The algorithm of noise subtraction for cochlear implant based on improved auditory perception wavelet packet[*]

Jingru Huang [1], Guanyu Tian [2], Guoqiang Qi[1], Shuzhong Bai [2], Shibin Du[1], Lan Tian[1*]

[1] School of information science and engineering Shandong University, Jinan 250100, China
[2] School of electrical engineering, Shandong University, Jinan 250061, China
*Corresponding author: Lan Tian，lant65@163.com

Abstract - Under the noisy background environment, the performance of cochlear implant (CI) will decrease rapidly. Wavelet transform is the effective analyzing tool for the complicated speech signal. But the traditional wavelet transform has the disadvantage of 'over threshold' that leads some effective speech parts lost in the enhanced speech. In order to raise the anti-noise ability of CI and retain richer and clearer speech signal, an improved speech enhancement algorithm is presented and applied into CI. In this method, by using the perceptual wavelet packet transform, the noisy speech signal was decomposed into wavelet packet node coefficients and the time adaptive threshold (TAT) was introduced in the process of noise subtraction. For each frame signal, based on the estimated speech-presence probability, the TAT of subtracting noise was adjusted along with time so that the noise signal can be effectively suppressed while much more useful speech component was retained in the recovered signal, especially for the endpoint parts of speech signal. Based on the ACE processing strategy, the enhanced speech signal was analyzed and synthesized in the simulation system of CI. The simulation experiment results show that the proposed algorithm can produce clearer and better synthesized speech in CI than that of the traditional wavelet transform algorithm.

Index Terms - Noise subtraction; Cochlear implant; Auditory perception; Wavelet packet transform.

1. Introduction

Electronic cochlear implant (also called CI) is an electronic device to restore the hearing for deepness and profound deaf people [1]. The ACE strategy is one of the best and widely used signal processing method in CI. It combines the advantages of SPEAK algorithm and CIS algorithm , which can both track the spectrum of the audio signal more effectively and also offer higher stimulation rate, thereby can obtain better clinical effects [2]. Wavelet transform is based on the theory of multi-resolution analysis and inhomogeneous dividing time-frequency space, which can meet the demands that different time components and different frequency components are taken into account at the same time. But whether binary wavelet, wavelet packet or M band wavelet transform, the fundamental frequency division of them is a double frequency relations, and it mismatch the inherent characteristics of human ear for the perception of voice signal in frequency domain. Based on the above analysis, this paper puts forward to a method based on Bark wavelet transform which agrees well with the auditory system to perform cochlear implant speech signal processing, and we also use the time frame energy that is based on wavelet packet nodes to estimate the probability of speech appeared, and then using the probability value adjustment denoising threshold value, which can effectively solve the problems of traditional wavelet excessive threshold processing. The algorithm is putted in the front of ACE scheme, and do cochlear implant computer simulation experiments, and then contrast to the traditional wavelet packet denoising algorithm so as to verify the improved wavelet packet denosing algorithm for the improvements of final synthetic speech quality.

2. ACE Processing strategy of CI

A. The Mechanism of CI

The electronic cochlea is based on that there are some enough auditory neural fibers can be stimulated existing nearby the electrode. Once the nerve fibers are stimulated, they generate and transfer nerve impulses to the brain, and the brain will explain these pulses as voice. The loudness of sound that can be perceived depends on the number of nerve fibers been stimulated and their intensity of excitement. The number of stimulated nerve fibers is proportional to the stimulated current amplitude, so the intensity of the sound can be controlled by changing the amplitude of stimulated current. On the other hand, frequency component and timbre feeling is related to the stimulated space in the cochlea, that in the top of the cochlea producing bass feeling and producing the high-frequency feeling on the base[3].We usually divide the frequency in this range into 24 frequency group, namely 24 Bark domain. The experimental research further indicates that the time frequency analyzing property of the cochlear is similar to a set of constant quality factor band pass filters.

B. The ACE Strategy of CI

Based on the mechanism of the cochlear implant, the audio signal can be divided in frequency domain according to the properties of auditory perception. The ACE strategy is a kind of waveform coding scheme, which is widely used in the Australia Nucleus 24 [4]. The basic process is shown in Fig.1. At first, the signal is restrained the noise after sampling, being pre-emphasized, and then passed through N band-pass filters. The envelopes of the filtered waveforms are extracted after filtering, then those envelope outputs are finally compressed and then used to modulate biphasic pulses. In the ACE

[*] This work was supported by Shandong University Creation Fund (2011JC018) and NSF of Shandong Province (ZR2012FM030).

strategy, the band-pass filtering and envelope extraction is done through FFT. The Synthetic speech signal has better quality, because the ACE strategy can keep a higher stimulation rate and also improve the frequency resolution of signal decomposition [5].

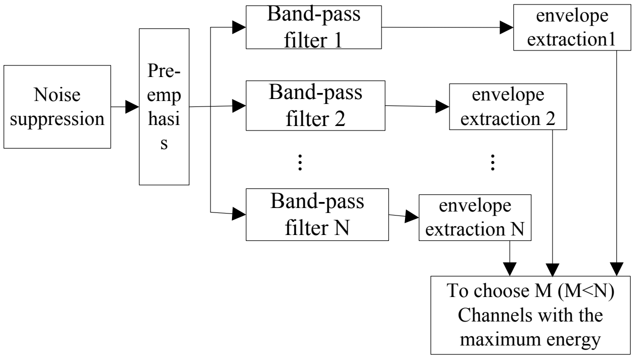


Fig. 1 The ACE strategy diagram of CI.

3. The Principle and Definition of Wavelet Packet Speech Enhancement Based on Time Adaptive Threshold

A. The Principle of Wavelet Packet Transform of Speech Denoising Based on the Human Ear Auditory Perception

The speech enhancement algorithm based on wavelet packet is similar to that based on wavelet. Firstly, determine the wavelet packet decomposition tree. The wavelet packet decomposition coefficient $w_{j,k}$ of each layer is obtained after the wavelet packet transform for the original speech signal. Then, each $w_{j,k}$ is performed by threshold processing .For every wavelet packet decomposition node, the threshold is

$$\lambda = \sigma\sqrt{2\log(N\log_2 N)} \tag{1}$$

Here σ is the noise variance, and N is signal length. Finally, the enhanced speech is obtained by using the wavelet packet inverse transformation for the wavelet packet coefficient which is performed by threshold processing.

In this speech enhancement algorithm, we use the auditory perception wavelet packet decomposition to analyze the speech signals. The basement membrane of the inner ear has the similar effects to the spectrum analyzer. The frequency from 20~16000Hz can be separated into 24 critical bands, each critical band corresponds to a Bark scale, and the frequency signals within the same critical band are being evaluated adding together. And what we call the auditory perception wavelet packet is to use the flexible frequency division method of the wavelet packet to design the wavelet packet decomposition structure which is similar to the division of the human ear 24 Bark domain. Many experiments prove that auditory perception wavelet packet division can not only match the human ear auditory characteristics, but also can use less Bark sub-band instead of a large number of frequency band calculation. And it reduces the computational complexity effectively. Therefore, auditory perception wavelet packet is adopted to decompose the speech signal in this paper [6].For the signal of 16 KHz sampling frequency, its frequency range corresponds to the critical band marked 1~21. Figure 2 shows the corresponding auditory perception wavelet packet decomposition tree. Here $w_{j,k}(j = 1,2,...8, k = 0,1,...2^j - 1)$ represent the kth sub-band coefficient of the jth layer of wavelet packet decomposition. And the center frequency of each adjacent end-nodes differs 1/4 frequency interval.

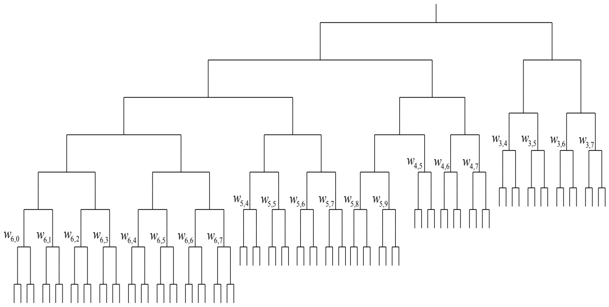


Fig. 2 The wavelet packet decomposition tree based on auditory perception

B. Adaptive Threshold Settings

A power spectrum estimation method is adopted in this paper，which can track the changes of signal power spectrum well. We use the time frame energy of the wavelet packet decomposition node coefficient to estimate the probability of the speech signal frame, and then use the probability to regulate wavelet packet denoising threshold value.

Set the wavelet packet decomposition coefficient as $w_{j,k}(n)$, the time frame energy of wavelet packet coefficient is represented as $E(\lambda)$. λ is the frame number. The length of each time frame is 10 ~ 40 ms, half of the frame overlaps. Firstly, make the first-order smooth processing for the energy of each frame with the following formula.

$$E_p(\lambda) = \eta E_p(\lambda - 1) + (1 - \eta)E(\lambda) \tag{2}$$

Here $E_p(\lambda)$ is the smoothed energy. Then the local minimum of the signal energy is obtained according to the following formula:

$$\begin{cases} E_{\min}(\lambda) = \gamma E_{\min}(\lambda - 1) + \dfrac{1-\gamma}{1-\beta}(E_p(\lambda) - \\ \qquad E_p(\lambda - 1)),\ E_{\min}(\lambda - 1) < E_p(\lambda) \\ E_{\min}(\lambda) = E_p(\lambda) \qquad ,\ \text{other} \end{cases} \tag{3}$$

If the main component part of a frame wavelet packet coefficient is noise coefficient ,then the energy $E_p(\lambda)$ and the corresponding local energy $E_{\min}(\lambda)$ differ a little .And when the wavelet packet coefficient of the frame is speech signal decomposition coefficient, there are lots of differences between $E_p(\lambda)$ and $E_{\min}(\lambda)$ .Therefore, the ratio of $E_p(\lambda)$ and local minimum energy $E_{\min}(\lambda)$ determines that the wavelet packet coefficient of the time frame is mainly speech signal decomposition coefficient or noise decomposition coefficient. Set $S_r(\lambda)=E_p(\lambda)/E_{\min}(\lambda)$, if $S_r(\lambda)$ is larger than the threshold value δ, then suggest that the frame is speech frame, otherwise the frame is noise frame. At the same time we get $p(\lambda)$ which is estimated to be the probability of the speech frame. Then

$$\begin{cases} p(\lambda)=\alpha_d p(\lambda-1)+(1-\alpha_d), & \mathrm{S}_r(\lambda)>\delta \\ p(\lambda)=\alpha_d p(\lambda-1) & \text{, other} \end{cases} \tag{4}$$

Here α_d is smoothing operator, then the wavelet packet coefficient of the frame $P(n)$ is $p(\lambda)$, that is $P(n)=p(\lambda)$, $n=(\lambda-1)\frac{M}{2}\sim\lambda\frac{M}{2}-1$, M is the length of each frame.

According to each wavelet packet node coefficient, the time adaptive threshold value is

$$Thr_{j,k}(n)=\sigma_{j,k}\sqrt{2\log(N)}*(1-P(n)) \tag{5}$$

Here $\delta_{j,k}=\frac{MAD_{j,k}}{0.645}$, $MAD_{j,k}$ is the mid-value of the absolute value of the wavelet packet coefficient of the kth sub-band of jth layer. N is the length of the signal.

4. The Experimental Procedure

The schematic diagram of this method is shown in Fig.3. We present the specific description of the algorithm as follows.

Step1: The noise speech signal is decomposed with the auditory perception wavelet packet which is shown in Fig 2. In order to improve the degree of details of the speech signal, it's necessary to further decompose each sub-band, making the frequency difference between the adjacent sub-bands be a quarter of the Bark sub-band.

Step2: We estimate the voice present possibility for these nodes decomposed by wavelet packets. Then the corresponding sequence can be gotten, here the parameters are measured in experiments. The parameters used in this paper are set as follows: the length of each frame is 32ms, $\eta=0.3$, $\gamma=0.998$, $\beta=0.8$, $\delta=8$, $\alpha_d=0.2$.

Step3: The time adaptive threshold can be obtained based on the coefficients of each wavelet packet node. These thresholds then are further used to be soft threshold to the wavelet coefficient $w_{j,k}$. Let $w'_{j,k}(n)$ be the wavelet coefficient after thresholding, then we get:

$$w'_{j,k}=\begin{cases} \operatorname{sgn}(w_{j,k}(n)(\left|w_{j,k}(n)-Thr(n)\right|) & , w_{j,k}\geq Thr_{j,k} \\ 0 & , w_{j,k}<Thr_{j,k} \end{cases} \tag{6}$$

Step 4: The enhanced speech signal is obtained by applying the inverse wavelet packet transform on $w'_{j,k}(n)$.

Step 5: Do ACE encoding and synthesizing on the enhanced signal.

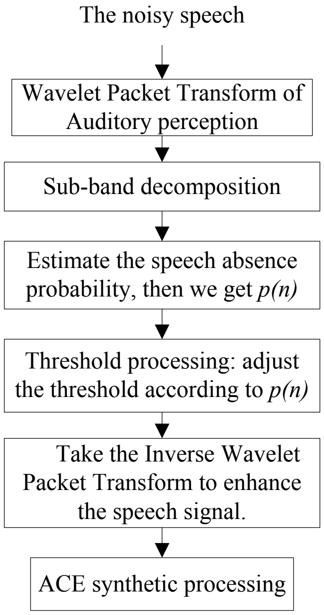
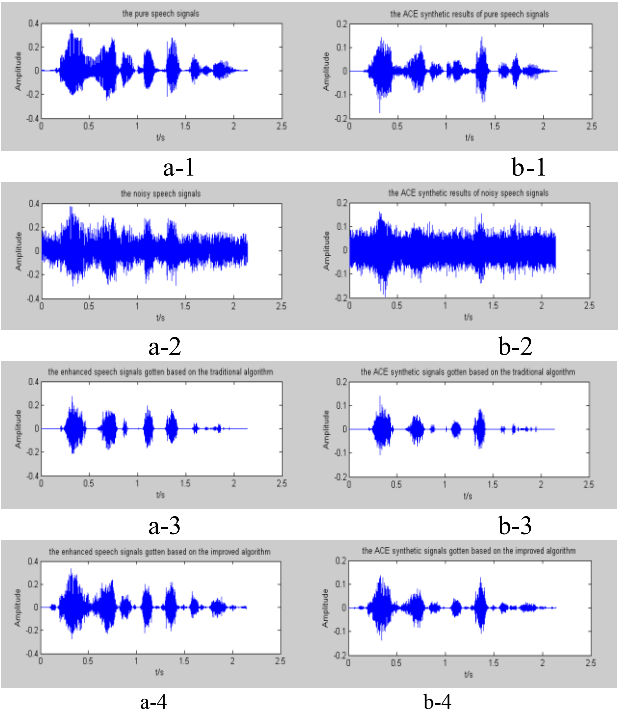

The noisy speech
Wavelet Packet Transform of Auditory perception
Sub-band decomposition
Estimate the speech absence probability, then we get p(n)
Threshold processing: adjust the threshold according to p(n)
Take the Inverse Wavelet Packet Transform to enhance the speech signal.
ACE synthetic processing

Fig. 3 The ACE processing scheme of the cochlear implant simulation based on Wavelet Packet Transform of Auditory perception

5. The Analysis of the Experiment Results

The sampling frequency of the pure speech signal and the noise, used in the experiment, are both 16 kHz. And the sampling precision is 16 bits. The noises that we selected are white noise, factory noise and f16 fighter noise from the Noisex-92 noise library. Firstly, the noise speech signal is operated by framing. The length of each frame is 32 ms, half of the frame overlaps. And then we sent the pure speech signal, the noise speech signal, the enhanced signal based on traditional wavelet enhancement method and the enhanced signal based on the method this paper proposed to the ACE strategy, in order to do CI speech synthesis simulation. There are 21 channels in the ACE coding scheme, in which 8 channels with the maximum energy are selected, and then are used as to generate corresponding stimulated current, and the electrode stimulation interval of each channel is 2ms.

In experiments, the waveform of simulation and SEG-SNR are adopted as the objective assessment method of the quality of speech enhancement. By adding factory noise with a SNR of 0 dB to the speech signal, we see the results shown in Fig 4. We can obviously see that the proposed algorithm achieves a better performance in noise suppression, and highlight the useful parts. In Fig 4, from b-1, b-2, b-3 and b-4, we could draw conclusions that the improved method makes the final synthetic signal is similar to the enhanced ACE synthetic signal of the pure speech signal, accordingly, the intelligibility of the composite signal can be improved.



a-1 the pure speech; a-2 the noisy speech; a-3 the enhanced speech based on the traditional algorithm; a-4 the enhanced speech based on the improved algorithm; b-1the ACE synthetic speech of pure speech; b-2 the ACE synthetic speech of noisy speech; b-3 the ACE synthetic speech based on the traditional algorithm; b-4 the ACE synthetic speech based on the improved algorithm.

Fig. 4 The waveform comparison of the four different kinds speech signal and theirs ACE synthesis signal

TABLE I The Contrast of Segmented Signal-noise Ratios Between the Two Speech Enhanced Methods（unit：dB）

Noise Type	White Noise			F16 Fighter Noise			Pink Noise		
Input SNR	5	0	-5	5	0	-5	5	0	-5
The Traditional Method	3.536	2.276	0.937	2.612	1.334	0.548	3.585	2.360	1.322
The Improved Method	6.581	4.66	1.996	4.387	1.515	-.2535	6.086	3.833	1.3

If $s(n)$ is the speech signal with noise, $\hat{s}(n)$ is the enhanced ACE synthetic signal, the segmented signal-to-noise ratio SRN is

$$SNR_{seg}=\frac{1}{L}\sum_{m=0}^{L-1}10\log 10\frac{\sum_{n=0}^{N-1}\hat{s}^2(n+Nm)}{\sum_{n=0}^{N-1}[s(n+Nm)-\hat{s}(n+Nm)]^2} \tag{8}$$

where L is the frame number, N is the length of each frame data .The table 1 shows the comparison of the segmented SNR of the enhancement results between the traditional wavelet packet denoising algorithm and the improved algorithm introduced in the paper..From the table we see that, in the environment with a variety of noise, improved wavelet packet enhanced result is obviously better than the traditional wavelet packet enhanced result in the segment signal-to-noise ratio.

6. Conclusion

This paper introduce an improved noise estimation algorithm, that is using the auditory perception wavelet packet transform which is consistent with the ear auditory perception characteristics of human to decompose the noise speech signal. And it uses the speech present probability which can better track the wavelet packet coefficient distribution to regulate the denoising threshold value. Then the enhanced speech signal is obtained through soft threshold processing and wavelet packet inverse transformation. A kind of improved time adaptive threshold of wavelet packet speech enhancement method is realized. Compared with the traditional noise estimation algorithm, this algorithm estimates stationary noise power spectrum well and can be well applied into the artificial cochlear spectrum reduction in speech enhancement. For the further research, we can divide the spectrum more closer, for instance, we can exchange 8 channels into 16 channels. Furthermore, the algorithm has a better denoising result for the utterences with many unvoiced sounds. But this method has limitations on some cases when the speech signal is interfered by non-stationary noise. These problems need further researching.

References

[1] Nie KB, St Ickney G, Zeng FG. “Encoding frequency modulation to improve cochlear implant performance in noise”, *IEEE Transact ions on Biomedical Engineering*, 2005, 52 (1): 64-73.

[2] Psarros CE, Plant KL, Lee K, et al, “Conversion from the SPEAK to the ACE strategy in children using the nucleus 24 cochlear implant system: speech perception and speech production outcomes”*, Ear and Hearing*, 2002, 23(18): 18

[3] Chen Wengang, Tian Lan, Jiang Xiaoqing, etc, “A speech enhancement method based on quickly tracking the noise spectrum”, *Journal of Shandong University*, 2006, 36(4): 26-28.

[4] Wang Zhizhong, Cao Keli translate, “Principles and practices of cochlear implants”, Beijing People's Medical Publishing House, 2003.

[5] Tian Yujing, Zuo Hongwei, Dong Yuming, Wei Desheng, “Bark subband wavelet package adaptive threshold speech de-noising method”, *Computer Application*, 1001- 9081(2010) 11- 3111- 04

[6] Sundarrajan R, Philipos CL, “A noise-estimation algorithm for highly non-stationary environment”, *Speech Communication*, 2006, 48(2): 220-231