The effect of gain thresholds on speech intelligibility for statistical model based noise reduction for cochlear implants:A simulation based verification

Wenzhi He[1]*,Nengheng Zheng*[1]*,Qinglin Meng*[2]

[1]College of Information Engineering,Shenzhen University
[2]Acoustic Lab., School of Physics and Optoelectronics,South China University of Technology

hewenzhi@email.szu.edu.cn, nhzheng@szu.edu.cn

Abstract

Noise corruption can dramatically decrease the speech intelligibility for listeners with cochlear implants (CI). Noise reduction is a key point in CI speech processing strategy. This paper proposes a statistical model based noise reduction algorithm forCIs. A realistic noise estimator, which requires no prior knowledge of the noise, is adopted for noise estimation. An improved method for determining the user-specific gain function is proposed, in which the apparent gain threshold is incorporated to compute the optimal parameters, with which the optimal gain function for noise suppression can be determined accordingly. Vocoder simulation perceptual experiments with normal hearing listeners shows that the proposed algorithm can significantly improve the speech intelligibility of the denoised speech.

Index Terms: noise reduction, cochlear implants, gain thresholds

1. Introduction

In speech communications, either human-human or human-machine, target speech is usually corrupted by acoustic interferences which reduce the intelligibility of the target speech. Listeners with normal hearing (NH) have the ability of understanding speech even in noisy environments, due to the high redundancy in speech signal. The robustness of human speech understanding can be explained by the perceptual process of listening, i.e., the auditory scene analysis (ASA) [1]. According to ASA, auditory streaming, glimpsing, spatial cues and linguistic knowledge of the target speech are all utilized to aid the segregation process [2][3]. The ability of noisy speech perception by listeners with cochlear implants (CI), however, is much poorer than that by NH listeners. This is most likely due to the limited frequency, temporal, and amplitude resolution transmitted by the CI devices [4].

Although single channel speech enhancement algorithms produce no significant improvements in speech intelligibility for NH listeners [5], they have shown significant improvements for CI ones [6]. Single channel noise reduction algorithms for CI could be divided into three classes, i.e., spectral subtraction based (SS) [7], statistical model based (SM) [8], and subspace based (SP) [9]algorithms. Although the SP ones have shown good performances in speech enhancement with stationary noise, they have not yet been applied in clinical CI devices due to its computational complexity and the degraded robustness to non-stationary noise [6]. Most of the current CI devices adopt the SS and SM as their noise reduction strategies[8]. The SS ones estimate the short-time spectral magnitude of the target speech by subtracting the spectral magnitude of the noise (adaptively estimated from speech pause segments [10]) from that of the noisy speech. The SM ones typically use the minimum tracking and recursive average methods to estimate the noise spectra that are used to estimate instantaneous signal-to-noise ratios (SNR) of the noisy signal, with which a gain function can be computed to suppress the noise component within the noisy speech [8]. Both SS and SM based methods use a gain function to determine a level of attenuation applied to noisy signal to optimally remove the noise. For example, Mauger et al [11] proposed an ideal binary gain function (IBGF) for SM based noise reduction. Speech perception results with CI users suggested that CI users prefer a larger degree of speech distortion to noise corruption, which is contrast to NH listeners, who prefer noise corruption to speech distortion. Instead of using IBGF, the second experiment of [11] used a smooth parametric Wiener gain function for noise suppression. Both experiments of [11] require prior knowledge of the noise, which is not suitable for real-world applications. What's more, in the second experiment of [11], pre-training experiment is required to determine the user-dependent gain function for noise suppression. To do so, each subject was asked to select the most suitable gain function by changing the threshold value (α) and slope value (β). It is a time-consuming task for subjects to select the most suitable values of α and β.

This paper proposed a more realistic noise reduction algorithm for CI devices.Instead of the ideal noise estimate, the improved minimum controlled recursive algorithm [12] which requires no prior knowledge of noise is adopted. Similar to [11], the parametric Wiener gain function is adopted for optimal gain function determination. Unlike in [11], the perceptual optimal gain function is not selected by varing α and β, but from a single metric called the apparent gain threshold (aGT)[11]. Vocoder simulation experiments were carried out to evaluate the effect of the aGT on the intelligibility of the denoised speech and to examined the potential effectiveness of the proposed noise reduction algorithm for CI speech perception.

2. Method

2.1. The statistical model based noise reduction

In this study, the noise and the target speech is additively mixed, i.e.,

$$y(n) = x(n) + d(n) \tag{1}$$

where $x(n)$, $d(n)$ and $y(n)$ denote the target speech, the noise and the noisy speech, respectively. In spectral domain, (1) can be reformulated as:

$$Y(k,l) = X(k,l) + D(k,l) \tag{2}$$

where Y(k, l), X(k, l) and D(k, l) represent the short-time Fourier transform of $y(n)$, $x(n)$ and $d(n)$ respectively, k and

l denote the frequency bin and the time frame index.

In statistical model based noise reduction, the target speech spectrum is estimated by multiplying the noisy speech spectrum with a gain function $G(k,l)$, i.e.,

$$\hat{X}(k,l) = G(k,l) \cdot Y(k,l) \tag{3}$$

Therefore, the key point is now to compute an optimal gain function such that the noise corruption and speech distortion of the estimated speech can be minimized, or in other words, the intelligibility of the denoised speech can be well maintained.

A classical way to do so is given by [13]. Given $\hat{\lambda}_d$ be an estimate of the power spectrum of noise, i.e., $|D(k,l)|^2$, a priori SNR estimate ξ can be calculated by:

$$\xi(k,l) = \begin{cases} \gamma(k,l) - 1, & \gamma(k,l) > 1 \\ 0, & \gamma(k,l) \leqslant 1 \end{cases} \tag{4}$$

where $\gamma(k,l)$ is the posteriori noise estimate computed as:

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\hat{\lambda}_d(k,l)} \tag{5}$$

A smoothed priori SNR estimate $\hat{\xi}(k,\ l)$ is calculated through the recursive average equation:

$$\hat{\xi}(k,l) = (1-a)\,\xi(k,l) + a\xi(k,l-1) \tag{6}$$

The value of the smoothing factor a was set to be 0.984 in [14]. The gain function can now be computed as:

$$G(k,l) = \left(\frac{\hat{\xi}(k,l)}{\hat{\xi}(k,l) + \alpha}\right)^{\beta} \tag{7}$$

where $G(k,l)$ is called the parametric Wiener gain function, $\hat{\xi}(k,\ l)$ is the smoothed priori SNR, and α and β are the paremetric Wiener variables. The α variable changes the threshold of the smooth gain function and the β variable changes the slope of the gain function.

Therefore, the three parameters, i.e., $\hat{\lambda}_d$, α, and β need to be determined for noise reduction. In [11], an ideal noise estimation, with the assumption that the prior knowledge of noise can be obtained, was applied to estimate the noise spectrum, i.e.,

$$\lambda_d(k,l) = \frac{1}{L}\sum_{i=0}^{L-1} |D(k,l-i)|^2 \tag{8}$$

Besides, user-dependent gain function was determined experimentally by perceptually assessing the quality of denoised speech with different α and β.

To compare the optimal α and β to the gain threshold as in the binary masking gain function, a single metric, named the apparent gain threshold (aGT) was defined to represent the overall function [11]. Experiments with CI users showed that although the optimal gain function varies across subjects, the average preferred gain function has an aGT of 6.8 dB. That is, the optimal gain function can be determined by varying the aGT, rather than by covarying the α and β parameters. However, this results were based on the ideal noise estimation. It might not be suitable for practical applications, where the noise spectrum should be estimated online.

2.2. The proposed algorithm

This study proposed an improved algorithm to overcome the problems above mentioned. Firstly, a realistic noise estimation algorithm, called the improved minimum controlled recursive algorithm (IMCRA) [12], was adopted for noise estimation. The IMCRA was shown to be able to track the noise spectrum variation without prior knowledge of the noise component.

As for the gain function, the aGT is adopted to compute the α (β is preset to be 1)[1], with which the gain function is computed as (7). The aGT is defined as the SNR that attenuates the input signal by half [11]. Replacing $\hat{\xi}(k,\ l)$ with aGT in (7), we have

$$\frac{1}{2} = \left(\frac{aGT}{aGT + \alpha}\right) \tag{9}$$

and

$$\alpha = 10^{(aGT/10)} \tag{10}$$

That is, if the perceptually optimal aGT for each subject can be experimentally determined, the α parameter can be computed according to (10). Finally, the user-dependent perceptual optimal gain function can be computed as (7).

3. Experiments and Results

To evaluate the effect of gain threshold on speech intelligibility and the potential effectiveness of the proposed noise reduction algorithm for noisy speech perception by CI recipients, vocoder simulation perceptual experiments were conducted with normal hearing listeners.

3.1. Experimental setups

Eight normal hearing college students (3 females and 5 males, 18-25 years old), participated in the experiment. Age and gender details of the 8 subjects are given in Table 1. All tests were peformed in sound-proof room using a notebook. The speech materials were played with the notebook via a Roland Quad-Capture UA-55 audio interface and a Sennheiser HD 650 headset, at a comfortable level.

The speech signals were sentences taken from the MHINT (Mandarin Hearing in Noise Test) database [15]. There are two training and twelve test lists, each containing twenty sentences. That is, there are 40 training sentences and 240 test sentences in total. Each sentence consists of ten Mandarin words. Two kinds of noise signal,i.e., babble noise and speech-spectrum shaped noise (SSN), were used as the interference. As in [11], user-dependent SNR was adopted for generating the noisy signal and the SNR is set to be SRT minus 1 dB, where SRT represents the speech reception threshold. The determination of SRT for each user under each noise followed the same procedure as in [16]. The training sentences were used to determine the SRTs. SRTs for each subject with the two noises are also given in Table 1.

For each type of noise, there were six test conditions, i.e., noisy speech without noise reduction (noted as Un), denoised speech with aGT of -5 dB, 0dB, 5 dB, 7.5 dB and 10 dB. Each condition had 20 test sentences. The denoised test sentences (or original noisy sentences for the Un condition) were vocoded and presented to the subjects. Each vocoded sentence could be presented up to 3 times upon the request of the subject. A test sentence was considered correctly recognized if 5 or more words in the sentence were recognized by the subjects. The

[1] According to [14], the optimal β values for most experimental conditions were all close to 1, the β value is therefore set to 1 in this study.

recognition rate for each condition was computed as the number of correctly recognized sentences over the total 160 test sentences (8 subjects, each with 20 sentences).

The vocoder simulation process is as in [17]. The input speech was first divided into 16 frequency bands between 80 and 7999 Hz using sixth-order Butterworth filters. The frequency range was divided equally in terms of the function of Greenwood [18]. The output of each band was half-wave rectified and then low-pass filtered, with cutoff frequency at 250 Hz, to generate the envelope. The resulting envelope was used to modulate the carrier: a sine wave located at the center frequency of the frequency band. Finally, the modulated carriers were level matched to (the output signal from the corresponding band) and summed to produce the vocoded speech.

Table 1: *Data for the 8 subjects who participated in this study. Age measured in years and SRT measured in decibels.*

Subject	Gender	Age	SRT(SSN)	SRT(Babble)
1	male	22	2.0 dB	-2.3 dB
2	male	23	3.2 dB	-1.0 dB
3	male	25	3.0 dB	-1.8 dB
4	female	23	1.8 dB	-1.5 dB
5	female	22	1.4 dB	-3.0 dB
6	female	24	2.6 dB	1.3 dB
7	male	26	2.0 dB	1.5 dB
8	male	25	2.7 dB	-2.1 dB

3.2. Results

3.2.1. Results with SSN

Figure 1 shows speech perception rate for the 6 denoising conditions. The target speech signals were corrupted by speech-spectrum shaped noise. Perception rates for each subject and their average and standard deviation are also demonstrated. As illustrated, the recognition rate of the denoised speech varies across different aGTs. That is, the gain threshold affects the intelligibility of the denoised speech. Although there is a certain degree of variation across individual subjects, the optimal aGT for most subjects are at around 0 or 5 dB. In average, best recognition rate of 74% can be obtained at aGT = 0 dB. In comparison with the undenoising condition, which has a recognition rate of 45%, there is a relative improvement of 29%.

3.2.2. Results with babble noise

Figure 2 shows speech perception rate for the 6 denoising conditions. The target speech signals were corrupted by 11-talker babble noise. Perception rates for each subject and their average and standard deviation are also demonstrated. As illustrated, similar results as those given in Fig.1 can also be observed. In average, best recogniton rate of 73% can be obtained at aGT = 5 dB. Compared with the recognition rate of 52% without denoising, there is a relative improvement of 21%.

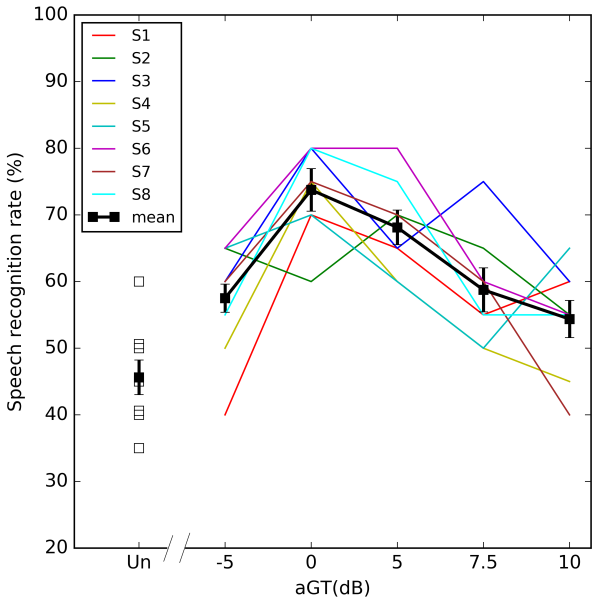

Figure 1: *Vocoded speech perception rate for the 8 subjects with different denoising conditions. The noise signal is SSN. Mean results of all subjects is shown. Error bars indicate the standard deviation cross the subjects.*

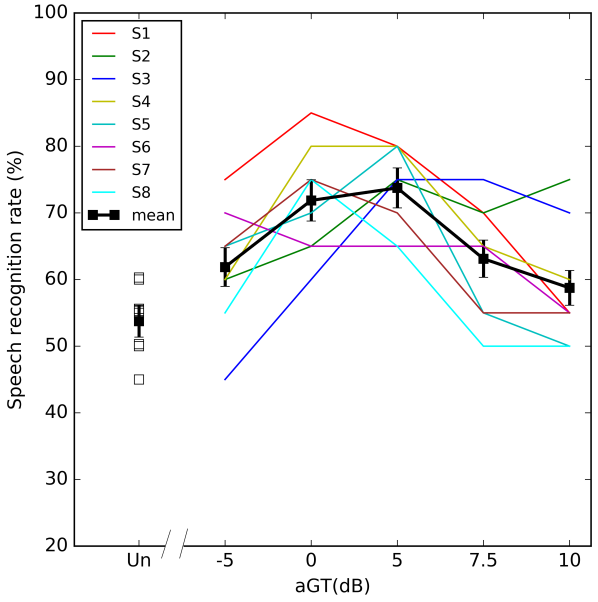

Figure 2: *Vocoded speech perception rate for the 8 subjects with different denoising conditions. The noise signal is babble noise. Mean results of all subjects is shown. Error bars indicate the standard deviation cross the subjects.*

4. Conclusion and Discussion

4.1. Discussion

The parametric gain function for noise suppression in the proposed algorithm is the same as that in [11]. However, there are two major differences between the two algorithms. Firstly, instead of using the ideal noise estimation as in (8) which requires prior knowledge of the noise component,the proposed algorithm adopted a more realistic noise estimator, which requires no prior knowledge of the noise. Secondly, in [11], the optimal α and β parameters for determining the optimal gain function were experimentally determined by varying both α and β parameters, which required intensive time consumption to find the optimal

parameters. In the proposed algorithm, the apparent gain threshold is incorporated and the optimal α (with β set to be 1) can be experimentally determined by varying the gain threshold.

It should be noted that in [11], the experiments were conduct with CI recipients, while in this study only vocoder simulation experiments on normal hearing subjects were carried out. Therefore, it is currently hard to compare the performance of the two algorithms on speech intelligibility for CI users. Experiments on CI subjects will be conducted in the next stage to evaluate the performance of the proposed noise reduction algorithm.

As shown in both Fig.1 and 2, the optimal gain threshold value may vary across individual subjects. Therefore, in clinical implementations, user-dependent gain threshold should be tuned to obtain the best denoising performance.

4.2. Conclusion

In this study, we proposed an improved statistical model based noise reduction algorithm for CI signal processing strategies. The improved minimum controlled recursive algorithm proposed by Cohen, which does not require prior knowledge of the noise, was adopted for realistic noise estimation. An apparent gain threshold was adopted to computed the perceptually optimal gain function for noise suppression. Vocoder simulation perceptual experiments with normal hearing listeners showed that the intelligibility of the denoised speech is highly related to the value of the gain thereshold. In average best performance was achieved with aGT = 0 dB for speech-spectrum shaped noise corruption, and with aGT = 5 dB for babble noise corruption.

5. Acknowledgements

This work is partially supported by Guangdong Natural Science Foundation(2014A030313557).

6. References

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[2] M. Cooke, "A glimpsing model of speech perception in noise." *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–73, 2006.

[3] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: effect of location and type of interferer." *Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

[4] M. K. Qin and A. J. Oxenham, "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 446–54, 2003.

[5] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms." *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1777, 2007.

[6] P. C. Loizou, A. Lobo, and Y. Hu, "Subspace algorithms for noise reduction in cochlear implants." *Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 2791–2793, 2005.

[7] L. P. Yang and Q. J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise." *Journal of the Acoustical Society of America*, vol. 117, no. 3 Pt 1, pp. 1001–4, 2005.

[8] P. W. Dawson, S. J. Mauger, and A. A. Hersbach, "Clinical evaluation of signal-to-noise ratio-based noise reduction in nucleus cochlear implant recipients." *Ear & Hearing*, vol. 32, no. 32, pp. 382–90, 2011.

[9] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *Signal Processing Letters IEEE*, vol. 9, no. 7, pp. 204–206, 2002.

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

[11] S. J. Mauger, P. W. Dawson, and A. A. Hersbach, "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," *Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 327–336, 2012.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *Speech & Audio Processing IEEE Transactions on*, vol. 11, no. 5, pp. 466–475, 2003.

[13] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1980.

[14] S. J. Mauger, K. Arora, and P. W. Dawson, "Cochlear implant optimized noise reduction." *Journal of Neural Engineering*, vol. 9, no. 6, pp. 2216–2221, 2012.

[15] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M. W. Huang, "Development of the mandarin hearing in noise test (MHINT)." *Ear & Hearing*, vol. 28, no. 2 Suppl, pp. 70S–74S, 2007.

[16] D. D. Dirks and J. R. Dubno, "A procedure for quantifying the effects of noise on speech recognition." *Journal of Speech & Hearing Disorders*, vol. 47, no. 47, pp. 114–23, 1982.

[17] N. A. W. Iii, S. F. Poissant, R. L. Freyman, and K. S. Helfer, "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," *Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2376–88, 2007.

[18] D. D. Greenwood, "A cochlear frequency-position function for several species-29 years later," *Acoustical Society of America Journal*, vol. 87, no. 6, pp. 2592–2605, 1990.