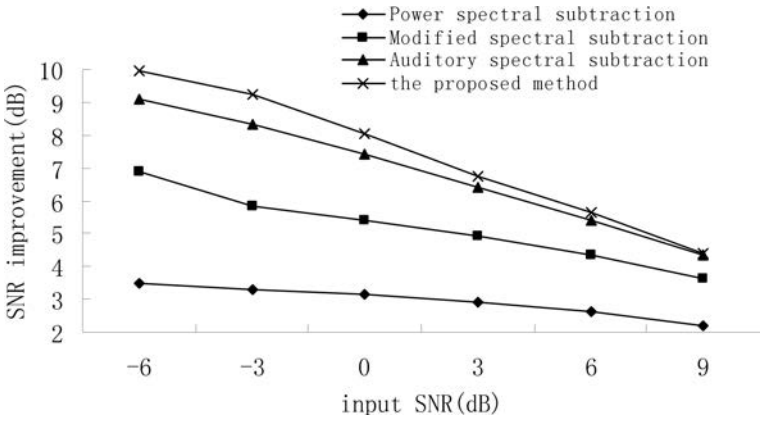

Fig. 4. The improvement comparison of the speech's SNR after the enhancement using the different methods.

4.3. The spectrogram

The objective test cannot indicate the details of the residual music noise, while the spectrogram can better display the details of the residual noise. In order to enhance the effect, this paper makes use of spectral subtraction, modified spectral subtraction, auditory imitating spectral subtraction, and the method of this paper, respectively. Figure 5 shows the spectrogram of the noisy speech "ta qu wu xishi wo dao Heilongjiang" (the SNR is −5 dB), and Fig. 6 shows the spectrograms after the enhancement using the different methods.

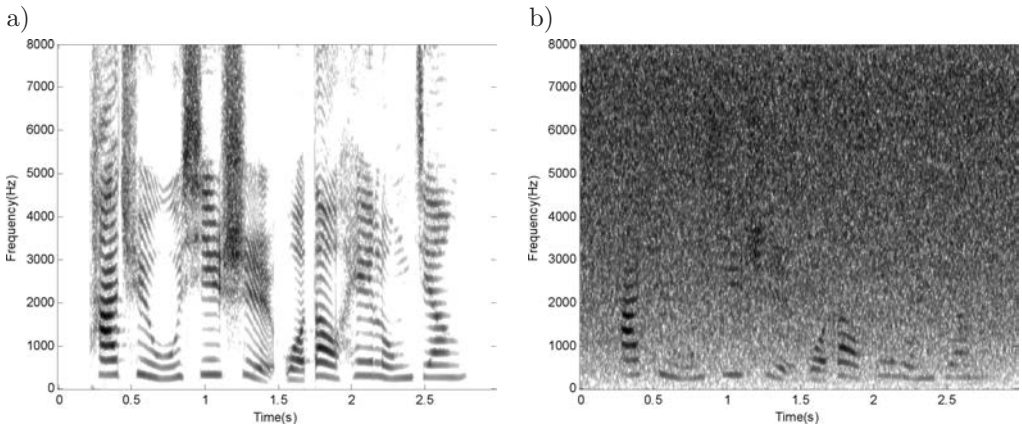

Fig. 5. The spectrograms of pure speech (a) and of noisy speech (b).

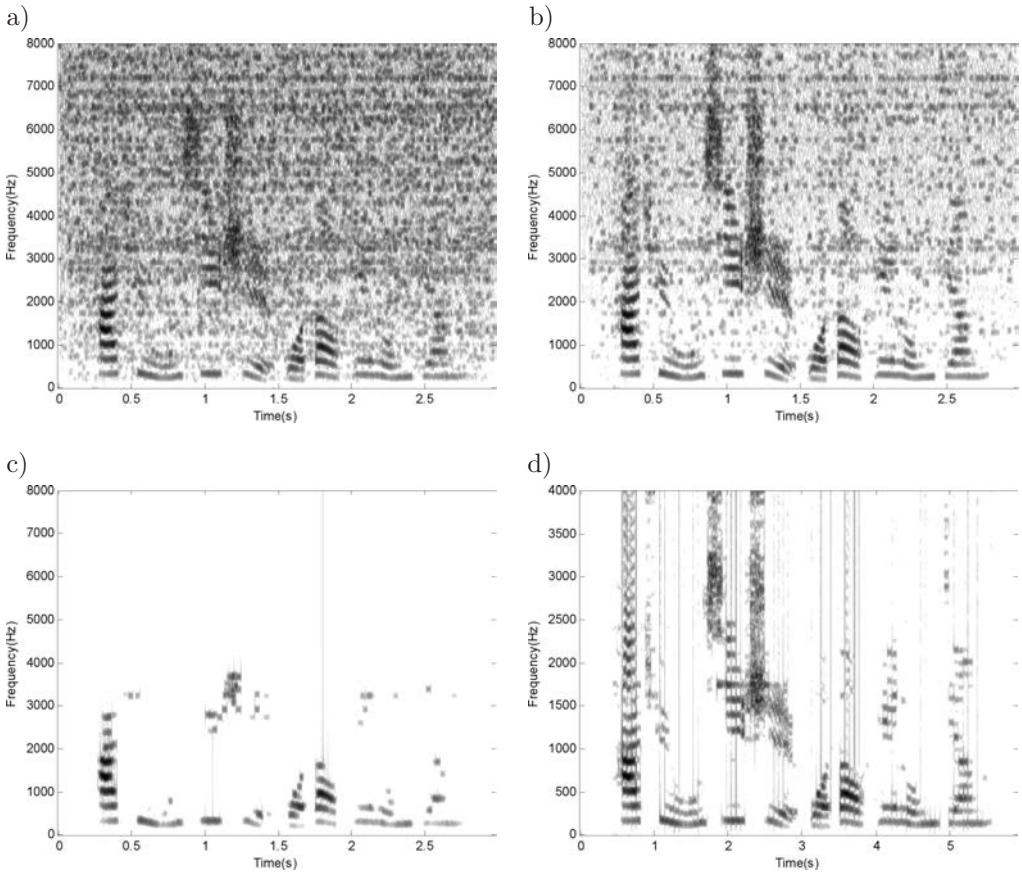

Fig. 6. The spectrograms after the enhancement using the different methods: a) spectral subtraction, b) modified spectral subtraction, c) auditory imitating spectral subtraction, d) the method of this paper.

After comparing the spectrograms after the enhancement using the different methods, we can see that the basic spectral subtraction introduces large amounts of residual musical noise and that the modified spectral subtraction has a slight improvement. The auditory imitating spectral subtraction reduces the noise and the residual noise significantly, but this method results in a serious loss in the unvoiced sound, so as to cause a decline in the speech enhancement's quality. The method of this paper can better make up for this shortfall. From Fig. 6d, we see that the unvoiced sound can mostly be kept in the high frequency.

4.4. The subjective evaluation

In addition to the objective evaluation, we also did a subjective evaluation of the enhanced speech. The test was conducted with 20 listeners (10 males and 10 females), with the aim being to understand the residual noise, the background noise, and the distortion of the speech. We adopted the mean opinion score (MOS) to evaluate the speech quality after the enhancement using the different methods. The test results are shown in Fig. 7.

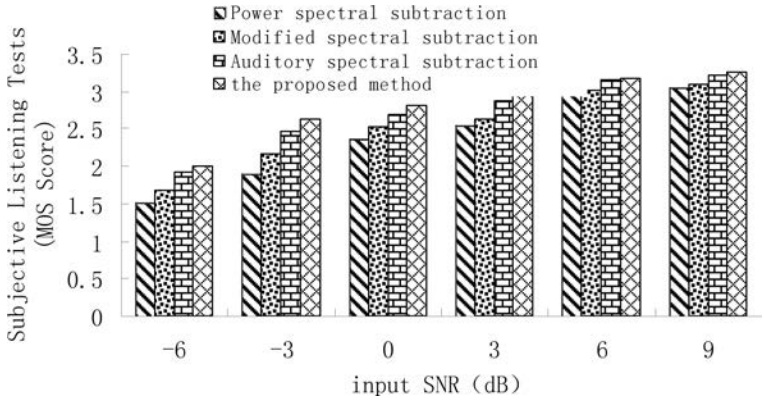

Fig. 7. Subjective hearing tests (MOS points).

The test results show a definite improvement in the auditory quality of the speech enhancement proposed by us in this paper. The speech enhancement's effect is also better than those of the other spectral subtractions. In the case of lower SNR, the speech's interferences affected by the residual noise are smaller than that of the other methods, and the quality of the enhanced speech has greatly improved.

4.5. Intelligibility test

In order to verify the audio quality of the speech enhancement algorithm, we make use of PESQ (Perceptual Evaluation of Speech Quality) as the speech qual-

ity's evaluation criterion when conducting the intelligibility test of the enhanced speech. The PESQ is most relevant for the sound quality assessment of the algorithm in the objective and subjective evaluation, and is here able to reflect the overall quality of speech. The PESQ represents the contrast speech quality of the tested speech and of the reference speech by a value between -0.5 and 4.5. If the quality of the tested speech is very close to the reference speech, the score of the PESQ is close to 4.5, if not, the score is very low. Table 2 compares the SNR of the output speech produced by the 4 algorithms and their PESQ scores. We see that speech quality after applying the suggested algorithm is better than the quality produced by ordinary spectral subtraction (SS), modified spectral subtraction (MSS) or auditory imitating spectral subtraction (AISS).

Table 2. PESQ scores of the intelligibility test.

Additive speech's SNR [dB]	PESQ scores				
	Noise	SS	MSS	AISS	MMWT
−10	1.075	0.714	1.031	1.142	1.537
−5	1.234	0.876	1.269	1.345	1.785
0	1.387	1.039	1.546	1.612	2.014
5	1.585	1.368	1.923	2.014	2.412
10	1.669	1.902	2.127	2.453	2.746

5. Conclusion

When the SNR of noisy speech is very low, it will produce an annoying "musical noise" as we conduct speech enhancement using spectral subtraction. In order to solve this problem, we propose a speech enhancement method based on the multi-scales and multi-thresholds of the auditory perception wavelet transform. By decomposing the noisy speech using the auditory perception wavelet transform, we achieve the goal of noise reduction while combining the speech signal's voicing identification and multi-thresholds processing. Simulation results show that the method of this paper has both a better denoising effect on voiced noise, and can also better keep the unvoiced signal in the high frequency band. Under the different conditions of input SNRs, the algorithm of this paper has a higher output SNR than other spectral subtractions, while the auditory quality of the enhanced speech also has been improved significantly.

References

1. Boll S. (1979), *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 2, 113–120.
2. Berouti M., Schwartz R., Makhoul J. (1979), *Enhancement of speech corrupted by acoustic noise*, Proc. IEEE ICASSP, Washington, DC, 208–211.

3. Donoho D.L. (1995), *De-noising by soft-thresholding*, IEEE Transactions inform Theory, **41**, 3, 613–627.

4. Ephraim Y., Malah D. (1984), *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **32**, 6, 1109–1121.

5. Ephraim Y., Malah D. (1985), *Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator*, IEEE Trans Acoust Speech Signal Processing, **33**, 2, 443–445.

6. Hu Y., Chen N. (2006), *A Method of Unvoiced/Voiced Classification and Pitch Detection Based on Wavelet Transform*, Audio Engineering, 11, 63–66.

7. Johnston J.D. (1998), *Transform Coding of Audio Signals Using Perceptual Noise Criteria*, IEEE Transactions on Selected Areas in Communication, **6**, 2, 314–323.

8. Lockwood P., Boudy J. (1992), *Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and Projection for Robust Recognition in Cars*, Speech Communication, 11, 215–228.

9. Mallat S.G., Zhong S. (1999), *Singularity detection and processing with wavelet*, IEEE Transactions Inform Theory, **38**, 2, 517–543.

10. Seok J.W., Bae K.S. (1997), *Speech enhancement with reduction of noise components in the wavelet domain*, Copyright 1997 IEEE, 1323–1326.

11. Shen Y.Q., Jin H.Z. (2000), *Speech enhancement based on wavelet transform*, Bulletin of Science and Technology, **16**, 3, 206–211.

12. Tao Z., Zhao H.M., Gong C.H. (2005), *Speech enhancement based on masking properties of human auditory system and bark wavelet transform*, Acta Acustica, **30**, 4, 367–372.

13. Tao Z., Zhao H.M., Gu J.H., Tan X.D., Wu J. (2008), *Speech feature extraction of cochlear implants on the basis of auditory perception wavelet transform*, IEEE ICALIP, Shanghai, 80–85.

14. Tao Z., Zhao H.M., Wu J., Gu J.H., Xu Y.S., Wu D. (2010), *A lifting wavelet domain audio watermarking algorithm based on the statistical characteristics of sun-band coefficients*, Archives of Acoustics, **35**, 4, 481–491.

15. Traunmuller H. (1990), *Analytical expression for the tonotopic sensory scale*, Journal of the Acoustical Society of America, 88, 97–100.

16. Virag N. (1999), *Single Channel Speech Enhancement Based on Masking Properties of Human Auditory System*, IEEE Transactions on Speech and Audio Processing, **7**, 2, 126–137.

17. Xu Y.S., Weaver J.B., Healy D.M., Lu J. (1994), *Wavelet transform domain filters: a sparial selective noise filtration technique*, IEEE Transactions Image Processing, **3**, 6, 747–758.

18. Zhu X.W., Yang D.C., Wang W., Mou F., Xu B.L. (2003), *The research on speech enhancement based on the simulation of auditory model using frame-synchronized combined wavelet packet transform algorithms*, Acta Acustica, **28**, 1, 12–16.