Psychoacoustic filtering for noisy speech enhancement

Sana Alaya
Signal, Image and Information Technology Laboratory, National Engineering School of Tunis, Tunis, Tunisia

Novlène Zoghlami
Signal, Image and Information Technology Laboratory, National Engineering School of Tunis, Tunis, Tunisia

Zied Lachiri
Signal, Image and Information Technology Laboratory, National Engineering School of Tunis, Tunis, Tunisia

Summary

A new denoising approach is introduced in this paper. It is based on the fact that denoising may be performed by mimicking the human ear function in order to improve the psychoacoustics appearance of speech signal. In the proposed method, the speech signal is decomposed by using a gammatone filterbank in accordance with ERB scale. Spectral attenuation filtering is then applied in each sub-band which is based on continuous noise estimation. In the output of the spectral attenuation filter, masking threshold is calculated by using the Johnston model and then inserted in the Psychoacoustics gain filter. Evaluation tests are realized using objective and subjective criterion such as Perceptual Evaluation of Speech Quality (PESQ) for the objectives scores and mean the quality rating of signal distortion (SIG), noise distortion (BAK) and overall quality (OVRL) for the subjective scores. The results show that our method gives best global quality of the enhanced speech signal while maximizing noise elimination and minimizing distortion.

1. Introduction

Noise reduction techniques are constrained by a compromise between robust noise reduction, minimizing distortions level and musical noise. Modeling the human auditory system helps effectively the problem of speech enhancement and more specifically in noise reduction to obtain a good quality and intelligibility signal [1]. Many techniques have been developed to reach this objective [1-2-3-4-5]. The proposed method processes the signal as follows. First, the speech signal is decomposed by using a gammatone filterbank in accordance with ERB scale [6]. This latter is characterized by her nonlinear frequency decomposition imitating the human ear decomposition. Second, spectral attenuation filtering [7] is applied in each sub-band which is based on continuous noise estimation. Third, masking threshold is calculated by using the Johnston model [8] in the output of the spectral attenuation filter. Fourth, the calculated threshold is then inserted in the Psychoacoustics gain filter. Finally, the Psychoacoustics filter will be applied on each output sub-band of the spectral attenuation filter.

This document is arranged as follows: Section 2 describes the method based on the psychoacoustics filtering. Section 3 presents the results of objective and subjective evaluation tests.

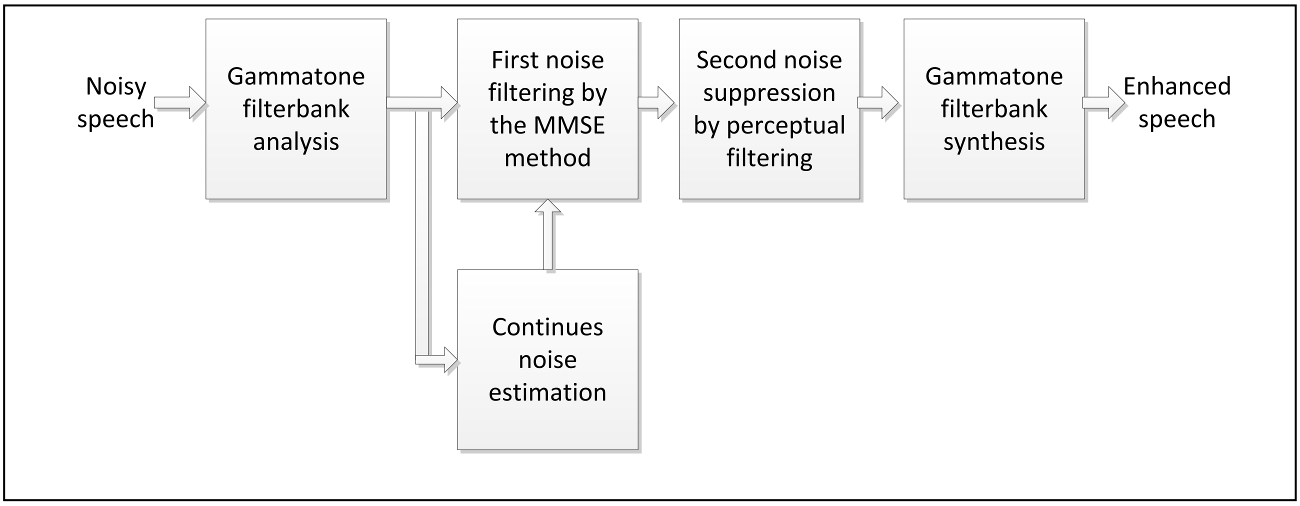

Noisy speech
Gammatone filterbank analysis
First noise filtering by the MMSE method
Second noise suppression by perceptual filtering
Gammatone filterbank synthesis
Enhanced speech
Continues noise estimation

Figure 1 : Principle of the psychoacoustic denoising speech signals method.

2. Psychoacoustic processing of noisy speech enhancement

Non-uniform decomposition of the speech signal $y(t) = s(t) + n(t)$ is reflected in using Gammatone filter banks where $s(t)$ is a clean speech, $n(t)$ the noise and $t= 0,1,\dots,M-1$ is a time vector. The Gammatone impulse response is defined as follows:

$$g(t) = At^{n-1} \exp(-2\pi b\, \mathrm{ERB}(f) t) \cos(2\pi f t + \varphi) \qquad (1)$$

Where (t>0), A the amplitude, n and b two invariant parameters that define the distribution. f is the asymptotic modulation frequency, φ the initial phase. $\ln t$ is the natural logarithm of the time vector. $\mathrm{ERB}(f)$ is a rectangular width of filter at frequency f [9] where $\mathrm{ERB}(f) = 21.4 \ln(0.00437f + 1)$ en Hz.
The noisy signal $y(t)$ is decomposed into sub-signals $y_i(t)$:

$$y_i(t) = y(t) * g_i(t) \qquad (2)$$

$g_i(t)$ is the impulse response of the ith Gammatone band.
Following the non-linear decomposition of the speech signal, the generalized spectral subtraction gain is applied to each sub-band followed by a psychoacoustic gain. The main idea of the psychoacoustic filter is to filter the audible musical noise which can be created following the enhancement by standard spectral subtraction. To determine whether a musical sound is audible, we compare it to the Johnston masking curve. If the noise is above the threshold, it is considered audible. Otherwise, it is inaudible. The auditory masking threshold calculated by the Johnston model, is denoted $JT_{i,k}(f)$. This threshold is calculated on the base of the output of the generalized spectral subtraction. Subsequently, it is inserted in the psychoacoustic gain $PG_{i,k}(f)$ [3] defined by the following equation:

$$PG_{i,k}(f) = \frac{|\tilde{S}_{i,k}(f)|^2}{(|\tilde{S}_{i,k}(f)|^2 + \max((\gamma_{i,k}(f) - JT_{i,k}(f)),0)} \qquad (3)$$

$\gamma_{i,k}(f)$ is the estimated power spectrum noise calculated by IMCRA noise estimation technique [10], f is the frequency, i the index of the sub signal and k the index of the frame.
$|\tilde{S}_{i,k}(f)|^2$ is the estimated enhanced power spectrum signal by the generalized spectral subtraction method $GSS_{i,k}(f)$ given by the equation:

$$\begin{cases} |\tilde{S}_{i,k}(f)|^2 = |Y_{i,k}(f)|^2 - \alpha|\gamma_{i,k}(f)|^2 \text{ si } |Y_{i,k}(f)|^2 > (\alpha + \beta)|\gamma_{i,k}(f)|^2 \\ |\tilde{S}_{i,k}(f)|^2 = \beta|\gamma_{i,k}(f)|^2, \text{sinon} \end{cases} \qquad (4)$$

With (α≥1) the subtraction factor that determines the amount of noise to remove. (0<β≤1) determines the minimum noise level, non-interfering, which can be present in the enhanced signal.
When $\gamma_{i,k}(f) < T_{i,k}(f)$, estimated noise after applying the spectral subtraction is inaudible. The gain is nearly equal to 1 $PG_{i,k}(f) \approx 1$ in order to minimize the risk of filtering masks sounds which may lead to the creation of residual noise. Psychoacoustic filtering, in this situation, is limited to the application of the spectral subtraction $GSS_{i,k}(f)$. The gain $PG_{i,k}(f)$ essentially depends on the value of the Johnston threshold $JT_{i,k}(f)$. Two cases arise; where $\gamma_{i,k}(f) > T_{i,k}(f)$ the estimated noise is audible by the human ear. We are therefore obliged to apply the $PG_{i,k}(f)$

following spectral subtraction in order to improve the perceptual quality of the enhanced signal. The final enhanced signal is obtained by applying the synthesis filter bank.
The block diagram in Figure 1 shows the psychoacoustic proposed denoising method.

3. Evaluation

The proposed method is tested by TIMIT speech signals sampled at 16kHz. These signals are degraded by environmental noise namely car noise and street noise at various SNR: 0, 5, 10 and 15dB. Windowing was carried out using the Hamming window of length 512 samples with an overlap of 50%. The proposed psychoacoustic method is compared with the classic spectral subtraction MMSE.
For evaluation we performed by calculating the signal to noise ratio (SNR) [11], by estimating the perceptual quality of the signal (PESQ) [12], by calculating the Itakura Saito distance (IS) [13] and computing the SIG BAK and OVRL values [14].

Figure 2 shows the results of the SNR.

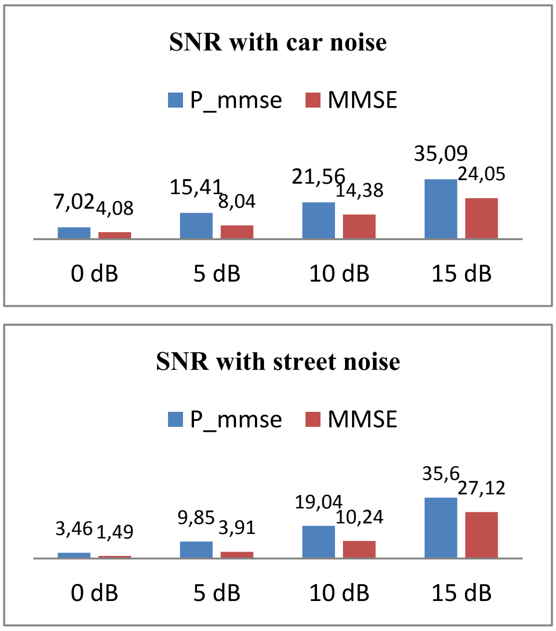


Figure 2 : SNR values of the proposed method P_MMSE compared to the SNR values of the classic spectral subtraction method MMSE

It is clear that the method makes significant improvements in enhanced signal quality by the proposed method. At 10 dB degradation, we obtained for the car noise 21.56 against 14.38 with the classic spectral subtraction.

Figure 3 presents the results found by the IS method.

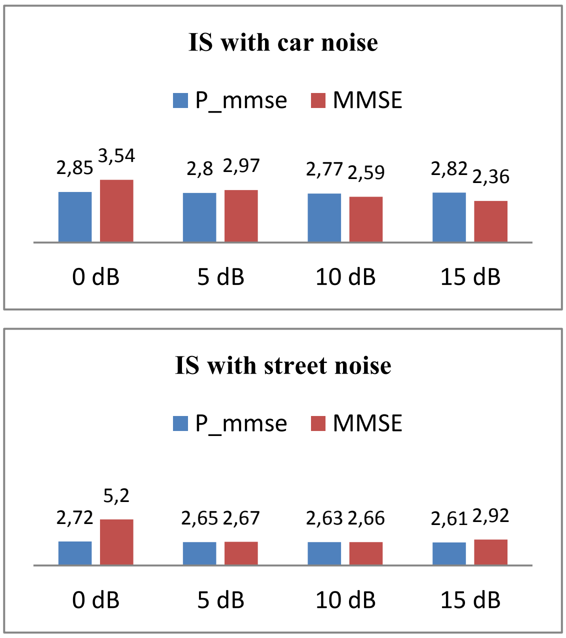
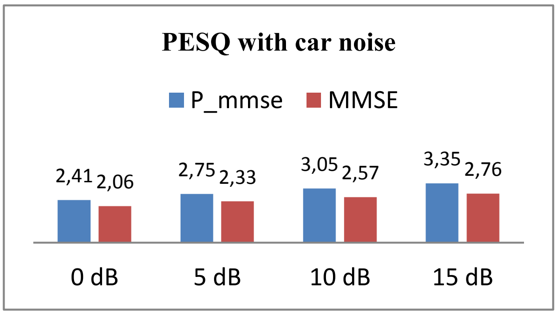


Figure 3 : The IS values of the proposed psychoacoustic method P_MMSE compared to the IS values of the classic spectral subtraction MMSE.

We find that the proposed psychoacoustic method minimizes distortions in the enhanced signal especially when signals are very noisy. Compared with the classic spectral subtraction, we find that in 0 dB of degradation psychoacoustic method provides only 2.72 distortions against 5.2 for classical spectral subtraction with street noise.

Figure 4 shows the results of the PESQ score given by the proposed psychoacoustic method P_MMSE compared with the classic spectral subtraction method MMSE.



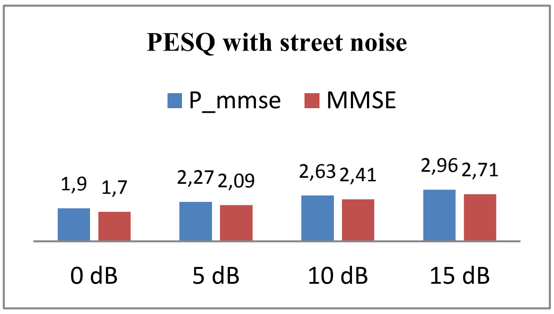
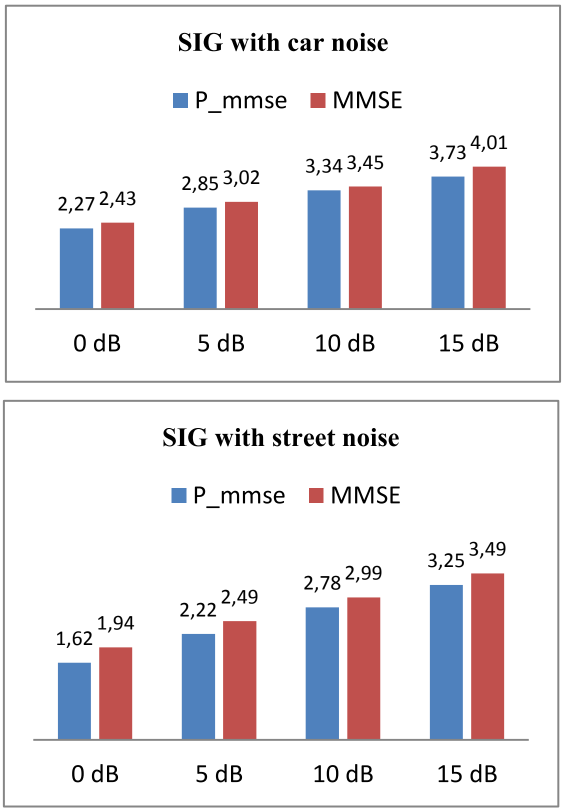


Figure 4 : The PESQ values of the proposed P_MMSE method compared with the PESQ values of the classic spectral subtraction method MMSE.

Based on the PESQ results found, it is concluded that the proposed method improves the perceptual quality of the signal significantly. At 15 dB degradation, we obtain for the car noise 3.35 against 2.76 with the classic spectral subtraction.

Figure 5 shows the results of the SIG BAK and OVRL scores given by the proposed psychoacoustic method P_MMSE compared with the spectral attenuation method MMSE



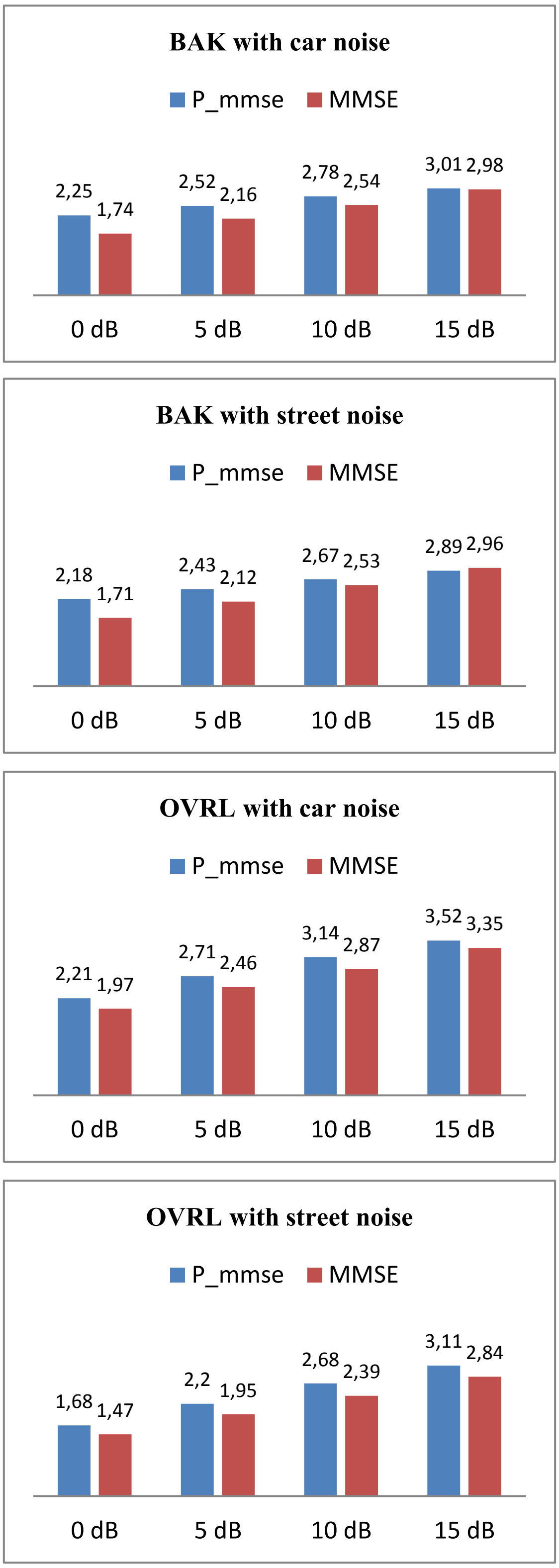


Figure 5 : The SIG BAK and OVRL values of the proposed P_MMSE method compared with the SIG BAK and OVRL values of the classic spectral subtraction method MMSE.

Based on the results of SIG BAK and OVRL found, it is concluded that the proposed method, significantly, improves the perceptual quality of the signal without creating distortions. For car noise at 0 dB, we obtain SIG=2.27 BAK=2.25 OVRL=2.21 for the proposed psychoacoustic method against SIG=2.43 BAK=1.74 OVRL=1.97 for the classic spectral subtraction.

Based on the SNR results, IS distance and PESQ value found, it is concluded that the proposed psychoacoustic method contributes to the improvement of quality and intelligibility of the enhanced signal especially with very noisy signal.

4. Conclusion

A new psychoacoustic enhancement of speech signals has been proposed in this paper in order to eliminate noise without creating distortion and residual noise. The method is based on integrating psychoacoustic gammatone filter that will ensure sub-band speech decomposition similar to that performed by the human ear. Following this decomposition, each sub-band will be treated independently by psychoacoustic spectral subtraction method. Comparing this method with the classic spectral subtraction, it is found that combining gammatone filter with the classic spectral subtraction and the Johnston model improves the quality and intelligibility of the noisy signal at different degradation levels.

References

[1] P. Loizou : Speech Enhancement: Theory and Practice. CRC Press, FL: Boca Raton, 2013.

[2] N. Zoghlami, Z. Lachiri, and N. Ellouze: Noise reduction based on perceptual speech analysis. Proc. EURONOISE 2009. 26-28.

[3] A. Amehraye, D. Pastor and A. Tamtaoui : Perceptual improvement of Wiener filtering. ICASSP 2008, 2081-2084.

[4] N. Virag : Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech and Audio Processing (1999) 126- 137.

[5] C.H. Taal, R.C. Hendriks and R. Heusdens : A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. ICASSP 2012, 4061 – 4064.

[6] V. Hohmann : Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica 88(3) (2002) 433-442.

[7] Y. Ephraim and D. Malah : Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process 32(6) (1984) 1109-1121.

[8] J. D. Johnston : Transform coding of audio signals using perceptual noise criteria. IEEE Jour. Selected Areas Commun 6 (1988) 314–323.

[9] B. C. J. Moore and B. R. Glasberg : A revision of Zwicker's loudness model. Acta Acustica 82 (1996) 335-345.

[10] I. Cohen : Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process (2003) 466-475.

[11] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements : Objective Measures of Speech Quality. Prentice Hall Advanced Reference Series, Englewood Cliffs, NJ, 1988.

[12] ITU-T recommendation P.862 : Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, 2000.

[13] D. O'Shaughnessy : Speech Communication: Human and Machine. Addison-Wesley, 1987.

[14] ITU-T recommendation P.835 : Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. International Telecommunication Union, 2003.