

# Speech Enhancement using a Wiener denoising technique and musical noise reduction

Md. Jahangir Alam<sup>1</sup>, Sid-Ahmed Selouani<sup>2</sup>, Douglas O'Shaughnessy<sup>1</sup>, Sofia Ben Jebara<sup>3</sup>

<sup>1</sup>INRS-Energie-Matériaux-Télécommunications, Université du Québec, Montréal, Canada

<sup>2</sup>Université de Moncton, Campus de Shippagan, NB, Canada

<sup>3</sup>Ecole Supérieure des Communications de Tunis, Tunisia

alam@emt.inrs.ca, dougo@emt.inrs.ca, selouani@umcs.ca, sofia.benjebara@supcom.rnu.tn

## Abstract

Speech enhancement methods using spectral subtraction have the drawback of generating an annoying residual noise with musical character. In this paper a frequency domain optimal linear estimator with perceptual post filtering is proposed which incorporates the masking properties of the human auditory system to make the residual noise distortion inaudible. The performance of the proposed enhancement algorithm is evaluated by the Segmental SNR, Log Spectral Distance (LSD) and Perceptual Evaluation of Speech Quality (PESQ) measures under various noisy environments and yields better results compared to the Wiener denoising technique.

**Index Terms:** speech enhancement, perceptual post filter, musical critical band, MMSE, noise masking threshold

## 1. Introduction

The removal of additive noise from speech has been an active area of research for several decades. Numerous methods have been proposed by the signal processing community. Among the most successful signal enhancement techniques have been spectral subtraction [1] [2] and Wiener filtering [3]. Although these techniques improve speech quality, they generally result in random narrowband fluctuations in the residual noise called musical noise caused by randomly spaced spectral peaks that come and go in each frame and at random frequencies, which is annoying and disturbing to perception of the enhanced signal. The quality and the intelligibility of the enhanced speech signal could be improved by reducing or in better cases eliminating this kind of musical residual noise.

Many variations have been developed to cope with the musical residual noise phenomena including spectral subtraction techniques based on masking properties of the human auditory system. A number of methods have been developed to improve intelligibility by modeling several aspects of the enhancement function present in the human hearing system [4, 5, 10]. These attractive methods use a noise masking threshold (NMT) as a crucial parameter to empirically adjust either thresholds or gain factors. This human hearing system is based on the fact that the human ear cannot perceive additive noise when the noise level falls below the NMT. Since the human ear builds critical bands around each frequency and behaves like a band-pass filter [6], in this paper, we have developed a post processing method with a modified Johnston model to reduce the musical residual noise in each critical band generated by classical speech enhancement methods. The tonality coefficient in each critical band is utilized to characterize the residual musical noise.

This paper is organized as follows: section 2 provides a description of the baseline speech enhancement system. In section 3, descriptions of the *a priori* SNR estimation, noise

estimation and proposed method are given. A discussion of the experimental results and a conclusion are drawn in section 4 and section 5, respectively.

## 2. Baseline speech enhancement method

Let the distorted signal be expressed as

$$y(n) = x(n) + d(n), \quad (1)$$

where  $x(n)$  is the clean signal and  $d(n)$  is the additive random noise signal, uncorrelated with the original signal. If at the  $m$ th frame and  $k$ th frequency bin,  $Y(m, k)$ ,  $X(m, k)$  and  $D(m, k)$  represent the spectral component of  $y(n)$ ,  $x(n)$  and  $d(n)$ , respectively, then the distorted signal in the transformed domain is

$$Y(m, k) = X(m, k) + D(m, k). \quad (2)$$

An estimate  $\hat{X}(m, k)$  of  $X(m, k)$  is given by

$$\hat{X}(m, k) = H(m, k)Y(m, k), \quad (3)$$

where  $H(m, k)$  is the noise suppression gain (denoising filter), which is a function of *a priori* SNR and/or *a posteriori* SNR, given by

$$H(m, k) = \left( \frac{\xi(m, k)}{1 + \xi(m, k)} \right). \quad (4)$$

The first parameter of the noise suppression rule is the *a posteriori* SNR given by

$$\gamma(m, k) = \frac{|Y(m, k)|^2}{\Gamma_d(m, k)}, \quad (5)$$

where  $\Gamma_d(m, k) = E\{|D(m, k)|^2\}$  is the noise power spectrum estimated during speech pauses. The *a priori* SNR, which is the second parameter of the noise suppression rule, is expressed as

$$\xi(m, k) = \frac{\Gamma_x(m, k)}{\Gamma_d(m, k)}, \quad (6)$$

where  $\Gamma_x(m, k) = E\{|X(m, k)|^2\}$ . The estimation of  $\xi(m, k)$  is given by the well known decision-directed approach [7] and is expressed as

$$\hat{\xi} = \hat{\xi}_{DD}(m, k) = \max \left( \alpha \frac{|H(m-1, k)Y(m-1, k)|^2}{\Gamma_d(m, k)}, (1 - \alpha)P'[\vartheta(m, k)], \xi_{\min} \right), \quad (7)$$

where  $P'[x] = x$  if  $x \geq 0$  and  $P'[x] = 0$  otherwise. In this paper we have chosen  $\alpha = 0.98$  and  $\xi_{\min} = .0032$  (i.e., -25 dB) by the simulations and informal listening tests. The instantaneous SNR can be defined as

$$\vartheta(m,k) = \frac{|Y(m,k)|^2}{\Gamma_d(m,k)} - 1. \quad (8)$$

The temporal-domain denoised speech is obtained with the following relation

$$\hat{x}(n) = \text{IFFT}\left(\left|\hat{X}(m,k)\right|e^{j\arg(Y(m,k))}\right). \quad (9)$$

### 3. Overview of the proposed perceptual Wiener denoising technique

The proposed technique to reduce musical noise is developed as a post-processing approach. It aims to detect not only musical peaks but also larger intervals of 1 bark where musical noise is dominant. In fact, it is important to recall that the human ear operates as filter bank, which subdivides the frequency axis into the so-called critical bands [6], frequency bands where two sounds are perceptually heard as one sound with energy equal to the sum of the two sounds' energies. We have exploited this property to detect residual musical noise in a whole critical band instead of separate tones.

#### 3.1 Estimation of *a priori* SNR

An important parameter of numerous speech enhancement techniques is the *a priori* SNR. In the well-known decision-directed approach, the *a priori* SNR depends on the speech spectrum estimation in the previous frame, which results in degradation of the speech enhancement performance. In order to alleviate this problem while keeping its benefits we have used the MMSE based two-step *a priori* SNR estimation approach proposed in [8], and which is expressed by

$$priori_{MMSE} = \frac{\hat{X}^2}{\Gamma_d} = \frac{\xi}{1+\xi} \left(1 + \frac{\xi}{1+\xi} \gamma\right), \quad (10)$$

where  $\xi$  is the *a priori* SNR is estimated using the DD approach given by (7) and  $\gamma$  is the *a posteriori* SNR given by (5).

#### 3.2 Noise Estimation

Noise estimation is also an important factor in speech enhancement systems. In this paper the noise power spectrum is estimated during speech pauses using the following recursive relation [6]:

if  $W(m,k) > 0$

$$\Gamma_d(m,k) = \lambda_d \Gamma_d(m-1,k) + (1-\lambda_d) W(m,k) |Y(m,k)|^2, \quad (11)$$

if  $W(m,k) = 0$

$$\Gamma_d(m,k) = \Gamma_d(m-1,k)$$

where  $\lambda_d$  is a smoothing factor satisfying  $0 < \lambda_d < 1$  and  $W(m,k)$  is the weighting factor on the noisy power spectrum. The weighting factor is designed so that it is almost inversely proportional to the estimated SNR (dB) given by

$$\tilde{\gamma}(m,k) = 10 \log_{10} \left( \frac{|Y(m,k)|^2}{\Gamma_d(m-1,k)} \right) \quad (12)$$

and the weighting factor is given by the following relation

$$W(m,k) = \begin{cases} 1 & \text{if } \tilde{\gamma}(m,k) \leq 0 \\ -\frac{1}{\tau} \tilde{\gamma}(m,k) + 1 & \text{if } 0 < \tilde{\gamma}(m,k) \leq \varepsilon \\ 0 & \text{if } \tilde{\gamma}(m,k) > \varepsilon \end{cases}, \quad (13)$$

where  $\tau$  is a slope deciding constant of the graph of (13) and  $\varepsilon$  is a threshold to eliminate an unreliable  $\tilde{\gamma}(m,k)$ . In this

paper we have used  $\lambda_d = 0.9$ ,  $\tau = 12$ , and  $\varepsilon = 6$  on the basis of simulations.

#### 3.3 Spectral Gain Calculation

The spectral gain  $H(m,k)$  for the Wiener denoising technique is given by (4). The spectral gain  $H_r(m,k)$  for the reference signal is given by

$$H_r(m,k) = \begin{cases} H(m,k) + \frac{\eta}{q} & \text{if } \left(H(m,k) + \frac{\eta}{q}\right) \leq 1 \\ 1 & \text{otherwise} \end{cases}, \quad (14)$$

where  $\eta$  and  $q$  are adjustment constants chosen experimentally. It is the shifting up of the Wiener denoising filter by  $\frac{\eta}{q}$ . In this paper we have used  $\eta = 0.35$  and  $3 \leq q \leq 10$

based on simulations. It is assumed that  $H_r(m,k)$  introduces minimum distortion and results in a reference signal which does not contain residual musical noise. The reference signal obtained using  $H_r(m,k)$  is used instead of the noisy speech signal to improve the accuracy of the musical noise detector [13].

#### 3.4 Proposed perceptual post filter

Musical noise is a major problem in speech enhancement. Several speech enhancement schemes have attempted to address the problem using various approaches. These have included time averaging [1], noise floors, over-subtraction of noise [2] or perceptual criteria [4, 10]. Our proposed method is based on perceptual criteria. The complete block diagram of our implemented denoising technique is shown in figure 1 whereas figure 2 shows the block diagram of the proposed perceptual post filter with a modified Johnston model.

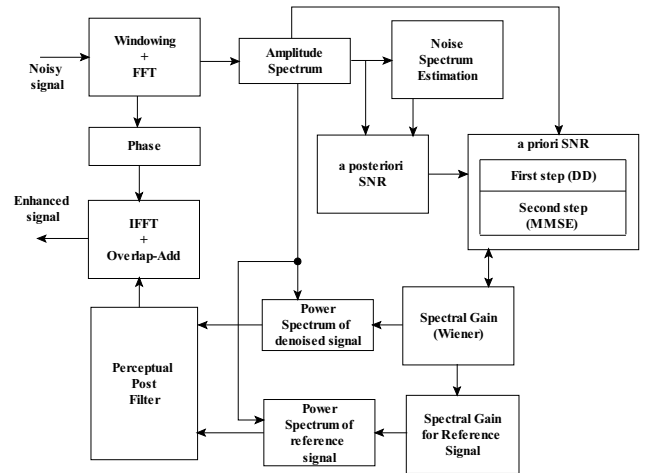


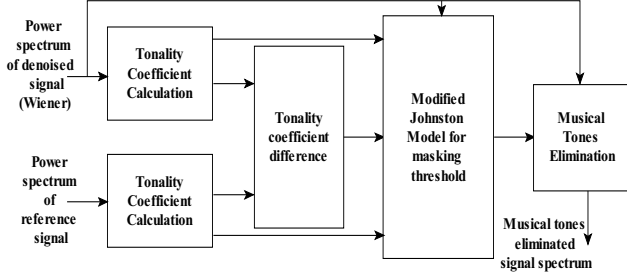
Figure 1 Block diagram of the perceptual Wiener denoising technique.

#### 3.4.1 Tonality coefficient calculation

In the denoised signal obtained by subtractive methods, annoying musical tones appear in the power spectrum, which leads to an increase of tonality coefficient. Thus it is possible to detect the presence or absence of musical tones by means of a tonality coefficient. The steps for calculating the tonality coefficient are taken from [11] and described below.

The first step is the frequency analysis of both signals along the critical band (CB). The power spectrum of the denoised signal and that of the reference signal are partitioned in critical bands.

We have considered CBs between 0 kHz and 4 kHz as we chose the Aurora database, having the sampling frequency of 8 kHz, for this experiment. In this paper we tried to detect CB musical noise for the CBs between 9 and 18 as the annoying musical noise is situated only in the frequency range between 1 kHz and 4 kHz. For frequencies under 1 kHz it is masked by the presence of real tones of clean speech [12].



**Figure 2** Block diagram of the Perceptual post filter.

In the second step, the tonality coefficient is measured using the ratio of the geometric mean (GM) and the arithmetic mean (AM) of the signal power spectrum, known as the spectral flatness measure (SFM). SFM is used to determine whether the signal is tone-like or noise-like. The coefficient of tonality is expressed as

$$tc(i) = \min\left(\frac{SFM_{dB}(i)}{-60}, 1\right) \quad (15)$$

where  $i$  is the CB index and  $SFM_{dB}(i)$  is given as

$$SFM_{dB}(i) = 10 \log_{10}\left(\frac{GM}{AM}\right). \quad (16)$$

The tonality coefficient of the pure tone is 1 and is close to zero for the noise-like signal. Using (15) and (16) the tonality coefficients for the denoised signal  $tc_d(i)$  and that of the reference signal  $tc_r(i)$  are computed for each CB. Then, the tonality coefficient difference for each CB is given by

$$\Delta tc(i) = tc_d(i) - tc_r(i). \quad (17)$$

Musical residual noise appears in the  $i$ th CB if  $tc_d(i) > tc_r(i)$  and it becomes audible if  $\Delta tc(i) > T'(i)$ , where  $T'(i)$  is the threshold for the  $i$ th CB, which depends on the order of the CB and masking properties of the human ear. Following the same procedure as described in [11], the threshold,  $T'(i)$  for all CBs is found to be approximately constant and is  $T'(i) = 0.06$  [13].

### 3.4.2 The modified Johnston model

The Johnston model for masking threshold, thanks to its computational simplicity, was developed for audio coding and is used to control the quantization process of spectral components [11]. The main difference between the Johnston model for masking threshold and the modified Johnston model for masking threshold lies in the calculation of the offset for the masking energy in each CB. The steps of the modified Johnston model for masking threshold are

- Partition of the signal power spectrum into CBs and the energies,  $E(i)$  in each CB are added.
- Calculation of the spread CB spectrum  $C(i)$  by convolving the spread function  $SF(i)$  and the bark spectrum  $E(i)$  in order to take into account the masking effect between different CBs.
- In the Johnston model [11], an offset  $O(i)$  is determined according to the tonality coefficient and the CB order as

$$O(i) = tc(i)(14.5 + i) + (1 - tc)5.5 \text{ dB}. \quad (18)$$

The total tonality coefficient used by Johnston gives a general idea about the nature of the power spectrum. In our context, we seek to detect musical noise in the selected CBs, i.e., between 9 and 18. The CB tonality coefficient of the denoised signal was taken into account when calculating the threshold offset  $O(i)$  in the proposed method.

In the modified model, a Boolean flag  $M(i)$  is constructed first based on the tonality coefficient difference  $\Delta tc(i)$ , and the threshold  $T'(i)$ , over which an additive tone becomes audible in the presence of narrow-band noise.  $\Delta tc(i)$  and  $T'(i)$  were determined in section 3.4.1. The Boolean flag  $M(i)$  indicates the presence or absence of CB musical noise and is given as

$$M(i) = \begin{cases} 1 & \text{if } \Delta tc(i) \geq T'(i) \\ 0 & \text{otherwise} \end{cases}. \quad (19)$$

In case of critical-band musical noise ( $M(i) = 1$ ), the used tonality coefficient to calculate  $O(i)$  is close to one. However, it shouldn't be for better estimation of masking threshold. For better estimation we proposed to correct the tonality coefficient of the  $i$ th musical critical band by replacing it with the tonality coefficient of the  $i$ th critical band tonality coefficient of the reference signal. Thus the corrected offset threshold for the modified Johnston model becomes

$$O(i) = tc_m(i)(14.5 + i) + (1 - tc_c(i))5.5 \text{ dB} \quad (20)$$

where  $tc_m(i) = \begin{cases} tc_d(i), & \text{for } M(i) = 1 \\ tc_r(i), & \text{for } M(i) = 0 \end{cases}$ . (21)

The threshold offset is then subtracted from the spread CB spectrum to yield the spread threshold estimate  $T(i)$

$$T(i) = 10^{\lceil \log_{10}(C(i)) - (O(i)/10) \rceil}. \quad (22)$$

- In the final step, the noise masking threshold (NMT) is estimated as

$$NMT(i) = \max(T_q(i), T(i)), \quad (23)$$

where  $T_q(i)$  is the absolute threshold of hearing.

## 4. Experimental Results and Discussion

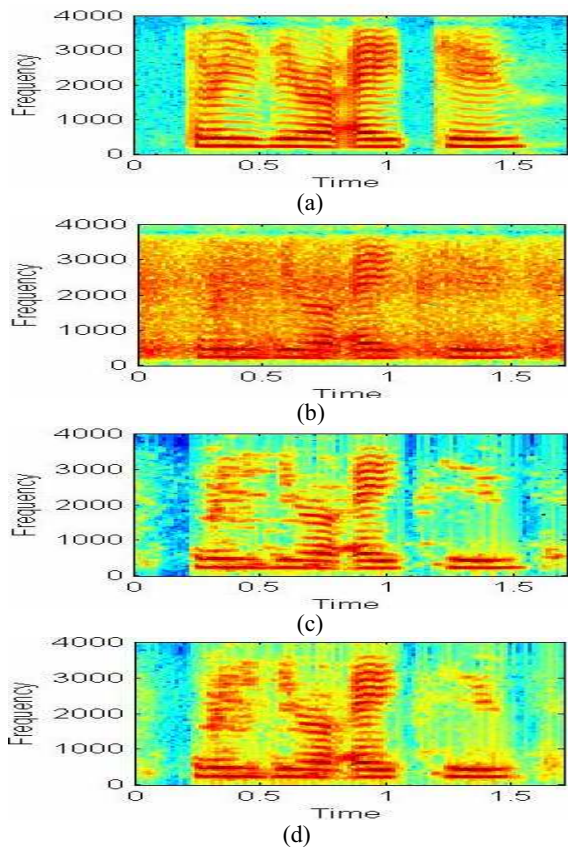
In order to evaluate the performance of the proposed perceptual Wiener denoising technique, we conducted extensive objective quality tests under various noisy environments. The frame sizes were chosen to be 256 samples (32 msec) long with 40% overlap; a sampling frequency of 8 kHz and a hamming window were applied. To evaluate and compare the performance of the proposed perceptual Wiener denoising technique, we carried out simulations with the *TEST A* database of Aurora [9]. Speech signals were degraded with five types of noise at global SNR levels of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB. The noises were N1 (Subway noise), N2 (Babble Noise), N3 (Car Noise), N4 (Exhibition Hall Noise) and WGN (White Gaussian Noise). Table 1 shows the average segmental SNR (<sup>1</sup>SegSNR), Log spectral distance (<sup>2</sup>LSD) and Perceptual Evaluation of Speech Quality (<sup>3</sup>PESQ) scores for the enhanced speech signals in various types of noisy environments [14]. It is observed that the proposed approach yields better SegSNR than that of the Wiener denoising technique under all tested noisy

<sup>1</sup>The higher value of the SegSNR indicates the weaker speech distortions.

<sup>2</sup>The higher LSD reflects the stronger speech distortions.

<sup>3</sup>The higher PESQ score indicates better perceived quality.

environments. In the case of the LSD measure the proposed method exhibits lower values of LSD for almost all noisy environments compared to those obtained by the conventional Wiener denoising technique and in the case of the PESQ measure, the proposed perceptual Wiener denoising technique gives better PESQ scores than the Wiener denoising technique. Figure 3 represents the spectrograms of the clean speech signal, noisy signal and enhanced speech signals obtained using the Wiener denoising technique and the proposed technique. The speech spectrograms provide more accurate information about the residual noise and speech distortion than the corresponding time domain waveforms. We compared the spectrograms for each of the methods and confirmed a reduction of the residual noise and speech distortion. Speech spectrograms presented in Figure 3 use a Hamming window of 256 samples with 50% overlap and the noisy signals include N1 (Subway Noise) with SNR = 5 dB. It is seen that the musical noise is almost removed for the most part in figure 3(d).



**Figure 3** Speech spectrograms, N1 (subway noise), SNR = 5 dB. (a) Clean signal, (b) noisy signal, (c) enhanced signal (Wiener), and (d) enhanced signal (proposed).

## 5. Conclusion

In order to improve speech enhancement performance by detecting and eliminating musical phenomena in the denoised signal using Wiener filtering, a new perceptual Wiener denoising approach is proposed. This method has the advantage to be applied as a post-processing for any subtractive denoising technique. The proposed approach is based on the detection of musical critical bands using the tonality coefficient and a modified Johnston masking threshold. Experimental results and plotted spectrograms show that proposed technique performs better in all tested objective quality measures; it does not introduce additional speech distortion, and results in significant reduction of the musical phenomenon.

Noise Type	Input SNR(dB)	SegSNR		LSD		PESQ	
		Wiener	Perceptual Wiener	Wiener	Perceptual Wiener	Wiener	Perceptual Wiener
N1	-5	-2.779	<b>-2.085</b>	2.277	<b>2.036</b>	1.316	<b>1.338</b>
	0	-1.199	<b>-0.357</b>	1.948	<b>1.912</b>	1.578	<b>1.751</b>
	5	2.124	<b>3.547</b>	1.519	<b>1.501</b>	2.351	<b>2.403</b>
	10	6.046	<b>6.456</b>	1.508	<b>1.490</b>	2.776	<b>2.852</b>
	15	9.074	<b>10.029</b>	1.269	<b>1.180</b>	3.169	<b>3.200</b>
N2	-5	-3.995	<b>-3.092</b>	2.192	<b>2.030</b>	1.075	<b>1.152</b>
	0	-2.206	<b>-1.945</b>	1.931	<b>1.822</b>	1.698	<b>1.783</b>
	5	0.862	<b>1.215</b>	1.602	<b>1.579</b>	2.295	<b>2.323</b>
	10	3.456	<b>3.795</b>	1.359	<b>1.299</b>	2.734	<b>2.757</b>
	15	4.679	<b>5.571</b>	1.381	<b>1.260</b>	2.780	<b>2.842</b>
N3	-5	-2.607	<b>-1.838</b>	1.803	<b>1.650</b>	1.503	<b>1.623</b>
	0	0.142	<b>1.025</b>	1.568	<b>1.478</b>	2.030	<b>2.150</b>
	5	3.034	<b>3.523</b>	1.409	<b>1.303</b>	2.581	<b>2.632</b>
	10	5.915	<b>6.188</b>	1.351	<b>1.213</b>	2.929	<b>3.012</b>
	15	8.801	<b>9.747</b>	1.332	<b>1.169</b>	3.194	<b>3.315</b>
N4	-5	-1.881	<b>-1.348</b>	2.024	<b>1.901</b>	0.599	<b>0.685</b>
	0	0.783	<b>0.845</b>	1.684	<b>1.637</b>	1.564	<b>1.687</b>
	5	2.332	<b>2.510</b>	1.734	<b>1.570</b>	1.631	<b>1.985</b>
	10	5.473	<b>6.310</b>	1.657	<b>1.596</b>	2.225	<b>2.245</b>
	15	8.937	<b>9.443</b>	1.369	<b>1.299</b>	2.549	<b>2.634</b>
WGN	-5	-1.989	<b>-1.325</b>	2.213	<b>1.990</b>	1.645	<b>1.952</b>
	0	0.735	<b>2.063</b>	2.066	<b>1.827</b>	2.179	<b>2.312</b>
	5	3.527	<b>5.393</b>	1.924	<b>1.501</b>	2.497	<b>2.631</b>
	10	6.682	<b>8.428</b>	1.509	<b>1.279</b>	2.957	<b>3.015</b>
	15	9.620	<b>10.691</b>	1.317	<b>1.241</b>	3.257	<b>3.323</b>

**Table 1** Experimental Results.

## 6. Reference

- [1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.
- [2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, (Washington, DC), pp. 208–211, Apr. 1979.
- [3] H. L. V. Trees, *Detection, Estimation, and Modulation: Part I - Detection, Estimation and Linear Modulation Theory*. John Wiley and Sons, Inc., 1st ed., 1968.
- [4] Y.M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing*, vol.39, no.9, pp.1943–1954, 1991.
- [5] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psychoacoustic criteria," *IEEE ICASSP*, pp.359–362, Minneapolis, MN, 1993.
- [6] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer-Verlag, 2nd ed., 1999.
- [7] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimation," *IEEE Trans. Acoust. Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
- [8] Md. Jahangir Alam, Douglas O'Shaughnessy and Sid-Ahmed Selouani, "Speech enhancement based on novel two-step *a priori* SNR estimators," to appear in *INTERSPEECH Conference*, Brisbane, Australia, September 2008.
- [9] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy environments," *ISCA ITRW ASR*, September 2000.
- [10] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol.7, no.2, pp.126–137, 1999.
- [11] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. on Selected Areas in Comm.*, vol. 6, pp. 314–323, Feb. 1988.
- [12] A. Ben Aicha, S. Ben Jebara and D. Pastor, "Speech denoising improvement by musical tones shape modification," *International Symposium on Communication, Control and Signal Processing ISCCSP*, Morocco 2006.
- [13] A. Ben Aicha, S. Ben Jebara, "Perceptual musical noise reduction using critical bands tonality coefficients and masking thresholds," *INTERSPEECH Conf.*, pp. 822–825, Antwerp, Belgium, August 2007.
- [14] J. H. L. Hansen and B. L. Pellom, "An effective evaluation protocol for speech enhancement algorithms," in *Proc. ICSLP*, vol. 7, pp. 2819–2822, Sydney, Australia, 1998.