Single Channel Speech Enhancement by Frequency Domain Constrained Optimization and Temporal Masking

Wen Jin, Michael Scordilis

Department of Electrical and Computer Engineering
University of Miami, Coral Gables, Florida 33146, U.S.A.
wjin@umsis.miami.edu m.scordilis@miami.edu

Abstract

A speech enhancement algorithm is proposed that exploits the masking properties of the human auditory system. The enhancement is formulated as a frequency domain constrained optimization problem. The noise components of the noisy speech are suppressed by a gain function subject to the constraint that both the signal distortion and residual noise should fall below the masking thresholds. Temporal as well as simultaneous masking effects are incorporated into the estimation of masking thresholds. The enhancement algorithm was tested with speech corrupted by white Gaussian and multitalker babble noise, respectively. Its performance was evaluated by ITU PESQ scores and segmental SNR. Experimental results indicate that the proposed gain function performs slightly but consistently better than a former perceptually motivated enhancement algorithm. Greater improvement is achieved by incorporating the temporal masking effects.

Index Terms: speech enhancement, psychoacoustical model, temporal masking.

1. Introduction

The goal of speech enhancement is to reduce listener's fatigue or to improve the speech signal prior to its presentation to automatic speech recognition systems. For single channel speech degraded by additive noise, it is advantageous to incorporate auditory models into enhancement methods so that the residual distortions are inaudible or less objectionable. Psychoacoustical models were initially proposed for high quality audio coding [1]. Their applications in speech enhancement can be found in [2–6]. In [2], the masking thresholds are derived and used as contraints in solving a frequency domain constrained optimization problem. [3] introduces an auditory model in a spectral subtractive-type of enhancement algorithm. [4] deduces a Frequency-to-Eigendomain Transformation (FET) that maps the masking thresholds to the eigen domain, enhancement is then performed by applying an eigenfilter. In [5], the Fourier domain masking thresholds are transformed to the Discrete Cosine Transform (DCT) domain by bark filtering and enhancement is achieved by a subspace approach.

The aforementioned enhancement algorithms [2–5] estimate the masking thresholds by exploiting simultaneous masking effects. Since the human auditory system has temporal as well as simultaneous masking properties [8], it is worthwhile investigating both of these masking effects on speech enhancements. In a previous attempt [6], Kalman filtering was used to enhance the noisy speech, and a post-filter that incorporates the temporal masking was then applied to the pre-enhanced speech.

In this paper, we introduce a frequency domain method that incorporates temporal as well as simultaneous masking effects into the enhancement process. The proposed scheme minimizes the speech distortion subject to the constraint that both the signal distortion and residual noise are inaudible. The magnitude spectra of the noisy speech are modified by a gain function such that the speech distortion and residual noise are suppressed below the masking thresholds. In the calculation of masking thresholds, temporal masking aspects are explicitly included in the auditory model by estimating the decay of the internal loudness in the human hearing system [8].

2. The proposed method

2.1. Deduction of the Optimal Gain

Consider a clean speech signal $\mathbf{x}$ corrupted by uncorrelated additive noise $\mathbf{n}$. The noisy observation $\mathbf{y}$ can be modeled as

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ are $N \times 1$ vectors. By applying a N-point Short Time Fourier Transform (STFT) to the noisy speech, we have

$$\mathbf{Y} = F^H\mathbf{y} = F^H\mathbf{x} + F^H\mathbf{n} = \mathbf{X} + \mathbf{N} \tag{2}$$

where F^H denotes the $N \times N$ Discrete Fourier Transform matrix, $(\cdot)^H$ is matrix Hermitian. $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{N}$ are the Fourier transforms of noisy speech, clean speech and noise respectively.

Let $\hat{\mathbf{X}} = G\mathbf{Y}$ be a linear estimator of the clean speech $\mathbf{X}$, where G is an $N \times N$ matrix. The estimation error is

$$\mathbf{r} = \hat{\mathbf{X}} - \mathbf{X} = \varepsilon_{\mathbf{x}} + \varepsilon_{\mathbf{n}} \tag{3}$$

where $\varepsilon_{\mathbf{x}} \triangleq (G - I)\mathbf{X}$ denotes the spectrum of speech distortion and $\varepsilon_{\mathbf{n}} \triangleq G\mathbf{N}$ is the spectrum of residual noise. Let $\bar{\varepsilon}^2_{\mathbf{x}} = E\{{\varepsilon_{\mathbf{x}}}^H\varepsilon_{\mathbf{x}}\}$ be the energy of the frequency domain speech distortion. The energy of the kth spectral component of speech distortion and residual noise is defined as $\varepsilon^2_{\mathbf{x},k} = E\{|\varepsilon_{\mathbf{x},k}|^2\}$ and $\varepsilon^2_{\mathbf{n},k} = E\{|\varepsilon_{\mathbf{n},k}|^2\}$ respectively. Finding the optimum linear estimator G can be formulated as solving the following constrained optimization problem:

$$\min_G \bar{\varepsilon}^2_{\mathbf{x}} \tag{4a}$$

$$\text{subject to: } \varepsilon^2_{\mathbf{x},k} + \varepsilon^2_{\mathbf{n},k} \le \alpha_k, \qquad k = 1, \ldots, N \tag{4b}$$

where α_k are some preset thresholds ($\alpha_k \ge 0$). For perceptually motivated speech enhancement, these thresholds α_k can be set equal to the masking thresholds T_k at frequency ω_k.

It should be noted that a similar problem formulation is proposed in [2]. Our method differs because both the signal distortion and residual noise are included in (4b), while the constraints used in [2] are only

$$\text{subject to: } \varepsilon_{\mathbf{n},k}^2 \leq \alpha_k, \qquad k = 1, \ldots, N. \tag{5}$$

We propose the use of constraint (4b) instead of (5) because it is perceptually more meaningful. In the best scenario, if both the speech distortion and residual noise are masked, there will be no audible distortions in the enhanced speech. However, as stated in [7], in most real cases a complete masking of both components cannot be guaranteed because of the fact that the minimum of the estimation error is greater than zero for non-trivial signals. The proposed approach masks both components of the estimation error whenever it is possible. This aggressive suppression of the estimation error will ensure the best possible quality of the enhanced speech.

The optimization problem (4) can be solved by using the method of Lagrangian multipliers [9]. Specifically, G is a stationary feasible point if it satisfies the gradient equation of the Lagrangian

$$J(G, \mu_k) = \bar{\varepsilon_{\mathbf{x}}^2} + \sum_{k=1}^{N} \mu_k (\varepsilon_{\mathbf{x},k}^2 + \varepsilon_{\mathbf{n},k}^2 - \alpha_k) \tag{6}$$

and

$$\mu_k (\varepsilon_{\mathbf{x},k}^2 + \varepsilon_{\mathbf{n},k}^2 - \alpha_k) = 0, \qquad \text{for} \qquad k = 1, \ldots N \tag{7}$$

where $\mu_k \geq 0$ is the kth Lagrangian multipler for the kth component of $\varepsilon_{\mathbf{x}}$ and $\varepsilon_{\mathbf{n}}$. From $\nabla_G J(G, \mu_k) = 0$ we obtain

$$(I + \Lambda_\mu) G F^H R_x F + \Lambda_\mu G F^H R_n F = (I + \Lambda_\mu) F^H R_x F \tag{8}$$

where $\Lambda_\mu = diag(\mu_1, \ldots, \mu_N)$ is a diagonal matrix. The matrices $F^H R_x F$ and $F^H R_n F$ are asymptotically diagonal provided the matrices R_x and R_n are Toeplitz [2]. The diagonal elements of matrices $F^H R_x F$ and $F^H R_n F$ are the power spectrum components $S_{\mathbf{x}}(\omega_k)$ and $S_{\mathbf{n}}(\omega_k)$ of the clean speech and noise respectively [2]. The optimum linear estimator can be obtained by solving the matrix equation in (8). One possible solution is obtained when G is a diagonal matrix with elements

$$g_{kk} = \frac{S_{\mathbf{x}}(\omega_k)}{S_{\mathbf{x}}(\omega_k) + S_{\mathbf{n}}(\omega_k) \frac{\mu_k}{(1+\mu_k)}} = \frac{\gamma(k)}{\gamma(k) + \beta_k} \tag{9}$$

where $\gamma(k) = S_{\mathbf{x}}(\omega_k)/S_{\mathbf{n}}(\omega_k)$ is known as the *a priori* SNR at frequency ω_k, and $\beta_k = \mu_k/(1 + \mu_k)$. For this G, the kth spectral component of speech distortion and residual noise is

$$(g_{kk} - 1)^2 S_{\mathbf{x}}(\omega_k) + {g_{kk}}^2 S_{\mathbf{n}}(\omega_k). \tag{10}$$

Assuming the constraints (4b) are satisfied with equality and the thresholds $\alpha_k = T_k$, then

$$(g_{kk} - 1)^2 \gamma(k) + {g_{kk}}^2 = Q_k \tag{11}$$

where $Q_k = T_k/S_{\mathbf{n}}(\omega_k)$. (11) is a quadratic equation in g_{kk} and has two roots. By imposing the condition $\mu_k \geq 0$, we retain only one root as the optimum gain

$$g_{kk}^b = \frac{\gamma(k) + \sqrt{\gamma(k)(Q_k - 1) + Q_k}}{\gamma(k) + 1}, \qquad Q_k \leq 1. \tag{12}$$

For this gain function (12), we have

$$\mu_k^b = \frac{\gamma(k) - \gamma(k)\sqrt{\gamma(k)(Q_k - 1) + Q_k}}{(\gamma(k) + 1)\sqrt{\gamma(k)(Q_k - 1) + Q_k}}. \tag{13}$$

Let $p_k = \gamma(k)(Q_k - 1) + Q_k$. It can be readily verified that $\mu_k^b \geq 0$ when $Q_k \leq 1$ and $p_k \geq 0$. The gain g_{kk}^b in (12) is complex when $p_k < 0$. In this case, the phase of the noisy speech spectral components will be modified. However, it has been a common practice in speech enhancement not to alter the phase of the noisy speech. Therefore, we use the gain $g_{kk} = g_{kk}^n$ when $p_k < 0$, where

$$g_{kk}^n = \sqrt{Q_k}, \qquad Q_k \leq 1 \tag{14}$$

is the gain obtained by solving the constrained optimization problem (4a), (5) and is used in [2]. If $Q_k > 1$, this means $S_{\mathbf{n}}(\omega_k) < T_k$. In other words, the residual noise is masked at frequency ω_k. We then use $g_{kk} = 1$. This choice of g_{kk} will result in zero signal distortion. In summary, the proposed gain function is

$$g_{kk}^p = \begin{cases} 1, & Q_k > 1 \\ g_{kk}^b, & Q_k \leq 1, p_k \geq 0 \\ g_{kk}^n, & Q_k \leq 1, p_k < 0. \end{cases} \tag{15}$$

It can be verified from (12), (14) and (15) that the proposed gain is bounded by $0 < g_{kk}^p \leq 1$. For a gain function that falls within this range, the minimum energy of the speech distortion and residual noise as given by (10) is achieved with the Wiener gain

$$g_{kk}^w = \frac{\gamma(k)}{\gamma(k) + 1}. \tag{16}$$

Comparing (9) with (16) and noticing that $0 \leq \beta_k < 1$, it can be verified that the proposed gain is larger than the Wiener gain g_{kk}^w. From (10), we can see that the larger the g_{kk} value, the smaller the signal distortion. On the other hand, increasing g_{kk} will increase the level of the residual noise. The motivation to use the gain as in (15) is to introduce minimum signal distortion while aggressively suppressing the speech distortion and residual noise below the masking thresholds.

2.2. Masking Thresholds

The simultaneous masking thresholds can be estimated by the MPEG4 psychoacoustical model [1]. The temporal masking phenomenon is known to occur when one sound (maskee) is masked some time before and after the presentation of another stronger signal (masker) [8] (*pre-masking* and *post-masking*). Post-masking plays a dominant role in non-simultaneous masking [8]. The amount of post-masking depends on the duration, energy level and frequency content of the masker [12]. The post-masking level decays faster for maskers with shorter duration and higher energy level, while longer post-masking is observed after signals with relatively long duration and low energy level. This effect can be modelled more easily in terms of the decay of the specific loudness against critical band index and time. The specific loudness can be estimated by (Chapter 8 of [8])

$$N = 0.08 (\frac{E_{TQ}}{E_0})^{0.23} [(0.5 + \frac{E}{2E_{TQ}})^{0.23} - 1] \frac{sone_G}{Bark} \tag{17}$$

where E_{TQ} is the excitation at the threshold in quiet and E_0 is the excitation of the reference sound with intensity $10^{-12} W/m^2$. E is

the excitation of the masker signal and N is the specific loudness it produces.

The duration-dependent decay of postmasking can be simulated by filtering the specific loudness in (17) with the RC circuit proposed in [12]. The total loudness is then obtained by

$$N^* = \sum_f W_f N_f \tag{18}$$

where N^* is the total loudness of the current frame, f is the subband index, and W_f is the bandwidth in bark of each subband as defined in the MPEG-4 standard [1].

The final masking thresholds for each subband are then determined by [6]

$$T(t,f) = max(T_s(t,f), T(t-1,f)e^{-\Delta t/\tau(f)N^*}) \tag{19}$$

where $T(t,f)$ is the final masking threshold of the current frame, and t is the frame index. $T_s(t,f)$ is the simultaneous masking thresholds of the current frame. $T(t-1,f)$ is the final masking threshold of the previous frame. Δt is the time shift between adjacent frames. $\tau(f)$ is the maximum decay time constant for each subband. The total loudness N^* is normalized by the total loudness of a 40 dB SPL uniform masking noise (UMN). If N^* is larger than one, it is set to one so that the value of N^* lies in the range of zero to one.

In summary, we take the following steps to incorporate the temporal masking effects into the psychoacoustical model. First, the simultaneous masking thresholds are computed via the MPEG-4 model. Second, the specific loudness is estimated by (17). Third, the specific loudness is processed by the RC circuit in [12]. Then, the total loudness is obtained by (18). Finally, the ultimate masking thresholds of current frame is determined by (19).

3. Implementation

3.1. Spectrum Estimation

As stated in [2], the accuracy in the estimation of clean speech spectrum and noise spectrum is crucial to the performance of the speech enhancement algorithm. In our implementation, the noisy speech spectrum is estimated by the multitaper wavelet-denoising method proposed in [2]. For noise spectrum estimation, the minimum-statistics tracking method proposed in [10] is used. The clean speech spectrum is then obtained by

$$\hat{S}_{\mathbf{x}}(\omega) = \begin{cases} S_{\mathbf{y}}^{wt}(\omega) - \hat{S}_{\mathbf{n}}(\omega), & S_{\mathbf{y}}^{wt}(\omega) > \hat{S}_{\mathbf{n}}(\omega) \\ \delta \hat{S}_{\mathbf{n}}(\omega), & S_{\mathbf{y}}^{wt}(\omega) \leq \hat{S}_{\mathbf{n}}(\omega) \end{cases} \tag{20}$$

where $S_{\mathbf{y}}^{wt}(\omega)$ is the spectra of noisy speech estimated by the multitaper wavelet-denoising method. $\hat{S}_{\mathbf{n}}(\omega)$ is the estimated noise power spectra and $\delta = 0.025$ is a zero-flooring parameter.

3.2. Masking Thresholds Calculation

The simultaneous masking thresholds are calculated from the estimated clean speech spectrum $\hat{S}_{\mathbf{x}}(\omega)$. The MPEG-4 psychoacoustical model at the sampling rate 8 kHz is used. The values for the RC circuit components are $R1 = 35k\Omega$, $R2 = 20k\Omega$, $C1 = 0.7\mu F$ and $C2 = 1.74\mu F$ [12].

4. Experimental Results

For comparison purposes, the enhancement method in [2] was also implemented. 60 sentences taken from the TIMIT database were downsampled to 8 kHz and used in the tests. The noise sources were downloaded from the IEEE Signal Processing Information Base [11]. Two types of noise were used, namely white Gaussian noise and multitalker babble noise. The noise was scaled in energy level and added to the downsampled clean speech to generate noisy speech with SNR in the range of -5 to 10 dB.

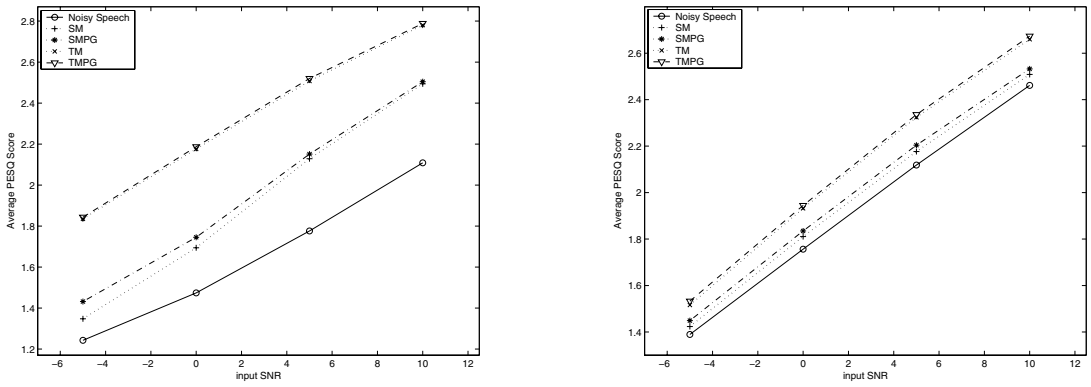
The enhancement was applied to 32 ms of noisy speech with a 50% overlap between adjacent frames. The enhanced speech was obtained by the overlap-and-add method. For a fair comparison, the MPEG-4 psychoacoustical model was used by both methods. The enhancement algorithms were evaluated by ITU-PESQ (Perceptual Evaluation of Speech Quality) scores and segmental SNR.

In order to assess the individual contributions of the proposed gain function (15) and the temporal post-masking (19), four implementations were tested, namely, the frequency domain simultaneous masking method (SM) [2], the frequency domain simultaneous masking with the proposed gain function (SMPG) (15), the proposed post-masking method (TM) (19) and the proposed gain plus post-masking method (TMPG).

Fig. 1 shows the average PESQ scores of 60 unenhanced noisy speeches degraded by white Gaussian/babble noise, and the PESQ scores of their enhancement outputs. Fig. 2 depicts the average segmental SNR of the same 60 enhanced speechs. Comparing Fig. 1 and 2, it can be seen that the improvement is more prominent in PESQ scores than in segmental SNR. This is because the proposed method is motivated by improving the perceptual quality rather than exactly reconstructing the speech waveforms, and the ITU-PESQ score is a more precise measurement of subjective quality than the segmental SNR. From Fig. 1(a) and 2(a), we can see that although the segmental SNRs converge at the input SNR of 10 dB for white Gaussian noise, there still exists notable differences in PESQ scores. Informal listening tests indicate that the speeches enhanced by the TMPG method sound more crispy than the SM method. From both Fig. 1 and 2, it can be verified that the methods with the proposed gain (PG) function perform slightly but consistently better than the methods without the PG. In other words, the SMPG is slightly but constistenly better than the SM method, so do the TMPG and TM methods. However, methods incorporating the post-masking effects (the TM methods) provides significant improvements at all input SNR for both white and babble noise.

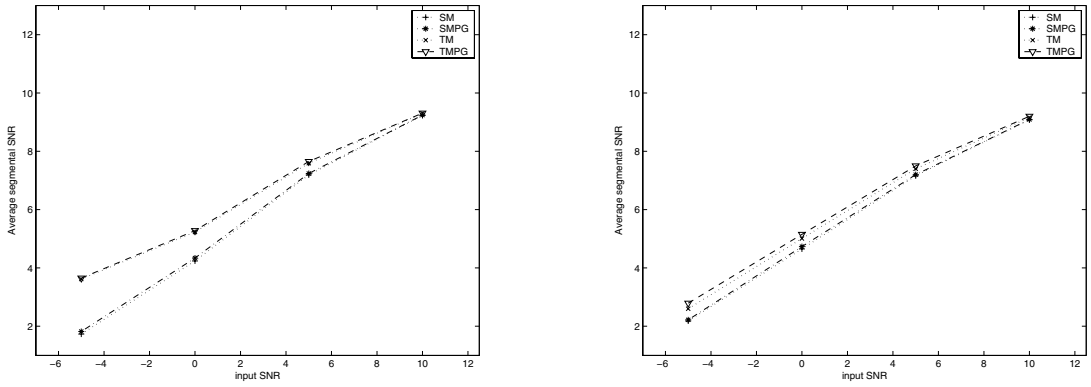
5. Conclusion

A frequency domain speech enhancement algorithm is proposed. The magnitude spectra of the noisy speech are weighted by a gain function that tries to keep both the signal distortion and residual noise below the masking thresholds. In the calculation of the masking thresholds, the proposed method incorporates not only the simultaneous masking, but also the temporal post-masking aspects of the human auditory system. The enhancement algorithms were tested with speech corrupted by white Gaussian and multitalker babble noise. Experimental results indicate that the proposed gain function achieves slight but consistent performance improvements over the method that suppresses the residual noise only. The proposed algorithm is shown to be more perceptually meaningful after incorporating post-masking effects. Moreover, enhancement in speech quality is more pronounced in terms of ITU PESQ scores. Further improvements could be obtained by implementing a more



(a) PESQ White Noise (b) PESQ Babble Noise

Figure 1: Average PESQ Scores of 60 sentences



(a) segSNR White Noise (b) segSNR Babble Noise

Figure 2: Average Segmental SNR of 60 sentences

robust noise power spectrum estimator for highly dynamic noise.

6. References

[1] ISO/IEC 14496-3, "Information technology - Coding of audio-visual objects - Part 3: Audio", International Standards, First Edition, December, 1999.

[2] Hu, Y. and Loizou, P. C., "Incorporating a psychoacoustical model in frequency domain speech enhancement", IEEE Signal Processing Lett., Vol.11:270–273, February 2004.

[3] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech Audio Processing, Vol.7:126–137, March 1999.

[4] Jabloun, F. and Champagne, B., "Incorporating the human hearing properties in the signal subspace approach for speech enhancement", IEEE Trans. Speech Audio Processing, Vol.11:700–708, November 2003.

[5] Vetter, R., "Single channel speech enhancement using MDL-based subspace approach in bark domain", ICASSP'01, Vol.1:641-644, May 2001.

[6] Ma, N., Bouchard, M. and Goubran, R. A. "Perceptual Kalman Filtering for speech enhancement in colored noise", ICASSP'04, Vol.1:717–720, May 2004.

[7] S. Gustafsson, P. Jax and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics", ICASSP'1998, pp. 397-400, Nov. 1998.

[8] Zwicker, E. and Fastl, H., Psychoacoustics-Facts and Models, Springer-Verlag, Berlin Germany, 1990.

[9] Ephraim, Y. and Van Trees, H. L., "A signal subspace approach for speech enhancement", IEEE Trans. Speech Audio Processing, Vol.3:251–266, July 1995.

[10] Martin, R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics", IEEE Trans. Speech Audio Processing, vol.9:504–512, July 2001.

[11] WebSite, "http://spib.rice.edu/spib/select_noise.html", IEEE Signal Processing Information Base, 1997.

[12] Zwicker, E., "Dependence of post-masking on masker duration and its relation to temporal effects in loudness", Journal of Acoustical Society of America, Vol.75:219–223, Jan 1984.