A PERCEPTUAL APPROACH TO REDUCE MUSICAL NOISE PHENOMENON WITH WIENER DENOISING TECHNIQUE

Sofia Ben Jebara

Research unit TECHTRA
Ecole Supérieure des Communications de Tunis
Route de Raoued 3.5 Km, Cité El Ghazala, 2083, Ariana, Tunisia
email: sofia.benjebara@supcom.rnu.tn

ABSTRACT

Traditional denoising techniques, powerful in term of noise reduction, have the drawback of generating an annoying musical noise. This paper addresses the problem of enhancing speech in highly noisy environments using perceptual considerations. The post-processing technique we develop, considers the masking threshold of both noisy speech and the denoised one, to detect musical noise components. Next, to make them inaudible, detected musical noise candidates are set under the noise masking threshold and their closest neighbors are smoothed. Extensive subjective and objective tests have shown that, after enhancement, the musical noise is well reduced even at very low signal to noise ratios.

1. INTRODUCTION

In many speech communication systems, such as mobile telephony, reducing noise in corrupted speech is a challenging task especially in high noise level. A large number of speech enhancement techniques have been proposed in the past. They are predominantly based on spectral subtraction [1] and Wiener filtering [2]. Although their improvement in term of noise reduction, the main drawback is the appearance of an annoying residual noise, often referred to as musical noise. Later proposals rely on psychoacoustical considerations. Mainly, they exploit the masking properties of the auditory system. For example, according to the enhancement scheme proposed in [3][4], only the audible noise components are estimated and suppressed. Other approaches introduce a perceptual modification on traditional denoising systems [5][6].

In this paper, we develop a post-processing method for reducing the musical residual noise imposed by classical spectral denoising systems. In this method, the auditory masking threshold is estimated twice, once for musical noise detection and once for musical noise reduction. The proposed idea to detect musical noise is based on the fact that musical noise components are present only in the denoised signal and lie above the noise masking threshold. On the other side, the same frequency components of noisy speech lie under their related noise masking threshold. Hence, to detect musical noise, some comparison rules will be applied. To reduce musical noise, detected musical noise candidates are set under the noise masking threshold and their closest neighbors are smoothed.

The remaining part of the paper is organized as follows. In section 2, the proposed denoising scheme is briefly outlined and useful backgrounds are given. Section 3 provides details about the musical noise detector based on masking threshold. Section 4 describes the way to reduce musical components. In section 5, evaluation results using temporal, spectral and perceptual criteria are presented. In the last section, we draw conclusions together with possible future improvements.

2. BASELINE SPEECH ENHANCEMENT SYSTEM

2.1. Notations

Let the corrupted speech signal $y(k)$ be presented as

$$y(k) = s(k) + n(k), \tag{1}$$

where $s(k)$ is the clean speech signal and $n(k)$ is the noise signal, they are assumed to be uncorrelated. The processing is done on a frame-by-frame basis. The Short Time Fourier Transform (STFT) is used and the previous model is re-written

$$\mathbf{Y}(m,f) = \mathbf{S}(m,f) + \mathbf{N}(m,f), \tag{2}$$

where m denotes the frame index and f is the frequency.

The denoised speech short-time magnitude $|\tilde{\mathbf{S}}(m,f)|$ is obtained using a spectral denoising approach. In this paper, we use Wiener principle [2]. The denoised speech is obtained as follows

$$|\tilde{\mathbf{S}}(m,f)|^2 = \mathbf{H}_{opt}(m,f)|\mathbf{Y}(m,f)|^2, \tag{3}$$

where $\mathbf{H}_{opt}(m,f)$ is the denoising filter, given by

$$\mathbf{H}_{opt}(m,f) = \frac{SNR_{prio}(m,f)}{1+SNR_{prio}(m,f)}, \tag{4}$$

where $SNR_{prio}(m,f)$ is the *a priori* Signal to Noise Ratio

$$\begin{aligned} SNR_{prio}(m,f) = \quad & (1-\alpha)\, P\left(SNR_{post}(m,f)\right) \\ & +\alpha \frac{|\mathbf{H}_{opt}(m-1,f)|\mathbf{Y}(m-1,f)|^2}{\mathbf{\Gamma}_n(m,f)}, \end{aligned} \tag{5}$$

where α is a real constant, $P(x) = \frac{1}{2}(x+|x|)$, $\mathbf{\Gamma}_n(m,f)$ is the noise power spectrum estimated during pause intervals and $SNR_{post}(m,f)$ is the *a postetiori* Signal to Noise Ratio

$$SNR_{post}(m,f) = \frac{|\mathbf{Y}(m,f)|^2}{\mathbf{\Gamma}_n(m,f)} - 1. \tag{6}$$

The proposed post-processing approach consists on reducing musical noise existing in denoised speech spectrum, yielding to an enhanced speech signal spectrum denoted $|\hat{\mathbf{S}}(m,f)|^2$. The temporal domain enhanced speech is obtained with the following relationship

$$\hat{s}(k) = IFFT\left[|\hat{\mathbf{S}}(m,f)|.e^{j\arg(\mathbf{Y}(m,f))}\right]. \tag{7}$$

2.2. Enhancement scheme outline

The block diagram of the enhancement technique is shown in Fig. 1. The different steps are described as follows.
• Wiener denoising algorithm is applied.
• In order to detect musical noise, we need an audible threshold of these components. The noise masking threshold (denoted NMT in scheme 1) is calculated for both noisy speech and the denoised one.
• A musical noise detector is used. For each frequency, it gives a boolean flag M which indicates the presence or not of musical noise.
• The musical noise reductor receives on its inputs the information of denoised signal spectrum, the two NMT, the two boolean flags VAD (described below) and M. It gives on its output a first version of the enhanced speech.
• Many precautions must be taken before restoring an enhanced speech. This is the task of the correction module.
It is important to note that a voice activity detector (VAD) is used for many purposes : to estimate power spectral density of noise needed for Wiener filtering, to detect and to suppress musical noise differently, according if it during speech or pause.

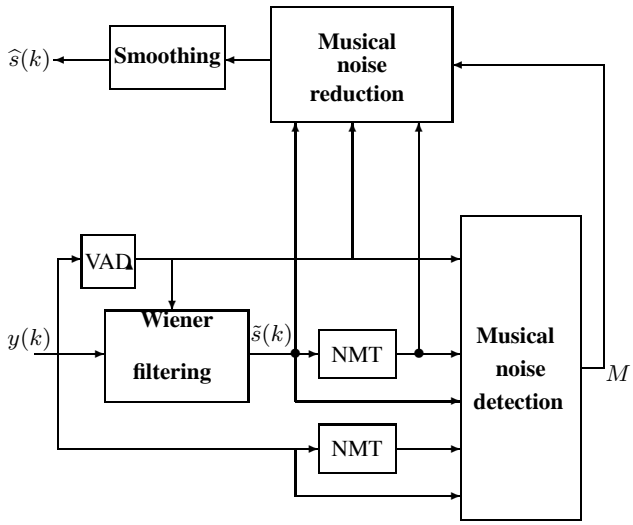

Fig. 1. Block diagram of the enhancement technique.

3. MUSICAL NOISE DETECTION

In order to detect audible musical noise, we use perceptual properties: a masked signal is made inaudible by the masker if the spectral magnitude of masked signal is below the perceptual masking threshold. In our case, the musical noise is made audible because its spectral magnitude is over a specified masking threshold. The masker is the clean speech signal. Since it is unknown, it must be estimated.
Globally, there are three steps in detecting musical noise : noise masking threshold calculation, tonal components detection in both noisy speech and denoised speech and musical noise selection.

3.1. Noise masking threshold calculation

In other applications, such as audio coding [7], using masking threshold, we distinguish between two situations : "tone masking noise" and "noise masking tone". In our context of musical noise detection, we consider only the situation of "noise masking tone". In fact, the musical noise is a tone signal which is audible during noise components of speech. We think then that the masking threshold must be computed only from noise speech components. We call it "Noise Masking Threshold" (NMT).

Table 1. Shifting constants for musical noise detection and correction constants for musical noise reduction.

	Frequency band (KHz)	$[0,1]$	$[1,2]$	$[2,3]$	$[3,4]$
Speech	$\tau(f)$ (dB)	0.5	1	2	2
	$\mathcal{C}(f)$ (dB)	0.5	2	5	10
Pause	$\tau(f)$ (dB)	1.5	2.5	3.5	4.5
	$\mathcal{C}(f)$ (dB)	10	10	10	10

The NMT is calculated according to principle used for MPEG audio coding [7]. However, some modifications are applied : use of only noise masking components, determination of masking threshold for each critical band and considering the maximum of the masking threshold in each critical band. This last idea is justified as follows. In audio coding, the minimum masking level is used in order to use minimum number of bits. In musical noise detection, maximum masking level is used in order to avoid overestimation of musical noise, for example by considering small tonal peaks due to speech or residual background noise as musical noise. This precaution is confirmed by experiments carried during this work.

3.2. Tonal components detection

The used approach for tonal components detection is inspired from [8], whose principle is described as follows. The power spectrum and noise masking threshold of both noisy speech and denoised speech are calculated. Components above NMT in noisy speech are marked as tonal and belong to speech components. Components above NMT in denoised speech are marked as tonal and belong to either speech components or musical noise components. Hence, musical noise candidates can be detected, they are the marked tonal components appearing in denoised speech and not appearing in noisy speech.

We improved the basic approach in order to avoid false detection by applying shifting rules for each frequency

$$M(m,f) = 1 \text{ if } \begin{cases} |\tilde{\mathbf{S}}(m,f)|^2 \geq NMT_{\tilde{s}}(m,f) - \tau(f) \\ \text{and} \\ |\mathbf{Y}(m,f)|^2 < NMT_y(m,f) + \tau(f). \end{cases} \quad (8)$$

The terms $\tau(f)$ are chosen empirically, they depend on the nature of processed frame (speech or noise) and the processed frequency band. After intensive tests, we retain the values resumed in Tab. 1.

It is important to note that the idea of shifting is well used in other applications, such as watermarking, in order to be sure that the masked signal is inaudible (see [9]).

3.3. Musical noise selection

In order to avoid considering isolated peaks of small bandwidth (due to tonal speech badly detected) as musical noise candidates, we carry some experiments to characterize minimum musical noise bandwidth. We notice that musical noise candidates must have at least 50 Hz bandwidth. Furthermore, instead of limiting detection to musical noise components having a bandwidth of 300 Hz approximately, we propose to extend the bandwidth. We hence consider both musical noise and other distortions of larger bands.

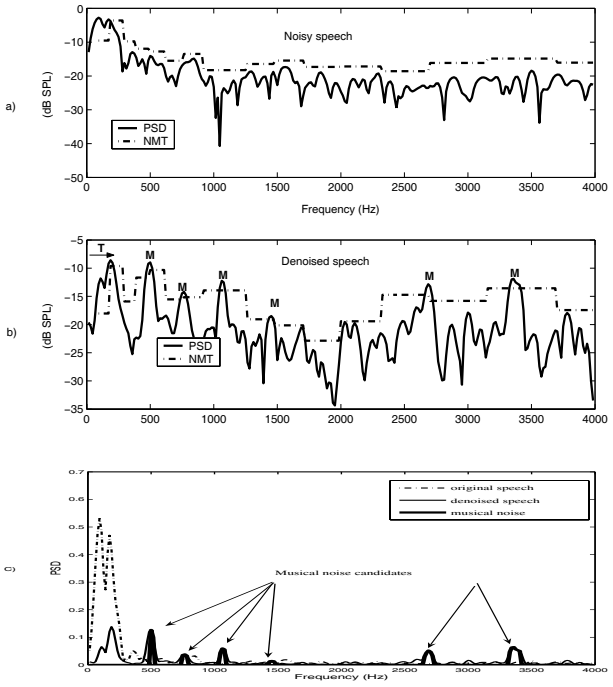

Fig. 2. Illustration of musical noise detection.

The principle of musical noise detection is resumed in Fig. 2, where Fig. 2.a (resp. Fig. 2.b) shows the power spectrum (denoted PSD) of a noisy speech frame (resp. denoised speech frame) and their related noise masking threshold. Tonal speech components are denoted by 'T' and musical noise components are denoted by 'M'. Fig. 2.c shows the original clean speech frame power spectrum, the denoised one and the detected musical noise components (in solid lines). This figure shows that in fact, tonal regions appearing in denoised speech and not appearing in clean speech are well detected.

4. MUSICAL NOISE REDUCTION

The main idea for musical noise reduction consists on shifting down the power spectrum of detected musical components under the denoised speech noise making threshold. Hence, they will be inaudible. However, different adjustments are carried to render the approach reliable. They are described in next subsections.

4.1. Correction

In order to be sure that musical noise is well reduced, we propose to include an correction term permitting to shift down sufficiently the power spectrum. The estimated power spectrum of corrected speech is written

$$|\bar{\mathbf{S}}(m,f)|^2 = \begin{cases} NMT_{\bar{s}}(m,f) - \mathcal{C}(f) & \text{if } M(m,f)=1 \\ |\tilde{\mathbf{S}}(m,f)|^2 & \text{otherwise} \end{cases} \quad (9)$$

where $\mathcal{C}(f)$ is the correction term chosen according to subjective listening tests. Tab. 1 resumes obtained values for speech and pause frames. The attenuation term is small for low frequency and important for high frequency. During pause, it is constant since distortion and musical noise appear in the same way in all frequency band.

4.2. Smoothing

Once, the power spectrum of corrected speech is estimated, we notice that novel bursts appear in the spectrum. In fact, because of the shifting down, the musical noise appearing as a peak on the spectrum will constitutes a valley. Just before and after musical noise components, the boundaries are kept untouched and they form new unwanted peaks on the modified spectrum. (see Fig. 3 for illustration). We propose to avoid them using median filters, well known to be powerful for impulsive noise reduction. In this paper, we choose a neighborhood of size three, which constitutes a good compromise between bursts reduction and moderate spectrum smoothing in regions around musical noise. The expression of the power spectrum of the enhanced speech $\hat{s}(k)$ is written

$$|\hat{\mathbf{S}}(m,f)|^2 = \begin{cases} \text{med}\left(|\bar{\mathbf{S}}(m,f-\Delta f)|^2, |\bar{\mathbf{S}}(m,f)|^2, |\bar{\mathbf{S}}(m,f+\Delta f)|^2\right) \\ \text{if } M(m,f)=0 \text{ and } M(m,f+\Delta f)=1 \\ \text{or } M(m,f)=0 \text{ and } M(m,f-\Delta f)=1 \\ |\bar{\mathbf{S}}(m,f)|^2 \quad \text{otherwise} \end{cases} \quad (10)$$

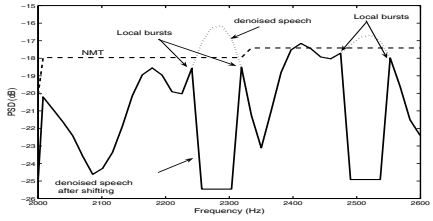

Fig. 3. Illustration of local bursts.

5. EVALUATION AND RESULTS

The proposed technique is evaluated using temporal, spectral and perceptual criteria. In term of quantitative temporal criteria, we use the Signal to Noise Ratio SNR and the Segmental Signal to Noise Ratio $SNRseg$ [10]. In term of spectral criteria, we use the cepstral distance CEP and the spectrograms [10]. In term of perceptual criteria, we use the Weighted Slope Spectral distance (WSS) [11] and the Modified Bark Spectral Distance ($MBSD$) [12].

In our simulations, we used a clean speech which was artificially corrupted with white Gaussian noise and Monte-Carlo simulations over 100 runs were performed. Table 2 compares the performances of the classical denoising scheme based on Wiener filtering and the proposed post-processing approach for different values of SNR. The prefix in (resp. out) in Tab. 2 is related to the distance between clean speech and the noisy one (resp. denoised one), while the prefix enh is related to the distance between clean speech and the enhanced one using our approach.

Tab. 2 permits the following interpretations.

• The proposed approach leads to better denoising quality for all criteria except the SNR. This fact is predictable since we used spectral and perceptual considerations to enhance speech. Moreover, the criterion has a small correlation with listening tests ($\rho = 0.38$).

• The improvement is well noticeable for spectral and perceptual criteria which have good correlation with listening tests.

• Although the improvement in term of WSS, the denoising approach didn't reach the original quality (between clean and noisy). However, it is well improved when compared to that of denoised speech using Wiener.

Table 2. Performances the speech enhancement system.

SNR_{in}	-10	-6	0	6	11
SNR_{out}	**2.8**	**3.85**	**9.18**	**21.45**	**42.38**
SNR_{enh}	**2.8**	3.65	8.06	16.01	24.9
$SNRseg_{in}$	-9.06	-7.5	-5.02	-1.9	1.02
$SNRseg_{out}$	-3.41	-1.8	0.34	3.64	6.5
$SNRseg_{out}$	**-3.18**	**-1.27**	**1.74**	**4.44**	**6.14**
CEP_{in}	12.34	9.72	7.06	4.88	3.44
CEP_{out}	8.48	7.45	7.19	7.21	6.88
CEP_{enh}	**1.33**	**1.05**	**1.06**	**1.27**	**1.45**
WSS_{in}	**662**	**593**	**509.32**	**449**	**401**
WSS_{out}	1089	926	814.39	748	697
WSS_{enh}	832	712	618.42	556	507
$MBSD_{in}$	1.19	0.837	0.55	0.34	0.214
$MBSD_{out}$	0.125	0.078	0.043	0.021	0.013
$MBSD_{enh}$	**0.069**	**0.038**	**0.020**	**0.016**	**0.018**

Spectrograms are considered in Fig. 4. The noisy speech signal is a speech sequence corrupted by a Gaussian noise whose $SNR = 10$ dB. It is worth pointing out that the denoised signal by the classical method is affected by a musical noise (isolated points randomly distributed in time and frequency). The amount of such noise is dramatically reduced by the proposed approach.

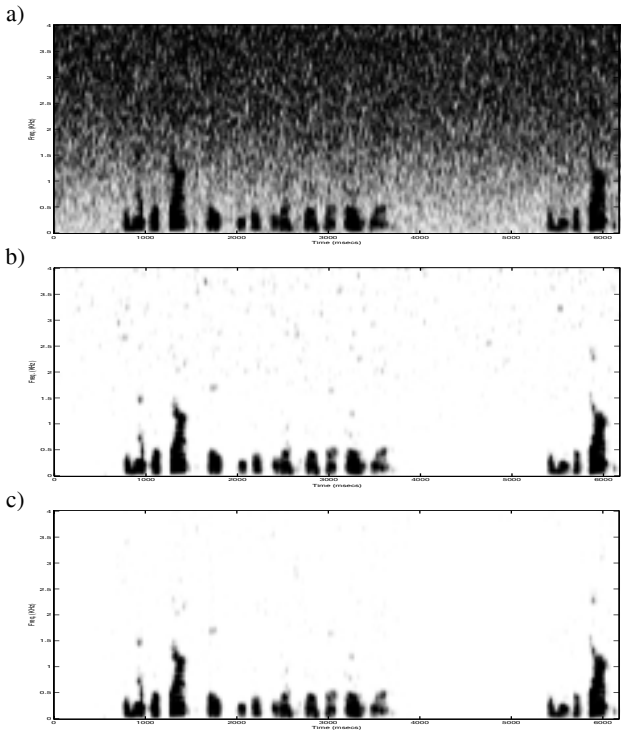

Fig. 4. Spectrograms of noisy speech (a), denoised speech (b), enhanced speech (c).

6. CONSLUSION

In this paper, a new post-processing method for reducing musical noise imposed by Wiener denoising approach is proposed. The method makes use of perceptual noise masking threshold to detect musical noise and then to reduce it. Simulation results show the improvement of this method in term of temporal, spectral and perceptual criteria. Further investigations should consider the improvement of NMT estimation and musical noise detection in order to reduce the small quantity of residual musical noise left in the enhanced speech.

7. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, April 1979.

[2] Y. Ephraim and D. Mallah, "Speech enhancement using optimal non-linear spectral amplitude estimation," on *Proceedings Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, IEEE, 1983, pp. 1118-1121.

[3] D. E. Tsoukalas, J. Mourjopoulos and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 6, pp. 497-514, November 1997.

[4] A. Akbari-Azrani, R. Le Bouquin Jannes and G. Faucon, "Optimzing speech enhancement by exploting masking properties of the human ear," on *Proceedings Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, IEEE, 1995, pp. 800-803.

[5] L. Lin, W. H. Holmes and E. Ambikairajah, "Speech denoising using perceptual modification of Wiener filtering," *IEE Electronic Letters*, vol. 38, no. 23, pp. 1486-1487, November 2002.

[6] N. Virag, "Single channel speech enhancement based on masking properties of human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, February 1999.

[7] ISO/IEC 11172-3, "International standard 11172-3:1993, Information technology, coding and moving pictures and associated audio for storage media at up to about 1.5 Mbits, Part 3, Audio," 1993.

[8] T. Haulick, K. Linhard and P. Schrogmeier, "Residual noise suppression using psychoacoustic criteria," *Proc. Eurospspeech*, 1997, pp. 1395-1398.

[9] L. Boney, A. H. Tewfik and K. N. Hamdy, "Digital watermarks for audio signal," on *Proceedings of Multimedia*, IEEE, 1996, pp. 473-479.

[10] S. R. Quanckenbush, T. P. Barnwell III and M. A. Clements, *Objective measures of speech quality*, Prentice Hall, Englewood Cliffs, 1988.

[11] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra : a first step," on *Proceedings Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, IEEE, 1982, pp. 1278-1281.

[12] W. Yan, M. Dixon and R. Yantorno, "A modified bark spectral distorsion measure which uses noise masking threshold," on *Proceedings of the Speech Coding Worshop IEEE*, 1997, pp. 55-56.