Perceptual improvement of Wiener filtering

Asmaa Amehraye, Dominique Pastor, Ahmed Tamtaoui

▶ **To cite this version:**

Asmaa Amehraye, Dominique Pastor, Ahmed Tamtaoui. Perceptual improvement of Wiener filtering. ICASSP 2008 : IEEE International Conference on Acoustics, Speech and Signal Processing, Mar 2008, Las Vegas, United States. pp.2081 - 2084, 10.1109/ICASSP.2008.4518051 . hal-02136069

HAL Id: hal-02136069
https://hal.science/hal-02136069v1

Submitted on 21 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PERCEPTUAL IMPROVEMENT OF WIENER FILTERING

A. Amehraye[1],[2], *D. Pastor*[1]

[1] Télécom Bretagne,
Technopôle Brest Iroise, 28238 Brest, France,
[2] GSCM-LRIT, Faculté des Sciences, Rabat, Maroc
asmaa.amehraye,dominique.pastor@enst-bretagne.fr

A. Tamtaoui[3]

[3] Institut National des Postes
et Télécommunications, 2 avenue Allal
El Fassi, Madinat Al Irfane, Rabat - Morocco
tamtaoui@inpt.ac.ma

ABSTRACT

This paper deals with musical noise resulting from subtractive type algorithms and especially Wiener filtering. We compare several methods that introduce perceptually motivated modifications of standard Wiener filtering and we propound a new speech enhancement technique. It aims to improve the quality of the enhanced speech signal provided by standard Wiener filtering by controlling the latter via a second filter regarded as a psychoacoustically motivated weighting factor. According to objective measures, the described process results in significant reduction of musical noise.

Index Terms— Wiener filtering, distortion, auditory system, speech enhancement, musical noise

1. INTRODUCTION

The objective of a speech enhancement process is to improve the quality and intelligibility of speech in noisy environments. The problem has been widely discussed over years. Many approaches have been proposed. Basic methods are subtractive type algorithms such as that described in [1]. Such methods return residual noise known as musical noise. This type of noise is quite annoying. In order to reduce the effect of musical noise, several solutions have been proposed. Some involve adjusting parameters of the spectral substraction so as to offer more flexibility as in [2] and [3]. Others, such as that proposed in [4], are based on signal subspace approaches. Despite the effectiveness of these techniques to improve the Signal to Noise Ratio (SNR), the problem of eliminating or reducing musical noise is still a challenge to many researchers.

In the last few decades, the introduction of psychoacoustic models has attracted a great deal of interest. The objective is to improve the perceptual quality of the enhanced speech signal. In [3], a psychoacoustic model is used to control the parameters of the spectral subtraction in order to find the best trade-off between noise reduction and speech distortion. To make musical noise inaudible, the linear estimator proposed in [5] incorporates the masking properties of the human auditory system. In [6], the masking threshold of tones and an intermediate signal, which is slightly denoised and free of musical noise, are used to detect musical tones generated by spectral subtraction methods. This detection can be used by a post-processing aimed at reducing the detected tones.

Even though the psychoacoustic models are usually developed in the frequency domain, signal subspace approaches can also involve perceptual models by resorting to some suitable frequency to eigendomain transform as described in [7], [8], [9].

In this work, we are particularly interested in methods related to standard Wiener filter for two reasons. First, the Wiener filtering is easy to implement. Second, it can reasonably be expected that if we succeed in reducing the perception of residual noise resulting from Wiener filtering, the quality of the denoised speech will be improved and yield a fairly satisfactory comfort of listening.

On the basis of such remarks, the authors in [10] propose to apply the Wiener filter only when noise is audible and, thus, to not process frequency components where noise is masked. Similarly, in [11], a perceptually motivated modification is applied to the Wiener filtering of the noisy speech signal sub-band components, these components being calculated via a filterbank.

In the present paper, we propose to control the standard Wiener filtering by a psychoacoustically motivated filter that can be regarded as a weighting factor. The purpose is to minimise the perception of musical noise without degrading the clarity of the enhanced speech. We compare the proposed method to those introduced in [10], [11] and [12] when the noisy speech signals are analysed in the Fourier domain. This is the reason why we adapt the method proposed in [11] to the frequency domain.

The organization of the paper is as follows. Section 2 recalls the basics concerning standard Wiener filtering of noisy speech signals. With the same notations and assumptions of section 2, section 3 introduces several enhancement processes, amongst which the new method we propose. Section 4 presents the performance evaluation by means of objective measures and the observation of spectrograms. Section 5 concludes this paper.

2. STANDARD FILTERING AND ITS LIMITATIONS

The observed noisy speech signal is assumed to be some speech signal additively corrupted by independent noise. The processing is performed frame by frame in the frequency domain. Each frame involves the same number M of samples. For the k^{th} frame, let $s_k(t), n_k(t)$ and $y_k(t), t = 0, 1, \ldots, M-1$, stand for the M samples of the speech signal, noise and the observed noisy speech signal, respectively. We thus have $y_k(t) = s_k(t) + n_k(t)$. Now, let $Y_k(\nu)$, $S_k(\nu)$ and $N_k(\nu)$, $\nu = 0, \ldots, M-1$, denote the Discrete Fourier Transform (DFT) coefficients of $y_k(t)$, $s_k(t)$ and $n_k(t)$, $t = 0, 1, \ldots, M-1$, respectively. For every $\nu = 0, 1, \ldots, M-1$, we have $Y_k(\nu) = S_k(\nu) + N_k(\nu)$.

Basic speech enhancement approaches involves estimating every frequency component $S_k(\nu)$ by $\widetilde{S}_k(\nu) = H_k(\nu)Y_k(\nu)$ where $H_k(\nu)$ is an estimator chosen according to a suitable criterion. The error signal generated by this estimator is

$$\begin{aligned} e_k(\nu) &= \widetilde{S}_k(\nu) - S_k(\nu) \\ &= (H_k(\nu) - 1)S_k(\nu) + H_k(\nu)N_k(\nu). \end{aligned} \quad (1)$$

The values $(H_k(\nu) - 1)S_k(\nu)$ are the DFT coefficients of the speech distortion due to the filtering and the frequency components $H_k(\nu)N_k(\nu)$ are the residual noise DFT coefficients. Musical noise then results from the pure tones present in residual noise. The Wiener filtering based on Malah's decision-directed approach [1] is one of the most famous methods aimed at reducing musical noise. In this case, the estimator is $H_k(\nu) = W_k(\nu)$ and $\widetilde{S}_k(\nu) = W_k(\nu)Y_k(\nu)$ is the Wiener estimate of $S_k(\nu)$ where $W_k(\nu)$ is hereafter called the Wiener gain function and is given by

$$W_k(\nu) = \xi_k(\nu)/(1 + \xi_k(\nu)) \tag{2}$$

where

$$\xi_k(\nu) = (1-\alpha)h\left(\chi_k(\nu) - 1\right) + \alpha\frac{|\widetilde{S}_{k-1}(\nu)|^2}{\gamma_k(\nu)} \tag{3}$$

is the so-called decision-directed estimate of the *a priori* SNR

$$\mathbb{E}\left[|S_k(\nu)|^2\right]/\mathbb{E}\left[|N_k(\nu)|^2\right]. \tag{4}$$

In Eq. (3), $\widetilde{S}_{k-1}(\nu) = W_{k-1}(\nu)Y_{k-1}(\nu)$ is the ν^{th} spectral component of the Wiener denoised speech signal in frame $k-1$; $\gamma_k(\nu)$ is the estimate of $\mathbb{E}\left[|N_k(\nu)|^2\right]$; $h(x) = x$ if $x \geq 0$ and $h(x) = 0$ otherwise; $\chi_k(\nu) = |Y_k(\nu)|^2/\gamma_k(\nu)$ is the estimate of the *a posteriori* SNR $|Y_k(\nu)|^2/\mathbb{E}\left[|N_k(\nu)|^2\right]$; the weighting factor α is set to 0.98 for a good compromise between musical noise and speech distortion [1] . The estimate $\xi_k(\nu)$ takes into account the current frame, with weight $(1-\alpha)$, and the result of the processing of the previous frame, with weight α. The smoothing character of the decision-directed approach reduces the level of musical noise, which, however, remains present and perceptually annoying.

3. IMPROVEMENT OF WIENER FILTERING

3.1. Proposed generalized block diagram

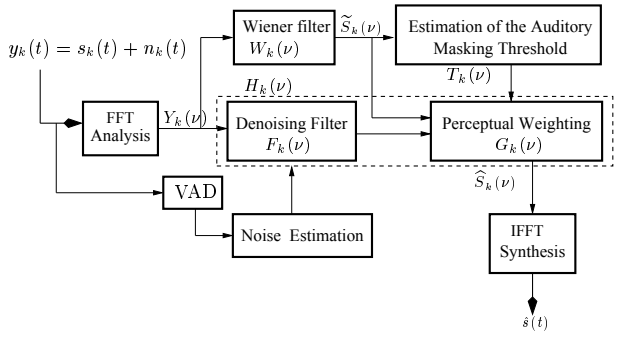

Fig. 1. Block diagram of the proposed enhancement process

The block diagram of figure 1 summarizes the different speech enhancement processes discussed in this section and compared in the next one. It allows some improvement of the Wiener filtering of frame k by choosing $F_k(\nu)$ equals to $W_k(\nu)$ and introducing perceptual criteria through the filter $G_k(\nu)$. The purpose is to achieve a good compromise, in a perceptual sense, between residual noise and speech distortion. For frame k, the resulting estimate $\widehat{S_k}(\nu)$ of $S_k(\nu)$ is $\widehat{S_k}(\nu) = H_k(\nu)Y_k(\nu)$, where $H_k(\nu) = G_k(\nu)F_k(\nu) = F_k(\nu)G_k(\nu)$.

Figure 1 also shows that the computation of the masking threshold $T_k(\nu)$ will be based on the Wiener estimate of the clean speech signal for all the methods described below. The masking threshold could also be estimated on the basis of the outcome of a spectral substraction as in [3]. However, the tone-like nature of musical noise increases the energy per critical band and the presence of too much musical noise can therefore induce an overestimation of the masking threshold. The Wiener estimate is thus preferable because the Wiener filter introduces less musical noise than spectral subtraction methods [1]. In this paper, the power spectrum of the noisy signal is estimated on the basis of signal-free time frames, which are detected using the Voice Activity Detector (VAD) of the ITU-T standard G729 (8kbits/s) [13].

3.2. Some previous works

Before introducing the speech enhancement method we propose, we describe two recent techniques that can be regarded as perceptually motivated modifications of the Wiener filter. The first one (A), described in [11], is a Wiener filtering of only the amount of noise that exceeds the masking threshold. In [11], this approach is applied to the sub-band components obtained by using an auditory filterbank. In fact, this method can easily be adapted to the usual case where the time-frequency analysis is performed by the standard DFT. With respect to the block diagram of figure 1, it involves choosing

$$\text{(A):}\begin{cases} F_k(\nu) = 1, \\ G_k(\nu) = \dfrac{|\widetilde{S}_k(\nu)|^2}{(|\widetilde{S}_k(\nu)|^2 + \max\left(\gamma_k(\nu) - T_k(\nu), 0\right))} \end{cases} \tag{5}$$

where $\widetilde{S}_k(\nu)$ is the Wiener estimate defined above.

In the second method (B), introduced in [10], the Wiener filtering is controlled by the result of the comparison between noise and the masking threshold. This comparison makes it possible to perform the denoising only for the noise frequency components that are audible in the sense that their amplitudes exceed the masking threshold. On the basis of the block diagram of figure 1 and with the same notations as those used so far, the perceptually motivated modification of the Wiener filter proposed in [10] involves writing

$$\text{(B):}\begin{cases} F_k(\nu) = 1, \\ G_k(\nu) = \begin{cases} W_k(\nu) & \text{, if } \gamma_k(\nu) > T_k(\nu) \\ 1 & \text{, otherwise.} \end{cases} \end{cases} \tag{6}$$

3.3. Proposed method

Remark: *Perceptual methods basically aim at reducing only audible noise to avoid much distortion. Therefore, noise components that are not audible, thanks to some maskers in the original noisy signal, are still present after denoising and can even become audible if the maskers are filtered. This is a limitation of some current perceptual methods.*

The method we propose is an attempt to overcome this type of drawback by using a filter G that acts as a perceptual weighting factor controlling the Wiener gain function. The proposed filter $H_k(\nu)$ is then a concatenation of two active filters. Therefore, comparing to the block diagram of figure 1, this double filtering is specified by

$$\text{Double filtering:}\begin{cases} F_k(\nu) = W_k(\nu), \\ G_k(\nu) = \dfrac{|\widetilde{S}_k(\nu)|^2}{(|\widetilde{S}_k(\nu)|^2 + \max\left(\gamma_k(\nu) - T_k(\nu), 0\right))} \end{cases} \tag{7}$$

The following analysis describes some properties of this "Double filtering" on the basis of Eq. (7).

If $\gamma_k(\nu) < T_k(\nu)$, which means that noise is inaudible in frame k, we have $G_k(\nu) = 1$. The Wiener filter is however applied for two reasons. First, it favours the gain in SNR. Second, it reduces the risk that non audible noise components might become audible after the filtering of audible maskers present in the original noisy signal (see the remark above). Note that if $\gamma_k(\nu) \ll T_k(\nu)$, that is, when the SNR is very good, the Wiener filter and the perceptual weighting factor both equal 1 so that no distortion is introduced.

When $\gamma_k(\nu) > T_k(\nu)$, we couple the high noise suppression capability of the Wiener filtering with the effect of the weighting factor so as to enhance the speech quality and reduce musical noise. In the limit case where $\gamma_k(\nu) \gg T_k(\nu)$, we have $\xi_k(\nu) \ll 1$ and $W_k(\nu)G_k(\nu)$ tends more quickly to 0 than $W_k(\nu)$. We can say that the proposed method accentuates the denoising when noise is perceptually significant.

We suggest applying a smoothing operation to avoid discontinuities, in the gain function $H_k = G_k W_k$, that result from the frequency selectivity of the filtering procedure. We choose the smooth correlogram, which is a circular convolution between H and a weighting window C. The values $C(\nu)$, $\nu = 0, 1, \ldots, N-1$, satisfy the two conditions: $\sum_{\nu=0}^{N-1} C(\nu) = 1$ and

$$C(\nu) = \frac{(0.5 + 0.5\cos(2\pi\nu/N))^q}{\sum_{\nu=0}^{N-1}(0.5 + 0.5\cos(2\pi\nu/N))^q} \quad (8)$$

The smoothing effect is illustrated by figure 2.

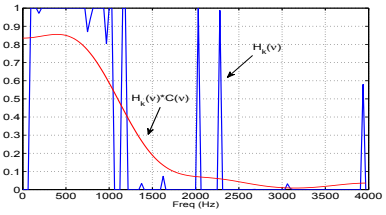

Fig. 2. Smoothing effect

For our experiments, we consider another perceptual filter (C) proposed in [12], which is not designed to improve Wiener filtering but which is interesting for comparison. This filter is designed so as to yield inaudible residual noise by forcing the residual noise spectral power to be below the masking threshold. It obeys the following equation

$$\text{(C)}: \begin{cases} F_k(\nu) = 1, \\ G_k(\nu) = \min\left(\sqrt{T_k(\nu)/\gamma_k(\nu)}, 1\right) \end{cases} \quad (9)$$

Summarizing, the common feature of the three perceptual filters defined by (5), (6) and (9) is to not process the noisy speech signal when noise is perceptually insignificant. In contrast, our approach (see Eq. 7) involves activating the Wiener filtering even when noise is not audible. By so proceeding, it is expected to reduce the amount of background noise that could result in audible musical noise after the filtering of adjacent maskers (see the remark above).

4. EXPERIMENTAL RESULTS

We have compared the five methods presented in the foregoing, that is the adaptation to the frequency domain of the (A) filter introduced in [11] (see Eq. (5)), the (B) filter propounded in [10] and specified by Eq. (6), the (C) filter proposed in [12] and summarized by Eq. (9), the standard Wiener filter (see Eq. (2)) employing the decision-directed approach of Eq. (3) and finally the "Double filtering" of Eq. (7). This comparison has been performed on speech signals from the TIDIGITS database downsampled to 8 KHz before adding white Gaussian noise or babble noise from the Noisex database at specific SNR's.

The experimental results of this section have been obtained via the following protocol. Short-time windows (32ms) of noisy speech, with 50% overlap, are transformed into the frequency domain using the short-time Fast Fourier transform. As mentioned above, the auditory masking threshold is computed on the basis of the Wiener estimate. The different calculation steps of the masking threshold are those described in [14]. The power spectrum of the noisy signal is estimated on the basis of signal-free time frames, detected by the G729 VAD as mentioned in section 3. The smoothing is applied to every perceptual method with $q = 20$ (see Eq.(8)), an experimental value which gives the best results presented below.

The enhanced speech signal $\hat{s}(t)$ in the time domain is obtained using the overlap-and-add approach after transformation back into the time domain via the Short-Time Inverse Fast Fourier Transform.

The five methods addressed in this paper have been assessed by

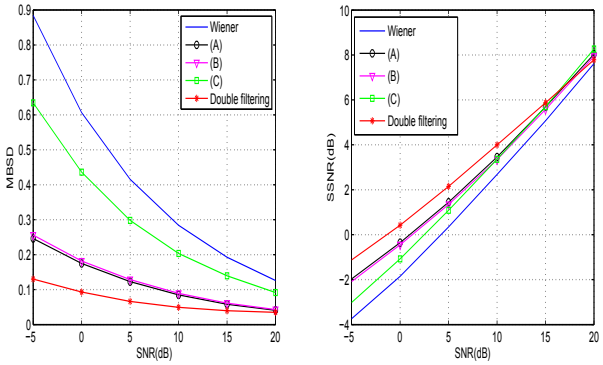

Fig. 3. Comparative performance, in terms of mean MBSD and segmental SNR measures, in the case of speech corrupted by white gaussian noise

means of objective measures, namely the standard Segmental Signal to Noise Ratio (SSNR) and the Modified Bark Spectral Distortion (MBSD). The SSNR is the average of the SNR values on short segments. The MBSD proves to be highly correlated with subjective speech quality assessment [15]. Figure 3 (resp. figure 4) presents the average MBSD and the average SSNR for 250 TIDIGITS sentences corrupted by additive white gaussian noise (resp. babble noise) with SNR from -5dB to 20dB.

According to these results, the proposed method achieves a significant improvement in comparison with the other described methods. Speech spectrograms are suitable for observing residual noise structure. The spectrograms presented in Figure 5 illustrate the capability of the double filtering to reduce musical noise without introducing too much distortion. From this figure, we can see that some

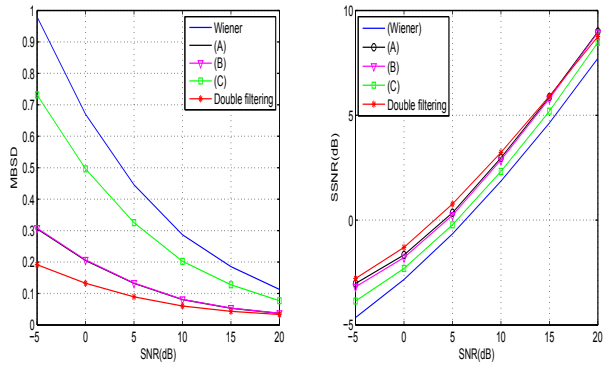

Fig. 4. Comparative performance, in terms of mean MBSD and segmental SNR measures, in the case of speech corrupted by babble noise

undesired random tone peaks present in (b) and (c) are practically non-existent in (d) thanks to the double filtering.

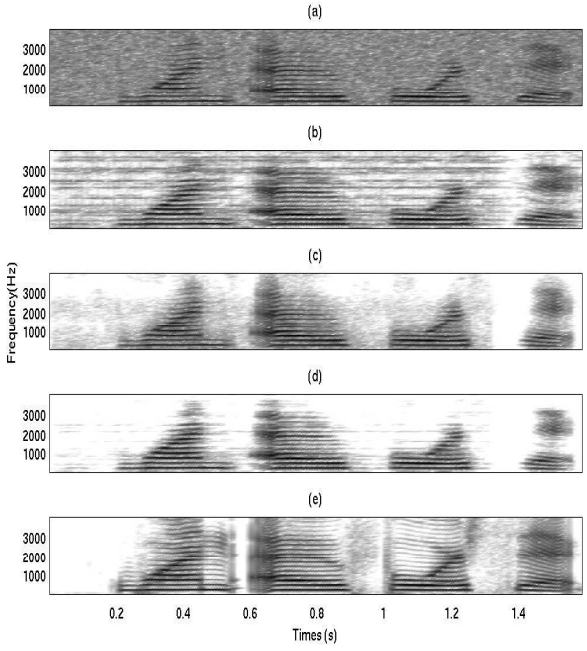

Fig. 5. Speech spectrograms of a test utterance. (a) Noisy speech corrupted by white gaussian noise at 10dB SNR. (b) Noisy speech enhanced by Wiener filtering, (c) Noisy speech enhanced by (A) filtering, (d) Noisy speech enhanced by "Double filtering" and (e) Clean speech. (We compare to (A), the best second method)

5. CONCLUSION

In this paper, an effective approach for suppressing musical noise present after Wiener filtering has been introduced. Based on the perceptual properties of the human auditory system, a weighting factor accentuates the denoising process when noise is perceptually significant and prevents that residual noise components might become audible in the absence of adjacent maskers. When the speech signal is additively corrupted by white Gaussian noise or babble noise, objective measure results showed the improvement brought by the proposed method in comparison to some recent filtering techniques of the same type.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec 1984.

[2] R. Schwartz M. Berouti and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. of ICASSP, 1979, vol. I, pp. 208–211.

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

[4] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251–266, 1995.

[5] Y. Hu and P. Loizou, "Incorporating a psychoacoustic model in frequency domain speech enhancement," *IEEE Signal Processing Letters*, vol. 11(2), pp. 270–273, 2004.

[6] A. Ben Aicha and S. Ben Jebara, "Utilisation de la courbe de masquage pour la détection des tonales musicales artificielles dans un signal de parole débruité par approche spectrales," Proc. of ISIVC, 2006, vol. I.

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

[8] You C. H, Rahardja S, and Koh S. N., "Audible noise reduction in eigendomain for speech enhancement," *IEEE Trans. Audio., Speech, and Language Processing*, vol. 15, pp. 1753–1765, August 2007.

[9] S. Kim J. Kim and C. Yoo, "The incorporation of masking threshold to subspace speech enhancement," Proc. of ICASSP, 2003, vol. I, pp. 76–79.

[10] P. Scalart C. Beaugeant, V. Turbin and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Processing*, vol. 64, pp. 33–47(15), Jan 1998.

[11] W. H. Holmes L. Lin and E. Ambikairajah, "Speech denoising using perceptual modification of wiener filtering," *IEE Electronic Letters*, vol. 38, pp. 1486–1487, Nov 2002.

[12] T. Lee and Kaisheng Yao, "Speech enhancement by perceptual filter with sequential noise parameter estimation," Proc. of ICASSP, 2004, vol. I, pp. 693–696.

[13] IUT-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear prediction (cs-acelp)," 1996.

[14] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Jour. Selected Areas Commun*, vol. 6, pp. 314–323, 1988.

[15] M. Dixon W. Yang and R Yantorno, "Modified bark spectral distortion measure which uses noise masking threshold," IEEE Speech coding Workshop, 1997.