

Speech Enhancement Using Temporal Masking in the FFT Domain

Yao Wang, Jiong An, Teddy Surya Gunawan, and Eliathamby Ambikairajah

School of Electrical Engineering and Telecommunications
The University of New South Wales, Australia
{wendy, jiongan, tsgunawan, ambi}@ee.unsw.edu.au

Abstract

Temporal masking models have not been previously applied in the Fast Fourier Transform (FFT) domain for speech enhancement applications. This paper presents a novel speech enhancement algorithm using temporal masking in the FFT domain. The proposed algorithm is suitable for the cochlear speech processor and for other speech applications. The input signal is analysed using FFT and then grouped into 22 critical bands. The noise power is estimated using a minimum statistics noise tracking algorithm. A short-term temporal masking threshold is then calculated for each critical band and a gain factor for each band is then computed. The objective and subjective evaluations show that the temporal masking model based speech enhancement scheme outperforms the traditional Wiener filtering approach in the FFT domain.

1. Introduction

In many speech communications, the presence of background noise causes the quality and intelligibility of speech to degrade, especially when the Signal-to-Noise Ratio (SNR) is low. Speech enhancement algorithms are of great interest because they have many applications such as speech recognition, hearing aids, and mobile communications, etc. The most popular method for enhancing speech in the FFT domain is Wiener filtering, which introduces speech distortion and a perceptually annoying residual noise known as “musical noise”, especially at low SNR.

Currently, a great research effort on denoising algorithms exploiting the human auditory hearing system has made considerable success, which has resulted in good quality speech with improved intelligibility and low level musical noise (Gustafsson, Nordholm & Claesson, 2001; Virag, 1999; Lin, Ambikairajah & Holmes, 2003). All these methods use simultaneous masking properties of the human auditory system, where the simultaneous masking threshold is calculated using the MPEG psychoacoustic model 1 (Black & Zeytinoglu, 1995).

Most recently speech enhancement using temporal masking and using fractional bark gammatone filters has been reported (Gunawan & Ambikairajah, 2004, 2006a) where a robust, flexible and versatile speech boosting

technique has been exploited. The authors reported that the algorithm has good potential for speech enhancement applications across many types and intensities of environmental noise.

Temporal masking is a time domain phenomenon and most of the recently developed temporal masking models operate on the output of a filter bank (Gunawan & Ambikairajah, 2004, 2006a). To the authors' knowledge, the temporal masking effect has not been previously applied in the FFT domain.

This paper presents a speech enhancement algorithm using temporal masking in the FFT domain where Martin's noise tracking algorithm (Martin, 2001) is used for noise power estimation and a perceptually modified Wiener filter (Lin et al., 2003) is used for gain calculation. The temporal masking model developed by (Gunawan & Ambikairajah, 2006a) is optimised to calculate the masking threshold. To evaluate the performance of our algorithm, the objective PESQ measure (ITU-T P.862) and subjective tests that adhere to ITU-T P.835 standard are utilised.

2. FFT Based Filter Bank

Our speech processing strategy in this paper is similar to the cochlear speech processor where the incoming signal is analyzed and separated into many frequency bands. The main spectral components of each critical band are then converted to stimuli of given amplitude

and the phases corresponding to each frequency in the critical band are discarded.

The incoming audio signal was sampled at 16 kHz and framed into 128 samples with 50% overlap and with the lowest and highest frequency bins discarded. The remaining frequency bins are then divided into 22 critical bands corresponding to the Bark scale. The FFT based filter bank, speech enhancement and the cochlear implant simulation are shown in Figure 1.

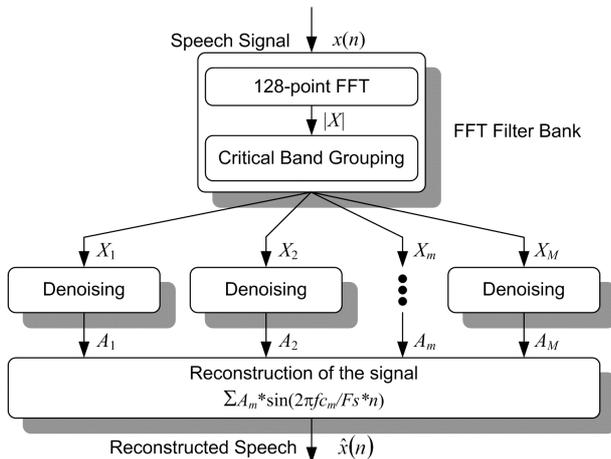


Figure 1: FFT filter bank

3. Temporal Masking in FFT Domain

Temporal masking has been used successively in speech enhancement (Gunawan & Ambikairajah, 2004) using a gammatone filter bank front-end. In this paper, the same method is reviewed while the calculation of temporal masking threshold for the FFT filter bank is derived.

Based on the forward masking experiments carried out by (Jesteadt, Bacon & Lehman, 1982), forward masking level FM can be well-fitted to psychoacoustic data using the following equation:

$$FM = a(b - \log_{10} \Delta t)(Lm - c) \quad (1)$$

where FM is the amount of forward masking in dB, Δt is the time difference between the masker and the maskee in milliseconds, Lm is the masker level in dB, and a , b , and c , are parameters that can be derived from psychoacoustic data. To simplify the masking calculation, a , b , and c were set empirically to 0.1, 2.3, and 20, respectively.

The temporal masking threshold is strongly influenced by the signals (maskers) in the previous frames. The temporal information is obtained by calculating the temporal distances (T_F) between frames. For 50% overlap between frames,

$$T_F = \frac{0.5 \cdot N_F}{F_S} \times 10^3 \quad \text{ms} \quad (2)$$

where N_F is the frame size. Since the longest duration of forward masking is 200 ms, the forward masking threshold is calculated over N_S successive frames as follows:

$$N_S = \lfloor 200/T_F \rfloor \quad (3)$$

The final temporal masking threshold for each critical band TM_m is then chosen as the maximum of FM_m over N_S previous frames:

$$TM_m = \max\{FM_{m,k}\}, \quad k = 1 \dots N_S \quad (4)$$

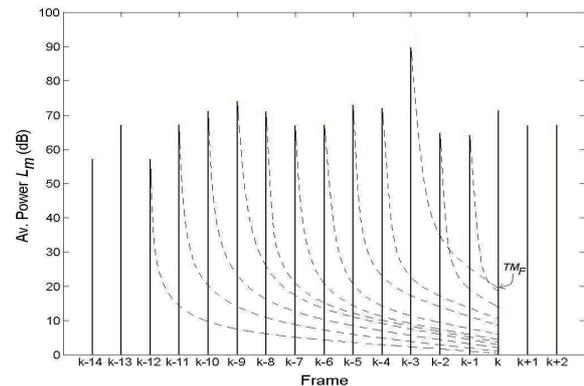


Figure 2: Temporal masking (--- is temporal masking level for previous frames)

4. Speech Enhancement

The objective in speech enhancement is to suppress the noise, thus resulting in an output signal $\hat{x}(n)$ that has a higher SNR. In this section, the basic Wiener filter and the proposed Wiener filter exploiting temporal masking threshold are explained. Furthermore, the noise estimation algorithm in the FFT domain for the Wiener filter approach is outlined.

4.1. Basic Wiener filter

In this method, a noisy signal $x(n)$ is decomposed into critical band signals using FFT filter bank. The objective here is to find a Wiener gain Γ_m (Lin et al., 2003) for each critical band. Subsequently, each noisy critical band signal is multiplied by the denoising gain Γ_m to obtain the denoised critical band amplitude $A_m = \Gamma_m \cdot X_m$ (Fig. 1). The signal can be reconstructed by using the denoised magnitude A_m and noisy phase ϕ_m corresponding to each frequency within the critical bands. In the case of cochlear speech processor, the phase information is discarded and as a result only the magnitude corresponding to centre frequency of each critical band is retained, thus providing centre

frequency with their respective magnitude. The following equation is used to reconstruct one frame of the signal.

$$\hat{x}(n) = \frac{1}{N} \sum_{m=1}^M A_m \cdot \sin\left(2\pi \frac{f_{cm}}{F_s} n + \phi_m\right) \quad (5)$$

where f_{cm} is the centre frequency in each critical band, A_m is the denoised critical band amplitude, and M is the number of critical bands for the particular sampling frequency F_s . In cochlear based speech processing, ϕ_m is the phase of each critical band and is set to zero.

The Wiener filter gain Γ_m can be calculated as follows (Lin et al., 2003):

$$\Gamma_m = \frac{\sigma_{\hat{x}_m}^2 - \alpha \cdot \sigma_{\hat{v}_m}^2}{\sigma_{\hat{x}_m}^2} \quad (6)$$

where $\sigma_{\hat{x}_m}^2$ is the noisy signal power, and $\sigma_{\hat{v}_m}^2$ is the estimated noise power, and α is the oversuppression factor.

In order to reduce the residual musical noise, a smoothing technique has been applied to the Wiener gain Γ_m as follows:

$$\tilde{\Gamma}_m(\bar{n}) = \gamma \cdot \Gamma_m(\bar{n}) + (1 - \gamma) \cdot \Gamma_m(\bar{n} - 1) \quad (7)$$

where $\tilde{\Gamma}_m$ is the smoothed Wiener gain for the current frame, $\Gamma_m(\bar{n})$ is the Wiener gain in the current frame, $\Gamma_m(\bar{n} - 1)$ is the Wiener gain in the previous frame, and γ is the smoothing factor.

4.2. Wiener filter incorporating temporal masking

Temporal masking has been successively used for speech enhancement (Gunawan & Ambikairajah, 2004, 2006a). In this paper, we modified Equation (6) to incorporate the temporal masking threshold as follows:

$$\Gamma_m = \frac{\sigma_{\hat{x}_m}^2 - \alpha \cdot \max(\sigma_{\hat{v}_m}^2 - \beta \cdot TM_m, 0)}{\sigma_{\hat{x}_m}^2} \quad (8)$$

where $\sigma_{\hat{x}_m}^2$ is the noisy signal power, $\sigma_{\hat{v}_m}^2$ is the estimated noise power, $\alpha = 2.4$ is the oversuppression factor, and $\beta = 0.85$ is the parameter that controls temporal masking threshold. Note that, α and β are the parameters that can be optimised empirically. The noise is included in Equation (8) only if it exceeds the masking threshold. Furthermore, the noise is weighted only by the amount that exceeds the masking threshold.

4.3. Noise estimation algorithm

An accurate estimate of noise power for each critical

band is obtained based on the optimal smoothing and minimum statistics strategy by Martin (Martin, 2001). Martin's algorithm is based on tracking the minimum of the noisy signal power spectral density. The minimum noise statistics noise tracking method is based on the observation that even during speech activity a short-term power spectral density estimate the noisy signal on a frequent basis and decays to values that are representative of the noise power level. The method rests on the fundamental assumption that during speech pauses, or within brief periods in between words and syllables, the speech energy is close or identical to zero. Thus by tracking the minimum power within a finite window large enough to bridge high power speech segments, the noise floor can be estimated.

The noise estimate is obtained by selecting the minimum value within a sliding window of 60 consecutive frames, regardless of whether speech is present or not. Since the minimum value of a set of random variable is smaller than their mean the minimum noise estimate is usually biased. The noise estimation technique of Martin's is by far the most accurate implicit noise estimation algorithm (Lin et al., 2003).

5. Performance Evaluation

In order to assess the performance of our proposed algorithm, objective tests using PESQ (Rix, Hollier, Hekstra & Beerends, 2002) and subjective tests conforming to ITU-T P.835 were performed. One speech signal sampled at 16 kHz is added with three noises at three SNR levels, i.e. 0 dB, 5 dB, and 10 dB. The frame size was 128 samples with 50% overlap.

A sentence spoken by an English female speaker is corrupted in three background noise environments (car, white, babble, street, pink and factory noises) from NOISEX-92 database. Two algorithms were compared, i.e. Wiener basic (**Wiener**) and Wiener with temporal masking (**WienerTM**).

5.1. Objective evaluation

Table 1: Objective evaluation using PESQ

SNR 0 dB	Wiener	WienerTM
Car Noise	1.412	1.537
White Noise	1.581	1.999
Babble Noise	1.321	1.42
Street Noise	1.517	1.63
Pink Noise	1.536	1.633
Factory1 Noise	1.487	1.727
SNR 5 dB	Wiener	WienerTM
Car Noise	1.843	1.871
White Noise	1.951	2.321
Babble Noise	1.704	1.814

Street Noise	1.945	2.071
Pink Noise	1.943	2.197
Factory1 Noise	1.881	2.066
SNR 10 dB	Wiener	WienerTM
Car Noise	2.194	2.268
White Noise	2.246	2.623
Babble Noise	2.106	2.215
Street Noise	2.282	2.347
Pink Noise	2.260	2.493
Factory1 Noise	2.219	2.399

The PESQ measure (ITU-T P862) was utilised for the objective evaluation. A total of 18 files from six noises and three SNRs for each method were simulated. As shown in Table 1, the Wiener filtering incorporating temporal masking outperforms the basic Wiener filtering method in all these noises. In addition, we found that this temporal masking based Wiener filter works better in white noise environment.

5.2. Subjective evaluation

The subjective evaluation conforming to ITU-T P.835 standard is performed. The standard uses separate rating scales to independently estimate the subjective quality of the speech signal alone, the background noise alone, and overall quality. The listener's uncertainty is reduced and the reliability is increased by using the above three dimensions of subjective speech quality. Moreover, the previously developed toolbox, i.e. P835tool, is employed (Gunawan & Ambikairajah, 2006b).

A subset of the files described in the objective test was selected to reduce the length of the subjective evaluations. Only the enhanced speech corrupted by car noise at 10 dB SNR was presented to the listeners. Furthermore, a high quality headphone, i.e. Sony MSRV700DJ, was utilized for the listening tests. A total of twenty subjects took part in the subjective listening test.

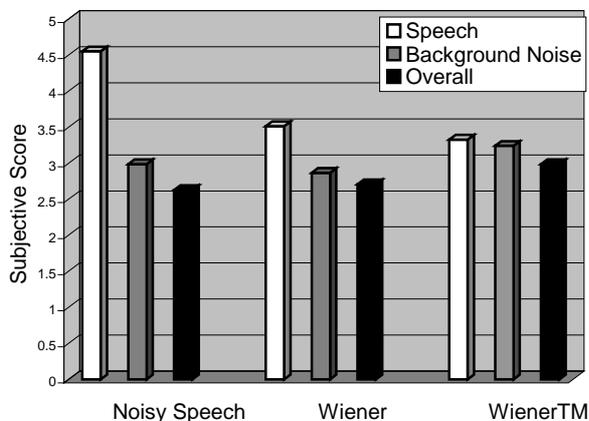


Figure 3: Subjective test results

Fig. 3 shows the score (MOS) for signal, background noise, and overall scale for the two methods of speech enhancement. The score for the noisy speech (unprocessed) files are also shown for references. Of the two methods examined, the Wiener filtering incorporating temporal masking performs better than the traditional Wiener filter in terms of background noise level and overall quality, while the speech quality is slightly lower.

5.3. Speech Spectrogram

Objective measures do not give indications about the structure of the residual noise. Speech spectrograms constitute a well-suited tool for observing this structure. The speech spectrogram for SNR of 10 dB is obtained by using a Hanning window of 128 samples with 50 % overlap. Fig. 4 shows the speech signals and its corresponding spectrograms.

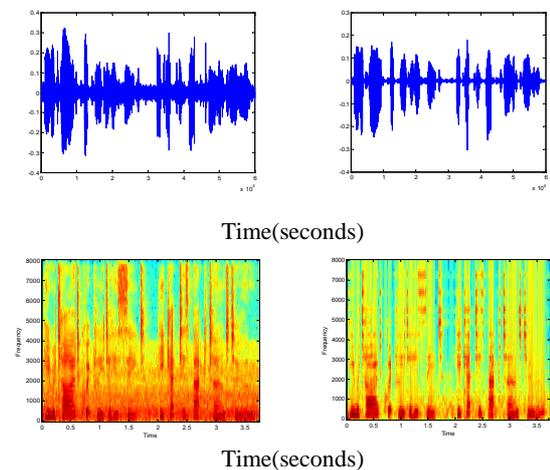


Figure 4: Speech waveform and its spectrogram

The original speech signal in English: “He retired quickly to his seat, he kept his back turn to them”, and its spectrogram is on the half left of the figure. The half right is the processed speech and its spectrogram. We can see that our proposed algorithm enhanced the noisy speech.

6. Conclusions

In this paper, we have incorporated a temporal masking model for speech enhancement in the FFT domain. A novel way of calculating the temporal masking threshold in the FFT domain was developed. The performance of the proposed speech enhancement algorithm was compared with the traditional Wiener filtering based speech enhancement technique. Subjective and objective results show that the proposed algorithm performs well under various noisy conditions. This algorithm can be

incorporated easily in a behind-the-ear cochlear speech processor as an alternative to the existing noise reduction algorithms.

7. References

- Black, M. & Zeytinoglu, M. (1995). Computationally efficient wavelet packet coding of wide-band stereo audio signals. *In proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 3075-3078, Detroit.
- Gunawan, T. S. & Ambikairajah, E. (2004). Speech Enhancement Using Temporal Masking and fractional bark gammatone filters. *In proceedings of the 10th International Conference on Speech Science & Technology*, pp.420-425, Sydney.
- Gunawan, T. S. & Ambikairajah, E. (2006a). A new forward masking model for speech enhancement. *In proceedings of IEEE International Conference on Acoustics, Speech, and Audio Signal Processing*, Vol. 1, pp. 149-152, Toulouse.
- Gunawan, T. S. & Ambikairajah, E. (2006b). Subjective evaluation of speech enhancement algorithms using ITU-T P.835 standard. *In proceedings of the 10th IEEE International Conference on Communication Systems (ICCS'06)*, Singapore.
- Gustafsson, H., Nordholm, S. E. & Claesson, I. (2001). Spectral Substraction Using Reduced Delay Convolution and Adaptive Averaging, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, 799-807.
- Jesteadt, W., Bacon, S. P. & Lehman, J. R. (1982). Forward masking as a function of frequency, masker level, and signal delay. *Journal of Acoustic Society of America*, Vol. 71, 950-962.
- Lin, L., Ambikairajah, E. & Holmes, W. H. (2003). Subband noise estimation for speech enhancement using a perceptual wiener filter, *In proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Vol. 1, pp. 80-83, Hong Kong.
- Martin, R. (2001). Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, 504-512.
- Rix, A. W., Hollier, M. P., Hekstra, A. P. and Beerends, J. G. (2002). Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I - time-delay compensation. *Journal of the Audio Engineering Society*, 50 (10), 755-764.
- Virag, N. (1999). Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, 126 – 137.