Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2012, Article ID 282019, 12 pages
doi:10.1155/2012/282019

Research Article

Application of Perceptual Filtering Models to Noisy Speech Signals Enhancement

Novlene Zoghlami[1] and Zied Lachiri[1,2]

[1] *LRSITI, Département Génie Electrique, Ecole Nationale des Ingénieurs de Tunis, BP 37, 1002 Le Belvédère, Tunisia*
[2] *Département de Génie Physique et Instrumentations, Institut National des Sciences Appliquées et de Technologies, Centre Urbain Nord, BP 676, 1080 Tunis Cedex, Tunisia*

Correspondence should be addressed to Novlene Zoghlami, novlene_zoglami@yahoo.fr

Received 20 March 2012; Revised 24 May 2012; Accepted 30 May 2012

Academic Editor: Raj Senani

This paper describes a new speech enhancement approach using perceptually based noise reduction. The proposed approach is based on the application of two perceptual filtering models to noisy speech signals: the gammatone and the gammachirp filter banks with nonlinear resolution according to the equivalent rectangular bandwidth (ERB) scale. The perceptual filtering gives a number of subbands that are individually spectral weighted and modified according to two different noise suppression rules. The importance of an accurate noise estimate is related to the reduction of the musical noise artifacts in the processed speech that appears after classic subtractive process. In this context, we use continuous noise estimation algorithms. The performance of the proposed approach is evaluated on speech signals corrupted by real-world noises. Using objective tests based on the perceptual quality PESQ score and the quality rating of signal distortion (SIG), noise distortion (BAK) and overall quality (OVRL), and subjective test based on the quality rating of automatic speech recognition (ASR), we demonstrate that our speech enhancement approach using filter banks modeling the human auditory system outperforms the conventional spectral modification algorithms to improve quality and intelligibility of the enhanced speech signal.

1. Introduction

The high quality sound of talking speech in real environment is very important for automatic speech processing systems and human- machine interfaces. However, the performance of these systems can be affected by background noise. Thus, there is a strong need to resolve this problem and improve the performance of these applications in high level noise environment by applying effective speech enhancement techniques able to suppress the undesirable noise. These techniques are concerned with improving some perceptual aspect, the quality and intelligibility of degraded speech. In a broad context, many methods are developed in order to remove the background noise while retaining speech intelligibility based on short time spectral estimation of the clean speech. These methods are able to reduce the noise and improve the quality, but at the expanse of introducing speech distortion which results in loss of intelligibility. Hence, the main challenge in designing effective speech enhancement algorithms is to suppress the noise without introducing any perceptible speech distortion. The spectral modification methods are historically one of the first algorithms proposed for noise reduction, especially the generalized spectral subtraction is the most popular technique [1]. This method is able to reduce the background noise using estimation of the short-time spectral magnitude of the speech signal by subtracting the noise estimation from the noisy speech. The spectral subtraction technique offers a high flexibility and simplicity in implementation. However, it needs to be improved since its major drawback, the introduction in the enhanced speech of residual noise called "musical noise" with unnatural structure, is composed of tones at random frequencies. The unnatural structure of the musical noise is perceived as nonstationary noise artifacts that depend on the

time and frequency changes of the noise, on one side and on the way that the human auditory system perceives these artifacts, on the other side. The minimum-mean-square-error-based-noise reduction proposed by Ephraim and Malah subtraction rule [2] exploits the average spectral estimation of the speech signal based on a prior knowledge of the noise variance, in the goal to mask and reduce the residual noise. In [3–8], the noise is reduced based on subtractive type algorithms according to a multibands and nonlinear spectral process. In [9–11], the authors exploit the human perceptual masking proprieties to improve the quality and intelligibility of the speech signal without introducing speech distortion. The difficulty with these approaches is that an estimate of the clean speech itself is necessary in order to calculate the masking threshold.

The solution proposed in this paper works towards achieving a high noise reduction with efficient residual noise elimination, at the same time, to preserve speech components. This is done by meeting several requirements to the speech analysis/synthesis system based on the knowledge of human perception proprieties. So it is proposed to adapt the spectral modification algorithms to a multibands analysis using human perceptual filter banks models according critical band concept and nonlinear frequency resolution. This allows to find the best tradeoff between the amount of noise reduction, the speech distortion and the level of musical noise in a perceptual view, and to overcome the limitation of spectral modification algorithms for speech enhancement in real-world listening situation where the background noise level and characteristics are constantly changing.

The paper is organized as follows: in Section 2, the principle of common spectral modification algorithms reviewed in the speech enhancement literature is described. In Section 3, the proposed enhancement approach is presented. Finally, an objective and subjective evaluation is performed in Section 4.

2. Spectral Modification Principle

The spectral modification techniques operate in the frequency domain. These methods are widely used for the enhancement of speech signals, which are corrupted by additive noise with constant or slowly varying spectral characteristics. The basic idea is to manipulate the magnitude of the noisy speech spectrum using fixed and uniform spaced frequency transformation. Consider a speech signal $x(n)$ degraded by additive background noise $d(n)$, the noisy speech $y(n)$ can be expressed as

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

The signal is divided into uniform frame using an adequate analysis window and it is processed in the frequency domain. The spectral analysis and synthesis are usually performed by a discrete Fourier transform and its inverse with overlap-add technique. The noise suppression process is a multiplication of the short-time spectral magnitude of the noisy speech $|Y(p,w)|$ by a gain function $G(p,w)$,

$$\left|\hat{X}(p,w)\right| = G(p,w) \quad \text{with } 0 \le G(p,w) \le 1. \quad (2)$$

With p is the frame index and w is the frequency index. $|\hat{X}(p,w)|$ is the magnitude spectrum of the processed speech. Each gain function corresponds to a given noise suppression rule that changes depending to the characteristics of the noisy signal spectrum and the estimated noise spectrum.

3. Using Perceptual Filtering Models for Speech Enhancement

The spectral modification techniques performed in noise reduction using short time spectral analysis based on fixed and uniform speech decomposition. This processing, however, creates small isolated fluctuations in the spectrum occurring at random frequency locations in each frame, converted in the time domain, these fluctuations sound similar to tones with frequency peaks that change randomly from frame to frame. These artifacts described as residual noise consists of tonal remnant noise component significantly disagreeable to the ear. Focusing on the perceptual processing based on how human listeners process tones and bands of noise, it is possible to suppress the background noise and completely attenuate the random peaks in the structure of musical noise. The human auditory system may be sensitive to abrupt artifacts changes and transient component in the noisy speech signal based on time-frequency analysis with a nonlinear frequency selectivity of the basilar membrane Thus, the human hearing process is modeled as a series of transformations of the acoustic signal via an array of overlapping band-pass filters known as perceptual filters. These filters occur along the basilar membrane and increase the frequency selectivity of the human ear. Hence, the speech component can be identified and the selectivity can be amplified. The idea behind this is that embedding the psychoacoustics models of human auditory system in perceptual filter banks may lead to improve intelligibility and perceptual quality of speech. Moreover, it is known that humans are capable of detecting the desired speech in noisy environment without any prior information of the noise type. Taking into account the psychoacoustic analysis and human perception properties, it is possible to make a successful speech enhancement system when we use a suitable perceptual model to obtain nonuniform filter banks representing the human ear processing and an appropriate spectral modification approach, such as the generalized spectral subtraction technique (GSS) and the minimum mean square Error (MMSE) for spectral enhancement of each nonuniform filter banks bands output.

The proposed enhancement scheme is presented in Figure 1.

Step 1. Speech decomposition via perceptual filter-bank analysis stage.

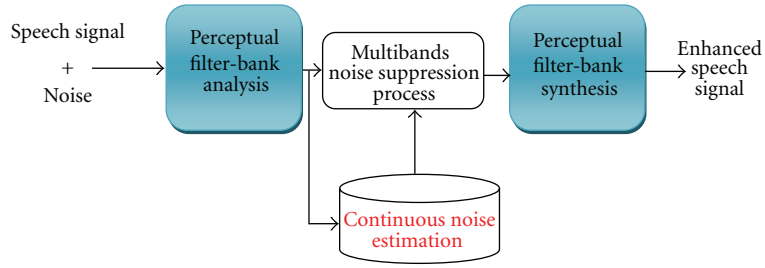

Figure 1: Proposed speech enhancement method based on perceptual filtering model.

Step 2. Speech enhancement process: multibands noise suppression process.

Step 3. Continuous noise estimation.

Step 4. Speech synthesis via perceptual filter banks synthesis stage.

3.1. Perceptual Filtering Models. The aim in perceptual modeling is to find mathematical model which represents some physiological and perceptual aspects of the human auditory system. Perceptual modeling is very useful, since the sound wave can be analyzed according to the human ear comportment, with a good mode. The simplest way to model the frequency resolution of the basilar membrane is to make analysis using filter banks. The simplest and the most realistic model is the gammatone filter banks [12], the impulsion response is based on psychoacoustics measurements, providing a more accurate approximation to the perceptual frequency response, and it is represented by a gammatone function defined in the temporal model by the following expression:

$$\mathrm{gt}(t) = At^{n-1}\exp(-2\pi bBc)\cos(2\pi f_c t + \varphi), \quad (3)$$

where A defines the magnitude normalization parameter, n is the filter order, fc is the center frequency of filters, B is filters bandwidths, and $bB(f_c)$ represents the filter envelop. The gammachirp filter bank is another perceptual model [13], it is an extension of the popular gammatone filter with an additional frequencymodulation term to produce an asymmetric amplitude spectrum. The complex impulsion response is based on psychoacoustics measurements, providing a more accurate approximation to the perceptual frequency response, and it is given in the temporal model as

$$\mathrm{gc}(t) = At^{n-1}\exp(-2\pi bBc)\cos(2\pi f_c t + c\ln t + \varphi), \quad (4)$$

where time $t > 0$, A is the amplitude, $n$and b are parameters defining the envelope of the gamma distribution, f_c is the asymptotic frequency, c is a parameter for the frequency modulation ($c = 3$), φ is the initial phase, $\ln t$ is a natural logarithm of time, and ERB (f_c) is the equivalent rectangular bandwidth of the perceptual filter at f_c.

The frequency resolution of human hearing is a complex phenomenon which depends on many factors, such as frequency, signal bandwidth, and signal level. Despite of the fact that our ear is very accurate in single frequency analysis, broadband signals are analyzed using quite sparse frequency resolution. The equivalent rectangular bandwidth (ERB) scale is an accurate way to explain the frequency resolution of human hearing with broadband signals. The expression used to convert a frequency f in Hz in its value in ERB is

$$\mathrm{ERB}(f) = 21{,}41 \cdot \log\left(\frac{4{,}37}{1000} + 1\right). \quad (5)$$

Figure 2 shows the correspondence between frequencies in Hz and its values in ERB and the frequency response of the gammatone and the gammachirp filter banks with $k = 27$ ERB bands.

3.2. Multibands Perceptual Process Using Perceptual Filter Bank. The proposed speech enhancement method is based on nonuniform decomposition of the degraded input waveform $y(n)$. The processing is done by dividing the incoming noisy speech into separate bands $y_{k,\mathrm{gt}}(n)$ that could be individually manipulated using spectral modification algorithms to achieve quality and intelligibility improvement of the overall signal. The analysis filter banks consists of 27-4th order gammatone filters and of 27-4th order gammachirp filters that cover the frequency range of the signal.

The filters bandwidth changes according the equivalent rectangular bandwidth ERB scale. The output of the k^h filter of the analysis gammatone filter banks can be expressed as

$$y_{k,\mathrm{gt}}(n) = y(n) * \mathrm{gt}_k(n), \quad (6)$$

where $\mathrm{gt}_k(n)$ is the impulse response of the kth, 4th-order gammatone filter. And the output of the k^h filter of the analysis gammachirp filter banks can be expressed as

$$y_{k,\mathrm{gc}}(n) = y(n) * \mathrm{gc}_k(n), \quad (7)$$

where $\mathrm{gc}_{k(n)}$ is the impulse response of the kth, 4th-order gammachirp filter.

The proposed speech enhancement method is based on nonuniform decomposition of the degraded input waveform $y(n)$. The processing is done by dividing the band k obtained by nonuniform gammatone decomposition $y_{k,\mathrm{gt}}(n)$ and obtained by nonuniform gammachirp decomposition $y_{k,\mathrm{gc}}(n)$ are divided into frames (10 ms–30 ms length) by multiplication with a sliding window $F(n)$. Nonuniform

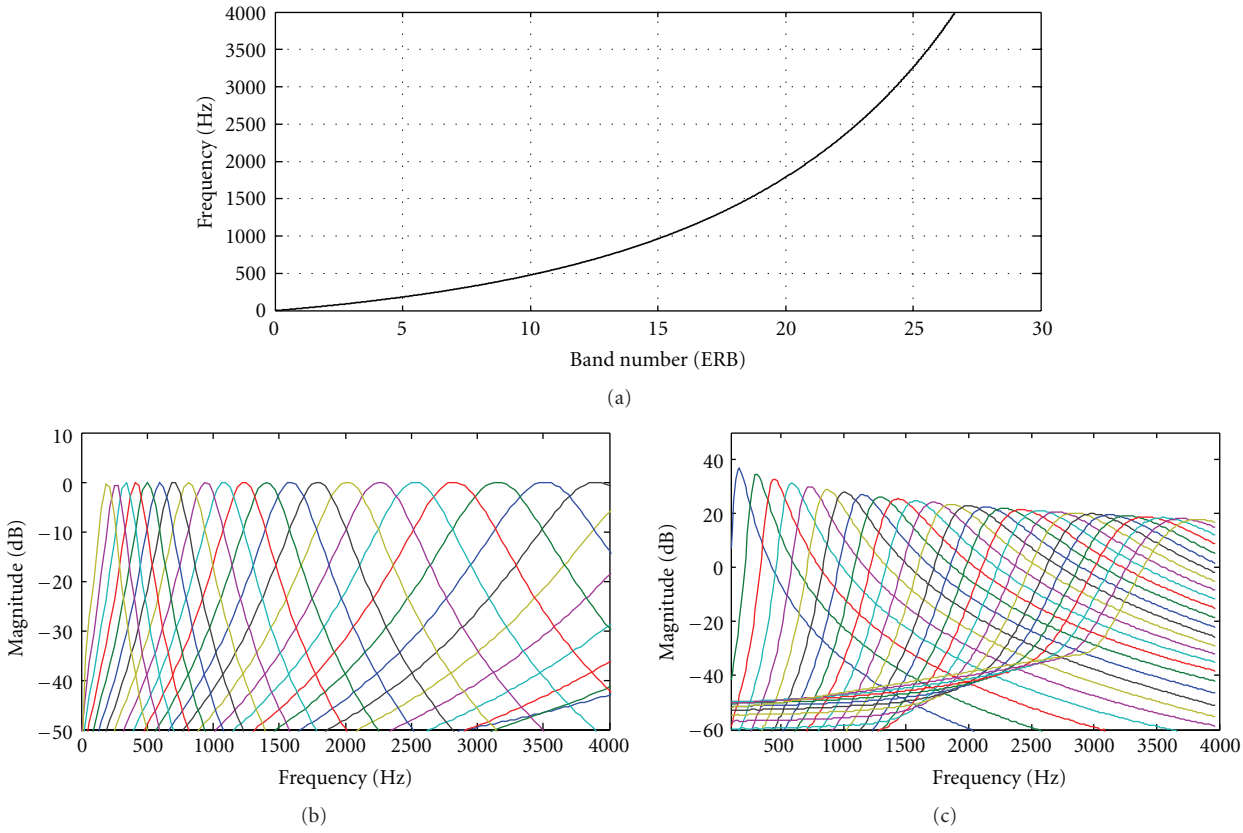

FIGURE 2: Frequency and ERB-scale correspondence (a) and the frequency response of the gammatone filter banks (b) and the gammachirp filter banks (c) with $k = 27$ ERB bands.

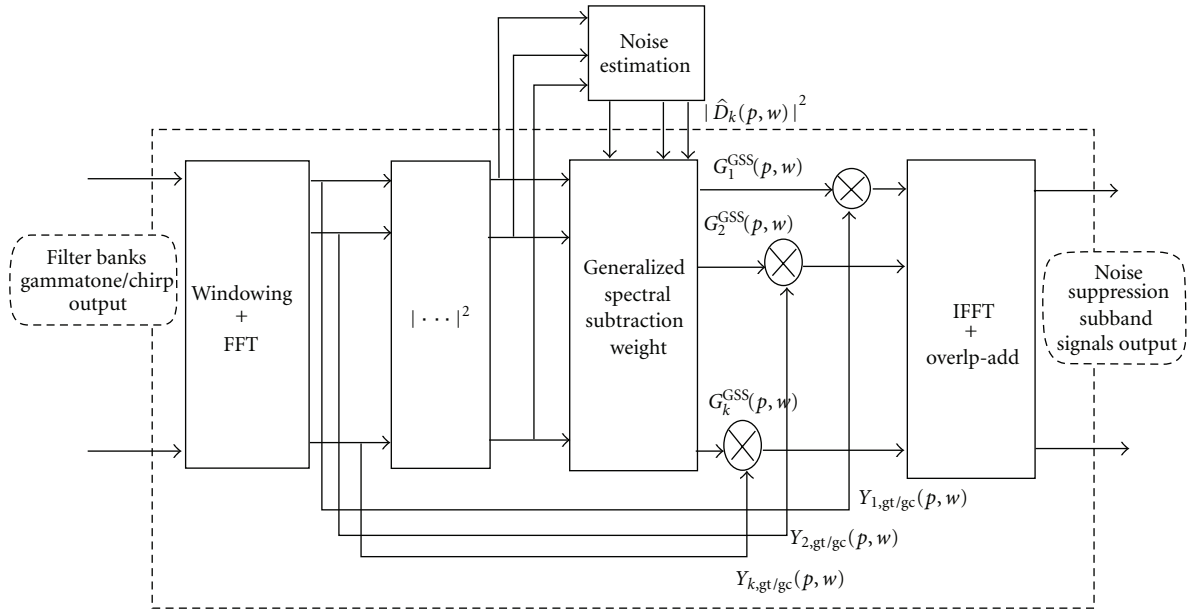

FIGURE 3: Proposed perceptual generalized spectral subtraction technique applied in a multirate system.

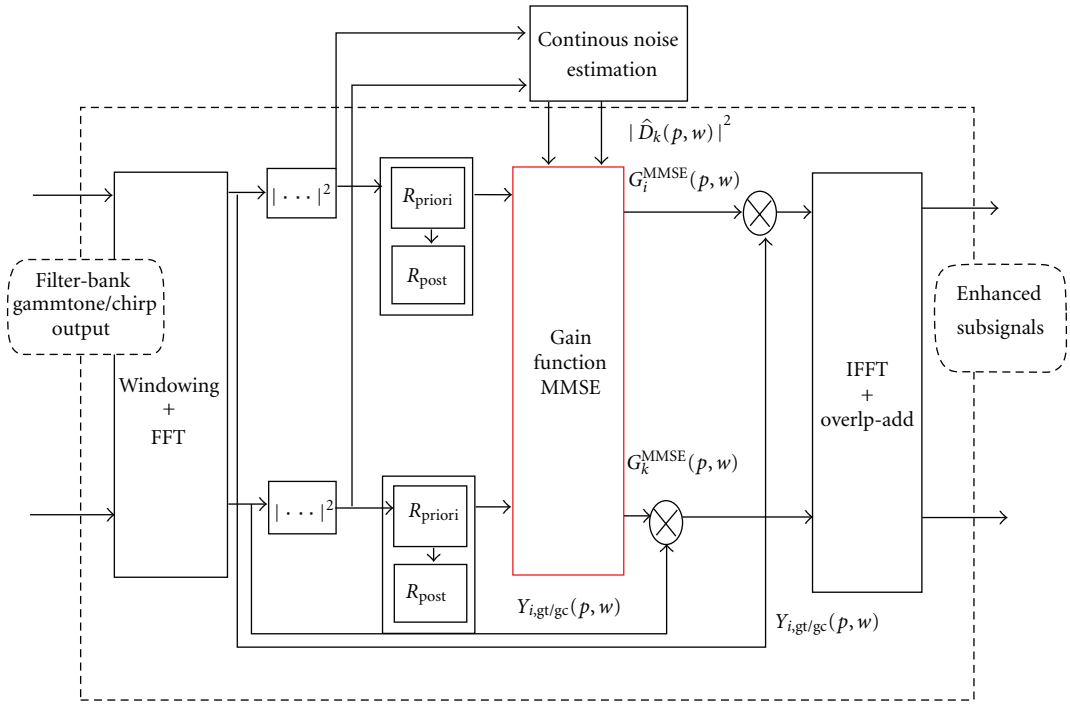

Figure 4: Perceptual MMSE spectral modification technique applied in multirate system.

subband signals $y_{k,\text{gt/gc}}(n,p)$ are transformed into the frequency domain with the fast Fourier transformation (FFT) and manipulated using the spectral gain given by the generalized spectral subtraction rule (GSS), on one side, and the Ephraim and Malah spectral rule (MMSE), on the other side.

3.2.1. Perceptual Generalized Spectral Subtraction Technique. The function gain of the generalized spectral subtraction rule is applied in a multirate system (Figure 3). The subbands spectrums of the noisy signal are multiplied by the general weights $G_{k,\text{gt/gc}}^{\text{GSS}}(p,w)$ in each subband k.

The multibands weights are calculated from the subbands magnitude spectrum of the noisy speech signal and the noise estimate in each frame p and for each frequency w. Using the generalized spectral subtraction technique, the enhanced speech spectrum $|\hat{X}_{k,\text{gt}}^{\text{GSS}}(p,w)|$ in each gammatone subband signal is given by

$$\left|\hat{X}_{k,\text{gt}}^{\text{GSS}}(p,w)\right| = G_{k,\text{gt}}^{\text{GSS}}(p,w)\cdot\left|Y_{k,\text{gt}}(p,w)\right|. \tag{8}$$

And in each gammachirp subband signal, the enhanced speech spectrum $|\hat{X}_{k,\text{gc}}^{\text{GSS}}(p,w)|$ is given by

$$\left|\hat{X}_{k,\text{gc}}^{\text{GSS}}(p,w)\right| = G_{k,\text{gc}}^{\text{GSS}}(p,w)\cdot\left|Y_{k,\text{gc}}(p,w)\right|, \tag{9}$$

where the gain functions $G_{k,\text{gt}}^{\text{GSS}}(p,w)$ and $G_{k,\text{gc}}^{\text{GSS}}(p,w)$ are expressed in each subband k as

$$G_{k,\text{gt/gc}}^{\text{GSS}}(p,w) = \begin{cases} \left(1-\alpha\left[\dfrac{\left|\hat{D}_{k,\text{gt/gc}}(p,w)\right|}{\left|\left|Y_{k,\text{gt/gc}}(p,w)\right|\right|}\right]^2\right)^{1/2} \\ \text{if } \left|\hat{X}_{k,\text{gc}}(p,w)\right|^2 > \beta\left|\hat{D}_{k,\text{gt/gc}}(p,w)\right|^2 \\ \beta\left(\left|\hat{D}_{k,\text{gt/gc}}(p,w)\right|^2\right)^{1/2} \\ \text{otherwise,} \end{cases} \tag{10}$$

where $|\hat{D}_{k,\text{gt/gc}}(p,w)|^2$ and $|Y_{k,\text{gt/gc}}(p,w)|^2$ are respectively the power spectrum of the noise estimate and the noisy speech signal in each nonuniform gammatone (gt) and gammachirp (gc) subband k. α is the over-subtraction factor ($\alpha \geq 1$), and β ($0 < \beta < 1$) is the spectral floor.

3.2.2. Perceptual MMSE Spectral Modification. In this section, we are interested in using the spectral gain $G_{k,\text{gt/gc}}^{\text{mmse}}(p,w)$ given by the spectral modification according to the Ephraim and Malah rule (MMSE) in each frame p and each frequency w (Figure 4) to obtain the enhanced speech spectrum $\hat{X}_{k,\text{gt/gc}}^{\text{mmse}}(p,w)$ in each gammatone subband signal as

$$\left|\hat{X}_{k,\text{gt}}^{\text{mmse}}(p,w)\right| = G_{k,\text{gt}}^{\text{mmse}}(p,w)\cdot\left|Y_{k,\text{gt}}(p,w)\right|. \tag{11}$$

And in each gammachirp subband signal, the enhanced speech spectrum $|\hat{X}_{k,\text{gc}}^{\text{mmse}}(p,w)|$ is given by

$$\left|\hat{X}_{k,\text{gc}}^{\text{mmse}}(p,w)\right| = G_{k,\text{gc}}^{\text{mmse}}(p,w)\cdot\left|Y_{k,\text{gc}}(p,w)\right|, \tag{12}$$

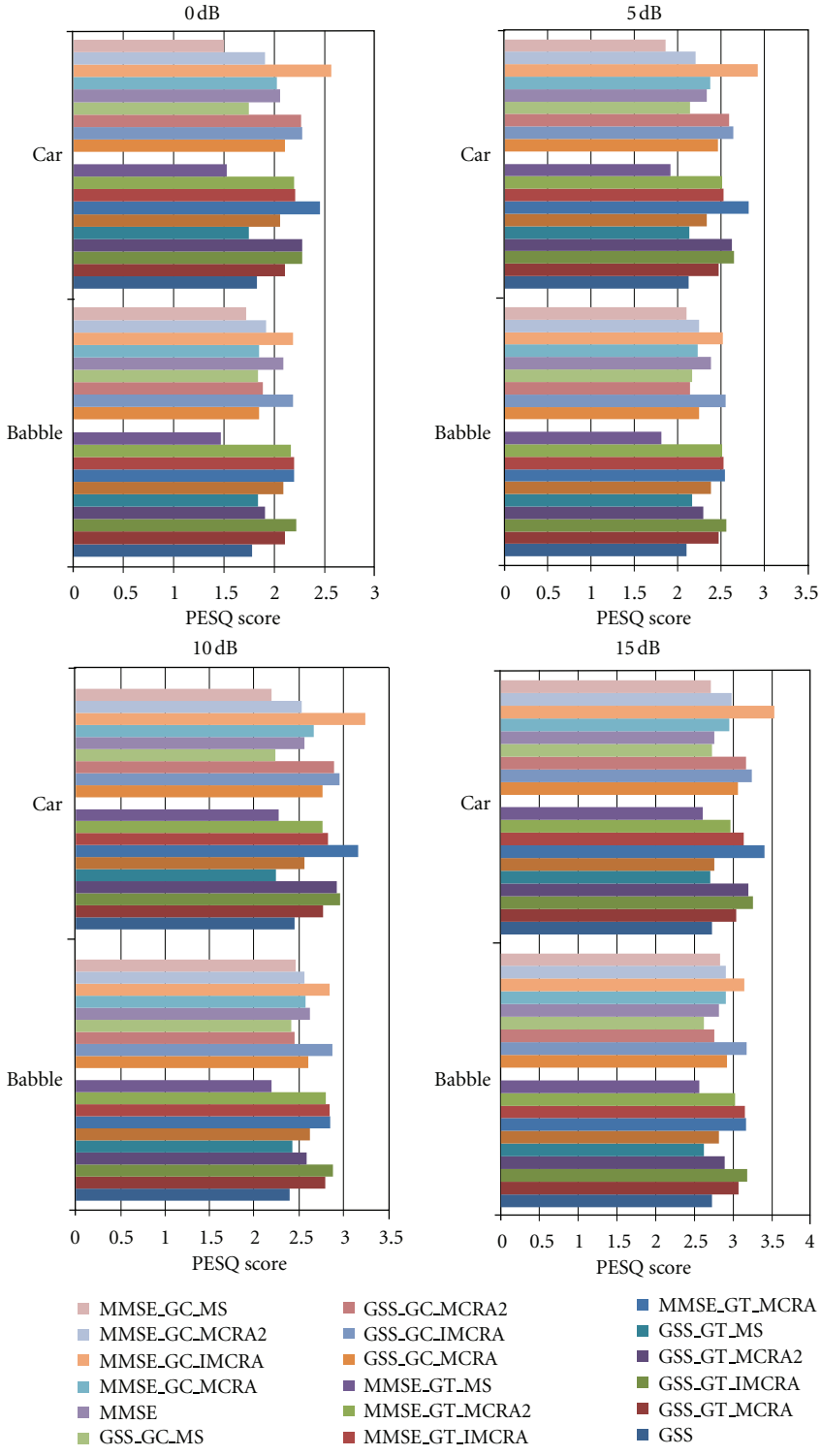

Figure 5: PESQ score for the proposed perceptual speech enhancement method based on the gammatone and the gammachirp filter banks decomposition at different signal-to-noise ratio for babble and car noise and compared to the classic algorithms.

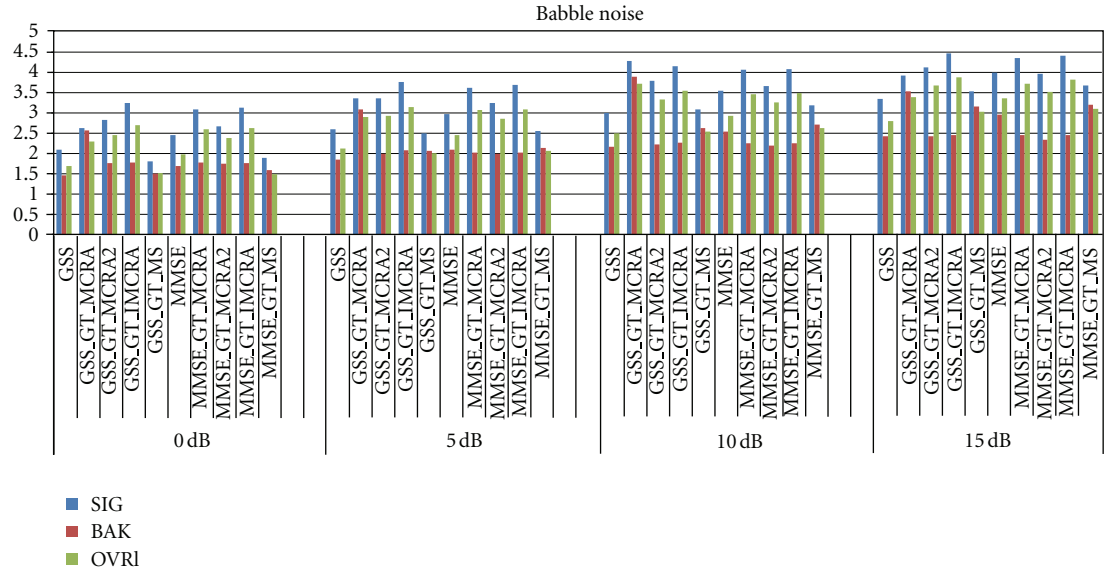

Figure 6: SIG-BAK-OVRL scores for proposed speech enhancement with the gammachirp decomposition compared to the GSS and the MMSE at different SNR input for babble noise.

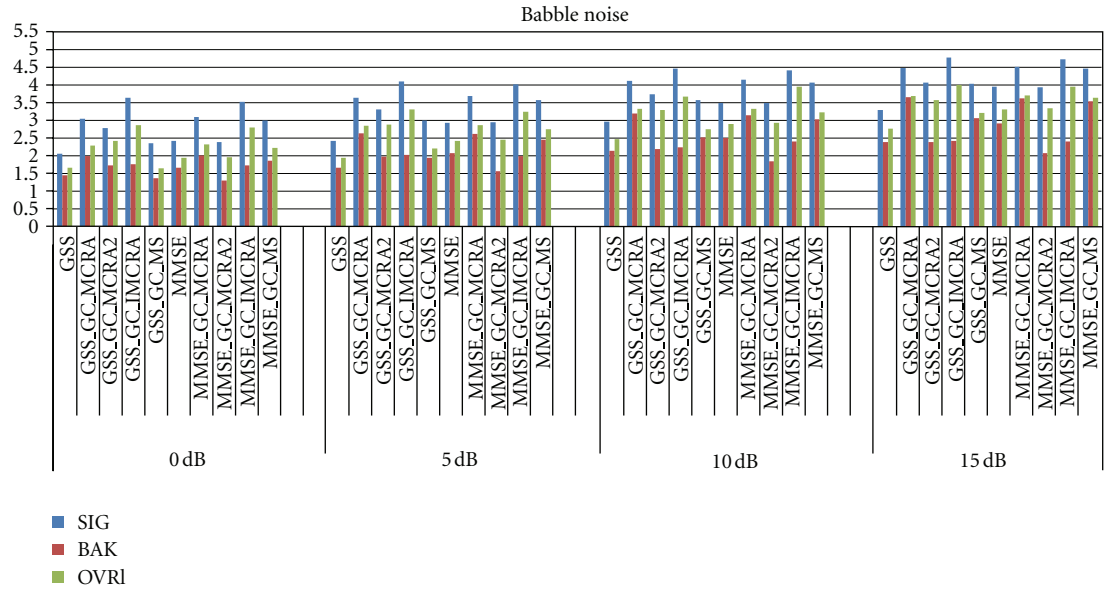

Figure 7: SIG-BAK-OVRL scores for proposed speech enhancement with the gammachirp decomposition compared to the GSS and the MMSE at different SNR input for babble noise.

where the gain functions $G^{\text{mmse}}_{k,\text{gt}}(p,w)$ and $G^{\text{mmse}}_{k,\text{gc}}(p,w)$ are expressed in each subband k as

$$G^{\text{mmse}}_{k,\text{gt/gc}}(p,w) = \frac{\sqrt{\pi}}{2}\cdot\sqrt{\left(\frac{1}{1+R\text{post}_{k,\text{gt/gc}}}\right)\times\left(\frac{R\text{priori}_{k,\text{gt/gc}}}{1+R\text{priori}_{k,\text{gt/gc}}}\right)\cdot F.1+R\text{post}_{k,\text{gt/gc}}\times\left(\frac{R\text{priori}_{k,\text{gt/gc}}}{1+R\text{priori}_{k,\text{gt/gc}}}\right)}. \tag{13}$$

The local and relative level a posterior and the prior signal to noise ratio in the current frame p and each gammatone (gt) and gammachirp (gc) subband are defined as:

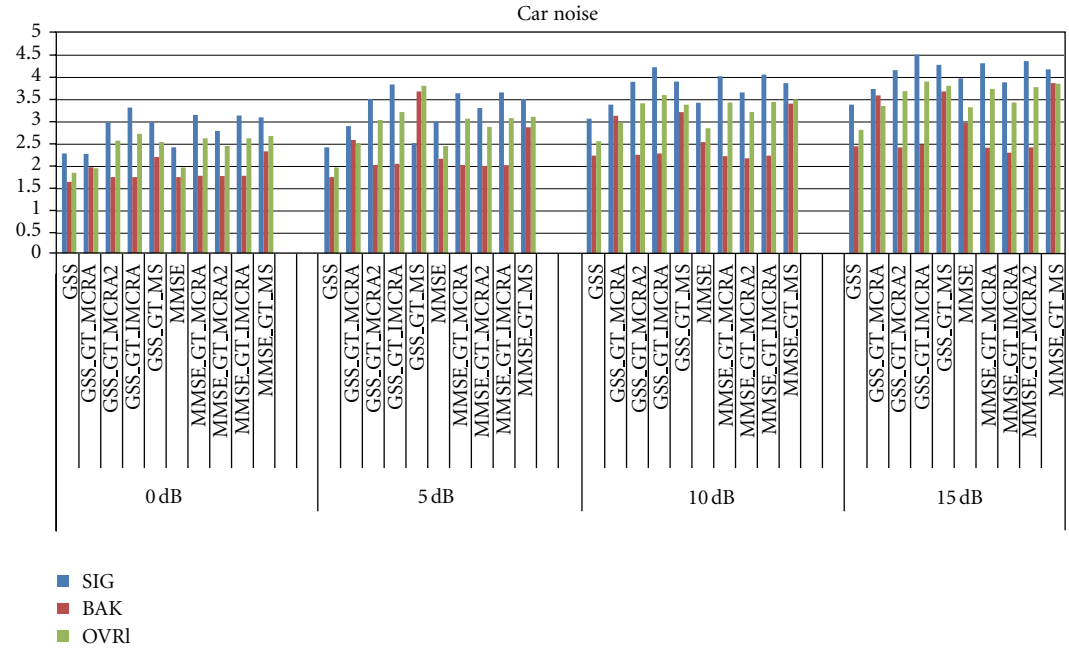

Figure 8: SIG-BAK-OVRL scores for proposed speech enhancement with the gammtone decomposition compared to the GSS and the MMSE at different SNR input for car noise.

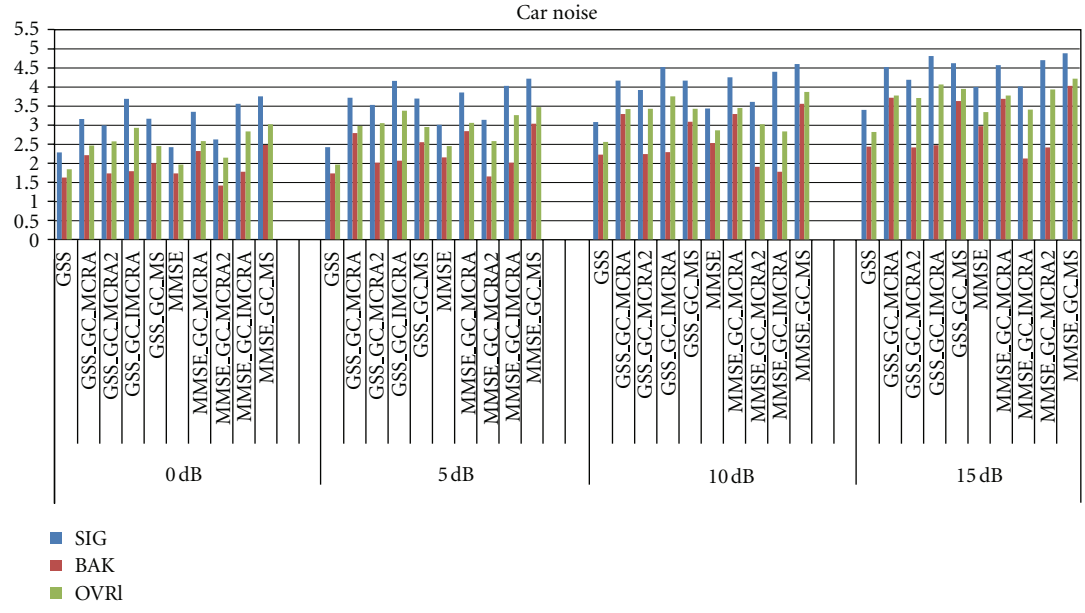

Figure 9: SIG-BAK-OVRL scores for proposed speech enhancement with the gammachirp decomposition compared to the GSS and the MMSE at different SNR input for car noise.

$$
\begin{aligned}
R\text{post}_{k,\text{gt/gc}}(p,w) &= \frac{\left|Y_{k,\text{gt}}(p,w)\right|^2}{\left|\hat{D}_{k,\text{gt}}(p,w)\right|^2},\\
R\text{priori}_{k,\text{gt/gc}} &= (1-\eta)\left[R\text{post}_{k,\text{gt/gc}}-1\right]\\
&\quad\times\frac{\left|Y_{k,\text{gt/gc}}(p-1,w)\right|^2}{\left|\hat{D}_{k,\text{gt/gc}}(p,w)\right|^2}.
\end{aligned}
\tag{14}
$$

γ is a parameter defined as $0 \le \eta \le 1$.

$|Y_{k,\text{gt/gc}}(p-1,w)|^2$ is the power spectral density defined in the frame $(p-1)$ and R_{post} is the relative level a posterior defined in each frame p and for each frequency w.

The temporal enhanced speech signal $\hat{x}_{k,\text{gt/gc}}(n)$ in each temporal subband k is estimated using the overlap-add technique and the inverse Fourier transform based on the assumption that phase distortion is not perceived by the human ear, the phase of the noisy speech is not processed and the enhanced speech signal in each subband k is obtained by using the inverse Fourier transform and the phase from the noisy speech signal.

The final enhanced output speech signal $\hat{x}_{\text{gt}}(n)$ from the gammatone synthesis filter banks and the gammachirp synthesis filter banks $\hat{x}_{\text{gc}}(n)$ are obtained by using the summation of the subband signals after processing

$$
\begin{aligned}
\hat{x}_{\text{gt}}(n) &= \sum_{k=1}^{M}\hat{x}_{k,\text{gt}}(n),\\
\hat{x}_{\text{gt}}(n) &= \sum_{k=1}^{M}\hat{x}_{k,\text{gt}}(n),
\end{aligned}
\tag{15}
$$

where $\hat{x}_{k,\text{gt}}(n)$ and $\hat{x}_{k,\text{gc}}(n)$ are given by

$$
\begin{aligned}
\hat{x}_{k,\text{gt}}(n) &= \text{IFFT}\left[\left|\hat{X}_{k,gt}(p,w)\right| e^{j\phi(Y_{k,gt}(p,w))}\right],\\
\hat{x}_{k,\text{gc}}(n) &= \text{IFFT}\left[\left|\hat{X}_{k,gc}(p,w)\right| e^{j\phi(Y_{k,gc}(p,w))}\right].
\end{aligned}
\tag{16}
$$

The noise estimate can have an important impact on the quality and intelligibility of the enhanced signal. If the noise estimate is too low, a residual noise will be audible; if the noise estimate is too high, speech will be distorted resulting in intelligibility loss. In the spectral subtraction algorithm, the noise spectrum estimate is updated during the silent moment of the signal. Although this approach might give satisfactory result with stationary noise, it will not with more realistic environments where the spectral characteristics of the noise change constantly. Hence, there is a need to update the noise spectrum continuously over time. Several noise-estimation algorithms have been proposed for speech enhancement applications [14]. In [15], the minimum statistics method for estimating the noise spectrum (MS) is based on tracking the minimum of the noisy speech over a finite window. As the minimum is typically smaller than the mean, unbiased estimates of noise spectrum were computed by introducing a bias factor based on the statistics of the minimum estimates. In [16], a minima controlled recursive algorithm (MCRA) is proposed; it updates the noise estimate by tracking the noise-only regions of the noisy speech spectrum. These regions are found by comparing the ratio of the noisy speech to the local minimum against a threshold. In the improved minima controlled recursive algorithm (IMCRA) approach [17], a different method was used to track the noise-only regions of the spectrum based on the estimated speech-presence probability. This probability, however, is also controlled by the minima. Recently, a new noise estimation algorithm (MCRA2) was introduced [18], the noise estimate was updated in each frame based on voice activity detection. The speech presence decision made in each frame is based on the ratio of the noise speech spectrum to its local minimum. In our work, the noise power spectrum is continuously estimated using these algorithms.

Table 1: Experimental parameters used in the noise suppression process.

Algorithms	Parameters
GSS	$\gamma = 2; \alpha = 1$
	$\beta = 2 \cdot 10^{-3}$
MMSE	$\eta = 0.95$

4. Results and Evaluation

The speech signals are obtained from TIMIT corpus. The sentences are sampled at 16 kHz. The noise is added to the original speech signal at different signal to noise ratio (0 dB, 5 dB, 10 dB, and 15 dB) from the AURORA database and includes multitalker babble and car noise. The database is used as it contains phonetically balanced sentences with relatively low word context predictability. To cover the frequency range of the signal, the analysis stage used in the multibands subtraction consists of 27-4th order gammatone/gammachirp filter banks according to the ERB scale. The parameters used in the noise suppression algorithms are set to Table 1.

The performance of the proposed speech enhancement method: the generalized spectral subtraction rule implemented on ERB gammatone/gammachirp filter banks (GSS_GTFB/GSS_GCFB) and the Ephraim and Malah spectral modification rule implemented on ERB gammatone/gammachirp filter banks (MMSE_GTFB/MMSE_GCFB) using continuous noise estimation algorithms based on the MCRA method (mcra), the IMCRA method (imcra), the MCRA2 method (mcra2), and the minimum statistics method (ms), are evaluated and compared with that the generalized spectral subtraction (GSS) and the Ephraim and Malah (MMSE) spectral modification basics techniques.

4.1. Objective Evaluation. In order to evaluate the performance, we measure the perceptual evaluation of speech quality PESQ [13]. The PESQ score is able to predict subjective quality with good correlation in a very wide range of conditions, the original and degraded signals are mapped

Table 2: Recognition rate for the proposed perceptual generalized spectral subtraction method based on the gammatone and the gammachirp filter banks decomposition (GSS_GT and GSS_GC) compared with the spectral subtraction rule (GSS).

	Babble noise			
Methods/SNR	0	5	10	15
GSS	16.88%	31.25%	44.19%	64.96%
GSS_GT				
MCRA	30.02%	39.76%	51.56%	80.98%
MCRA2	28.12%	36.88%	47.69%	64.25%
IMCRA	30.68%	45.77%	63.21%	84.24%
MS	26.63%	33.86%	50.03%	79.58%
GSS_GC				
MCRA	33.80%	41.62%	56.23%	89.01%
MCRA2	29.99%	37.80%	48.89%	71.53%
IMCRA	33.89%	56.07%	68.21%	91.22%
MS	27.38%	35.96%	55.83%	83.98%
	Car noise			
Methods/SNR	0	5	10	15
GSS	47.88%	61.25%	65.19%	73.56%
GSS_GT				
MCRA	67.90%	82.83%	77.63%	88.98%
MCRA2	58.61%	77.76%	67.42%	74.09%
IMCRA	70.68%	86.40%	87.87%	90.09%
MS	57.08%	70.85%	75.29%	88.08%
GSS_GC				
MCRA	69.09%	87.99%	79.80%	89.97%
MCRA2	68.78%	85.60%	79.04%	80.43%
IMCRA	71.08%	88.86%	89.89%	91.95%
MS	59.18%	72.15%	78.90%	89.01%

onto an internal representation using a perceptual model to predict the perceived speech quality of the degraded signal. The subjective experiments used in the development of the PESQ uses the absolute category rating opinion scale.

According to the results illustrated in Figure 5, we note that the approach based on nonuniform filter banks decomposition using two different models of the human perceptual comportment is performed in speech enhancement. We observe that the PESQ score is consistent with the subjectively perceived trend of an improvement in speech quality with the proposed speech enhancement approach over that the spectral modification (GSS) algorithm alone.

This improvement is particularly significant in the case of car noise at 15 dB, and we register a score of 3,26 for the proposed GSS_GT (using gammatone decomposition) in spite of 2,73 for the GSS alone; the PESQ improvement is also observed using the GSS_GT at 0 dB (2,22) for babble noise continuously estimated with the MCRA2, contrary to the GSS (1, 78). On the other hand, the gammachirp filter banks decomposition in association with the MMSE spectral modification rules (MMSE_GC_MCRA) contributes significantly in the enhancement of speech signal corrupted by car noise (3,54 PESQ score at 15 dB). In order to strengthen the objective evaluation, we measure the scores relative to the standard norm P. 835 [19].

This norm attends and rates successively the enhanced speech signal on the distortion of the speech signal alone using five-point scale of signal distortion (SIG), the noise distortion using a five-point scale of background intrusiveness (BAK), and the overall quality effect (OVRL). This process is designed to integrate the effects of both the signal and the background in making the rating of overall quality.

Figures 6 and 7 list at different signal to noise ratio the subjective overall quality the OVRL measure that includes the naturalness of speech (SIG) and intrusiveness of background noise (BAK) for babble noise. Figures 8 and 9 list the SIG-BAK and OVRL scores for the car noise. the proposed perceptual spectral modification using different continuous noise estimation algorithms performed significantly better than the classic spectral subtractive algorithms.

Lower signal distortion (higher SIG score) is observed with the proposed approach in most condition with significant differences at 10 dB for car noise: a SIG score of 3,09 given by the GSS, and improved by the GSS_GT to 4,27 using IMCRA noise estimation and a score of 4,62 registered by the proposed MMSE_GC with the MS noise estimation. This demonstrates the performance of our approach based on nonuniform gammatone/gammachirp filter banks decomposition to reduce the noticeable of the background noise and minimize the signal distortion. We

Table 3: Recognition rate for the proposed perceptual generalized spectral subtraction method based on the gammatone and the gammachirp filter banks decomposition (MMSE_GT and MMSE_GC) compared with the spectral modification rule (MMSE).

	Babble noise			
Methods/SNR	0	5	10	15
MMSE	64.81%	79.41%	89.36%	90.13%
MMSE_GT				
MCRA	67.69%	84.79%	93.18%	94.97%
MCRA2	65.90%	83.74%	92.24%	93.93%
IMCRA	68.76%	84.41%	93.80%	94.62%
MS	66.24%	82.15%	92.07%	93.07%
MMSE_GC				
MCRA	57.36%	74.40%	87.66%	93.00%
MCRA2	61.29%	74.73%	87.32%	92.89%
IMCRA	67.91%	84.08%	91.83%	93.24%
MS	65.78%	75.33%	90.33%	91.10%
	Car noise			
Methods/SNR	0	5	10	15
MMSE	75.88%	85.25%	91.40%	92.96%
MMSE_GT				
MCRA	80.68%	89.56%	92.64%	95.13%
MCRA2	79.03%	86.83%	91.51%	93.84%
IMCRA	82.40%	90.51%	92.06%	94.99%
MS	79.78%	90.15%	91.88%	93.35%
MMSE_GC				
MCRA	74.77%	86.71%	91.95%	93.69%
MCRA2	72.64%	82.49%	90.33%	93.08%
IMCRA	80.18%	90.20%	92.29%	94.89%
MS	79.13%	89.75%	92.57%	93.80%

notice also that incorporating continuous noise estimation in particularly the IMCRA and the MCRA continuous noise estimation in the perceptual spectral modification approach performed better than the generalized spectral subtraction and the MMSE rules in the overall quality improvement.

This indicates that the proposed perceptual spectral modification for speech enhancement is sensitive to the noise spectrum estimate.

4.2. Subjective Evaluation. Significant gains in noise reduction are accompanied by a decrease in speech intelligibility. Formal subjective test is the best indicator of achieved overall quality. So the subjective evaluation used in our work is based on an automatic recognition system (ASR) developed under the HTK platform [20]. Thus, we used a standard continuous density HMM recognizer with 3 Gaussian mixtures per state, diagonal covariance matrices, and 5 emitting states per word model.

The parameterise step is consisted of 12 MFCC coefficients.

Tables 2 and 3 show the world recognised rate in percent (%) at different SNRs for the proposed approach using the two auditory filtering models compared to the classic spectral modification rules.

We observe that the proposed multibands approach gives the best world rate recognition. It can be seen that the amelioration is significant, especially, in the case of car noise at different level of degradation.

5. Conclusion

In this paper, we proposed a new speech enhancement method which consists of integrating psychoacoustics proprieties of the human auditory system, especially perceptual filters modeling. It is based on decomposing the input signal in nonuniform subbands using an analysis/synthesis gammatone and gammachirp filter banks that are manipulated in each nonlinear block with the generalized spectral subtraction process and the MMSE spectral modification technique. We noticed that the use of the two perceptual filter banks models with frequency resolution according to the ERB scale allowed obtaining, from the perceptive point of view and from the vocal quality, better results than those supplied by the classic spectral modification algorithms to

improve the quality and intelligibility of the enhanced speech signal.

References

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 208–211, April 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error-log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[3] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 4, pp. 4160–4164, May 2002.

[4] R. M. Udrea, S. Ciochină, and D. N. Vizireanu, "Multi-band bark scale spectral over-subtraction for colored noise reduction," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '05)*, pp. 311–314, Iasi, Romania, July 2005.

[5] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, pp. 537–540, May 2002.

[6] R. M. Udrea, N. Vizireanu, S. Ciochina, and S. Halunga, "Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale," *Signal Processing*, vol. 88, no. 5, pp. 1299–1303, 2008.

[7] R. M. Udrea, N. D. Vizireanu, and S. Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter," *Digital Signal Processing*, vol. 18, no. 4, pp. 581–587, 2008.

[8] M. Udrea and S. Ciochina, "Speech enhancement using spectral over-subtraction and residual noise reduction," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '03)*, vol. 2, pp. 311–314, 2003.

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

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

[11] A. Amehraye, D. Pastor, A. Tamtaoui, and D. Aboutajdine, "From maskee to audible noise in perceptual speech enhancement," *International Journal of Signal Processing*, vol. 5, article 2, 2009.

[12] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[13] T. Irino and M. Unoki, "An analysis/synthesis perceptual filterbank based on an IIR gammachrp filter," in *Computational models of Perceptual Function*, S. Greenberg and M. Slaney, Eds., vol. 312 of *NATO ASI Series*, IOS Press, 2001.

[14] P. Loizou, "Speech Enhancement: Theory and Practice," CRC Press, Boca Raton, Fla, USA.

[15] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[16] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[17] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[18] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

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

[20] S. J. Young, *The HTK Book 3.1*, Entropic, 2002.