An Improved Multi-band Speech Enhancement Method for Colored Noise Estimation and Reduction

Radu Mihnea Udrea, Dragos Nicolae Vizireanu, Claudia Cristina Oprea, Ionut Pirnog
Telecommunications Department
"Politehnica" University of Bucharest
313, Splaiul Independentei, Sector 6, 060042
Bucharest, Romania
mihnea@comm.pub.ro, nae@comm.pub.ro, cristina@comm.pub.ro, ionut@comm.pub.ro

***Abstract*—There are many situations where speech is affected by different kind of acoustic noise. We propose an improved spectral subtraction method for the reduction of colored acoustic noise added to the speech. Our implementation uses a multi-band spectral over subtraction method to reduce the colored noise. We use a non-linear Bark scale distribution to estimate the over-subtraction factor. The noise power spectral density is estimated, using a time-recursive algorithm, by tracking the minimum of the noisy speech spectrum in each frequency band. Simulations show a better quality in terms of Itakura Saito distance and perceptual evaluation of quality for the enhanced speech. Using the proposed speech enhancement method, a very good speech quality with less musical noise and with minimal speech distortion is obtained.**

Keywords-speech enhancement; spectral subtraction; noise estimation; critical band.

I. INTRODUCTION

The speech signal is often accompanied by the background noise of the environment. There are many negative effects when processing the degraded speech for some applications like: voice command systems, voice recognition, speaker authentication, hands-free systems.

The main objective of speech enhancement is to improve the perceptual aspects of speech such as overall quality or intelligibility. Enhancement techniques can be classified as single channel and dual channel or multi channel enhancement techniques. Single channel enhancement techniques apply to situations in which only one acquisition channel is available. In multi channel enhancement techniques, a reference signal for the noise is available and hence adaptive noise cancellation technique can be applied.

The spectral subtraction method is a well-known single channel noise reduction technique. The basic spectral subtraction technique proposed by Boll [2] apply subtraction of the noise spectrum estimate over the speech spectrum. The conventional power spectral subtraction method substantially reduces the noise levels in the noisy speech. However, it also introduces an annoying distortion in the speech signal called musical noise. Due to the inaccuracies in the short-time noise spectrum estimate, large spectral variations exist in the enhanced spectrum causing these distortions.

Berouti [3] proposed an important variation of spectral subtraction for reduction of residual musical noise. The proposed method subtracts an overestimate of the noise power spectrum from the speech power spectrum. This operation minimizes the presence of residual noise by decreasing the spectral excursions in the enhanced spectrum. The over-subtraction factor provides a degree of control over the noise removal process between periods of noise update.

A nonlinear approach to the subtraction procedure was proposed in recent studies [4], [5], [6], [7], which takes into account the variation of the signal-to-noise ratio across the entire speech spectrum. The real-world noise spectrum is not flat, therefore the noise signal does not affect the speech signal uniformly over the whole spectrum. Hence, it becomes imperative to estimate a suitable factor that will subtract just the necessary amount of the noise spectrum from each frequency sub-band, to prevent destructive subtraction of the speech while removing most of the residual noise.

Noise spectrum estimation is also a challenging situation. Several noise-estimation algorithms have been proposed for speech enhancement applications. For rather stationary noise sources, the noise power spectral density (PSD) can be estimated by tracking the minimum of the noisy speech spectrum in each frequency band [8], [9]. However, in case of non-stationary noise sources, more advanced methods can be used [10].

In this paper we used a modified spectral over-subtraction approach that allows better and more suppression of the noise. We propose to use the noise PSD estimation to compute the *a posteriori* SNR in each frequency subband. Then we calculate the corresponding over-subtraction factor and we apply the nonlinear multi-band spectral subtraction that reduces colored noise, using a different over-subtraction factor in each frequency band.

This paper is organized as follows: a overview of the spectral subtraction methods is presented in Section II. Section III presents the human auditory system and the critical band and Bark scale model of speech analysis. Section IV describes the time-recursive averaging type of algorithms in which the noise spectrum is estimated in order to be used by the spectral subtraction method. Section V presents the improved multi-band spectral over-subtraction

method proposed to reduce the colored noise. Section VI shows implementation details and experimental results.

II. SPECTRAL SUBTRACTION METHODS

A. Basic Spectral Subtraction Method

The spectral subtraction method proposed by Boll [2] consists in obtaining an estimate of the noise-free signal spectrum by subtracting an estimate of the noise spectrum from the input noisy signal spectrum. The background noise is considered acoustically added to the speech. It is assumed that the background noise remains locally stationary to the degree that its spectral magnitude expected value prior to speech activity equals its expected value during speech activity.

The noise is assumed to be uncorrelated and additive to the speech signal. An estimate of the noise signal is measured during silence or non-speech activity in the signal. We assume that the speech signal $s(n)$ has been degraded by the additive noise signal $d(n)$,

$$y(n) = s(n) + d(n). \tag{1}$$

Taking the Discrete Fourier Transform (DFT) of $x(n)$ gives

$$Y(k) = S(k) + D(k). \tag{2}$$

The estimate of the noise spectrum is obtained during speech pauses (SP) when

$$y_{SP}(n) = d(n). \tag{3}$$

Noise spectrum can be estimated as the average value of $|Y_{SP}(k)|$ over the speech pauses frames

$$\left|\hat{D}(k)\right| = \frac{1}{M}\sum_{i=0}^{M-1}\left|Y_{SP_i}(k)\right|, \tag{4}$$

where M is the number of consecutive frames of SP.

The estimate of the clear speech spectrum magnitude can be obtained as

$$\left|\hat{S}(k)\right| = \left|Y(k)\right| - \left|\hat{D}(k)\right|. \tag{5}$$

The phase $\theta_Y(k)$ of the input signal is used for reconstruction of the estimated signal spectrum based on the fact that for human perception the short time spectral magnitude is more important than the phase for intelligibility and quality. This conclusion was made by Lim and Wang in their work [11], when using the actual phase rather than the degraded speech phase does not improve the quality of the enhanced speech.

Therefore,

$$\hat{S}(k) = \left|\hat{S}(k)\right| e^{j\theta_Y(k)}. \tag{6}$$

The time reconstructed speech signal is obtained taken the Inverse Discrete Fourier Transform of $\hat{S}(k)$.

Since the noise spectrum cannot be directly obtained, there are some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. This residual spectral content manifest themselves in the reconstructed time signal as varying tonal sounds resulting in a musical disturbance of an unnatural quality. This sounds like a musical noise and is the main drawback of the spectral subtraction method.

B. Generalized Spectral Subtraction

A generalized form of the basic spectral subtraction was proposed by Berouti [3]. The estimate of the enhanced speech is given by

$$\left|\hat{S}(k)\right| = \left(\left|Y(k)\right|^p - \alpha\left|\hat{D}(k)\right|^p\right)^{1/p}, \tag{7}$$

where p is the exponent of the spectrum and α is a multiplication factor used for over-subtraction of noise spectrum estimate. For $\alpha = 1$ and $p = 2$ we have the *Power Spectral Subtraction* method.

Power spectral relation after taking DFT from (1) gives

$$\begin{aligned} |Y(k)|^2 = &|S(k)|^2 + |D(k)|^2 + \\ &+ S(k)D^*(k) + S^*(k)D(k) \end{aligned} \tag{8}$$

where $S^*(k)$ and $D^*(k)$ are complex conjugates of $S(k)$ and $D(k)$ respectively.

Because in our system only the power of the input noisy signal $|Y(k)|^2$ can be evaluated, the rest of terms are approximated by their average during non-speech activity period.

If $d(n)$ is uncorrelated with $s(n)$, then

$$E\{S(k)D^*(k)\} = 0 \text{ and } E\{S^*(k)D(k)\} = 0. \tag{9}$$

The short time power spectrum of the noisy speech can be approximated by

$$\left|Y(k)\right|^2 \approx \left|S(k)\right|^2 + \left|D(k)\right|^2. \tag{10}$$

The noise PSD $\hat{\sigma}_d^2(k)$ is estimated as the average value of the noise power spectrum taken during non-speech activity periods.

$$\hat{\sigma}_d^2(k) \simeq E\left\{\left|D(k)\right|^2\right\}. \quad (11)$$

A significant improvement to minimize the presence of residual noise and musical noise in the processed speech was proposed by Berouti et al. [3]. The average noise power spectrum is multiplied by the over-subtraction factor α and subtracted from the noisy speech spectrum in order to minimize the residual and musical noise:

$$\left|\hat{S}(k)\right|^2 = \left|Y(k)\right|^2 - \alpha \cdot \hat{\sigma}_d^2(k), \quad \alpha \geq 1. \quad (12)$$

This method improves the noise suppression better than basic spectral subtraction technique and also eliminates the musical noise. Besides it adapts to wide range of signal to noise ratios.

After subtracting an overestimate of the noise power spectrum the resulting estimated speech spectrum is down-limited at a minimum β level (spectral floor):

$$\left|\hat{S}(k)\right|^2 = \begin{cases} \left|\hat{S}(k)\right|^2, \text{ if } \left|\hat{S}(k)\right|^2 > \beta \cdot \hat{\sigma}_d^2(k) \\ \beta \cdot \hat{\sigma}_d^2(k), \text{ otherwise} \end{cases}, \quad (13)$$

where the spectral floor parameter was set to $\beta = 0.001$.

These modifications lead to minimizing the perception of the narrow spectral peaks by decreasing the spectral excursions and thus lower the musical noise perception.

To reduce the speech distortion caused by large values of α, its value is adapted from frame to frame. The basic idea is to take into account that the subtraction process must depend on the *a posteriori* SNR of the frame, in order to apply less subtraction with high *a posteriori* SNR and vice versa.

The *a posteriori* SNR is calculated for every frame with:

$$\gamma = 10\log_{10}\frac{\sum_{k=0}^{N-1}\left|Y(k)\right|^2}{\sum_{k=0}^{N-1}\hat{\sigma}_d^2(k)}, \quad (14)$$

where N is the number of frequency bins of DFT.

The over-subtraction factor α can be calculated [8] as:

$$\alpha = \begin{cases} 1 & \gamma \geq 20\text{dB} \\ \alpha_0 - \frac{3}{20}\gamma & -6\text{dB} \leq \gamma < 20\text{dB} \\ 4.9 & \gamma < -6\text{dB} \end{cases}, \quad (15)$$

where $\alpha_0 = 4$ is the desired value of α at $\gamma = 0\text{dB}$.

The over-subtraction factor gives a degree of adaptation from frame to frame, but it may reduce speech spectral information in the same frame for frequency domains where noise PSD is lower, if noise spectrum is not flat.

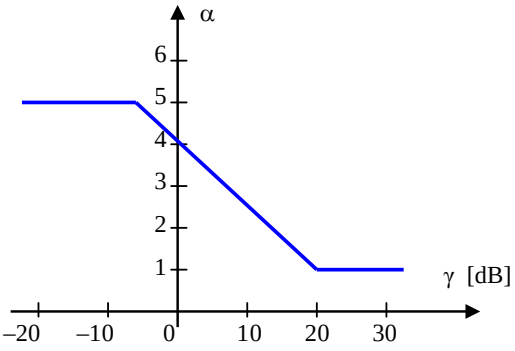

Figure 1. The over-subtraction factor α dependence over the a posteriori SNR γ.

C. Multi-Band Spectral Subtraction

In real environments, the noise spectrum is not uniform for all frequencies. For example, in the case of engine noise, most of the noise energy is concentrated in the low-frequency area. To take into account the fact that colored noise affects the speech spectrum differently at different frequencies, a multi-band linear frequency spacing approach to spectral over-subtraction was proposed in [5].

The speech spectrum is divided into a number of non-overlapping bands, and spectral subtraction is performed independently in each band. The estimate of the clean speech spectrum in the i-th band is obtained by:

$$\left|\hat{S}_i(k)\right|^2 = \left|Y_i(k)\right|^2 - \alpha_i \cdot \hat{\sigma}_d^2(k), \; \nu_i < k < \nu_{i+1}, \quad (16)$$

where k is the frequency bin, ν_i and ν_{i+1} are the beginning and ending frequency bins of the i-th frequency band and α_i is the over-subtraction factor of the i-th band. The over-subtraction factor α_i is a function of the *a posteriori* SNR γ_i of the i-th frequency band.

$$\gamma_i = 10\log_{10}\frac{\sum_{k=\nu_i}^{\nu_{i+1}}\left|Y_i(k)\right|^2}{\sum_{k=\nu_i}^{\nu_{i+1}}\hat{\sigma}_d^2(k)}. \quad (17)$$

The over-subtraction factor α_i may be calculated for each frequency band as:

$$\alpha_i = \begin{cases} 1 & \gamma_i \geq 20\text{dB} \\ 4 - \frac{3}{20}\gamma_i & -6\text{dB} \leq \gamma_i < 20\text{dB} \\ 4.9 & \gamma_i < -6\text{dB} \end{cases}, \quad (18)$$

The negative values of the estimated spectrum were floored using (13).

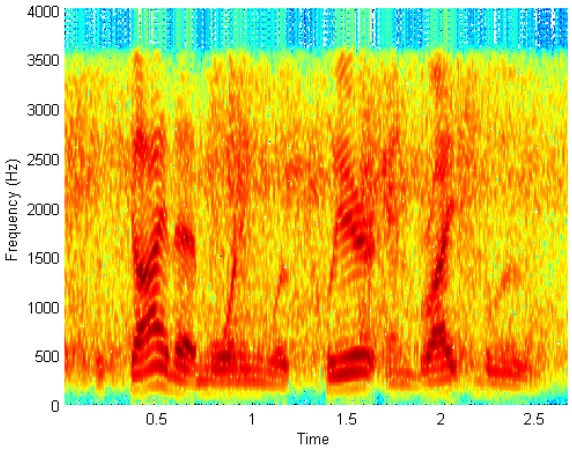

Figure 2. Spectrogram of speech affected by colored noise.

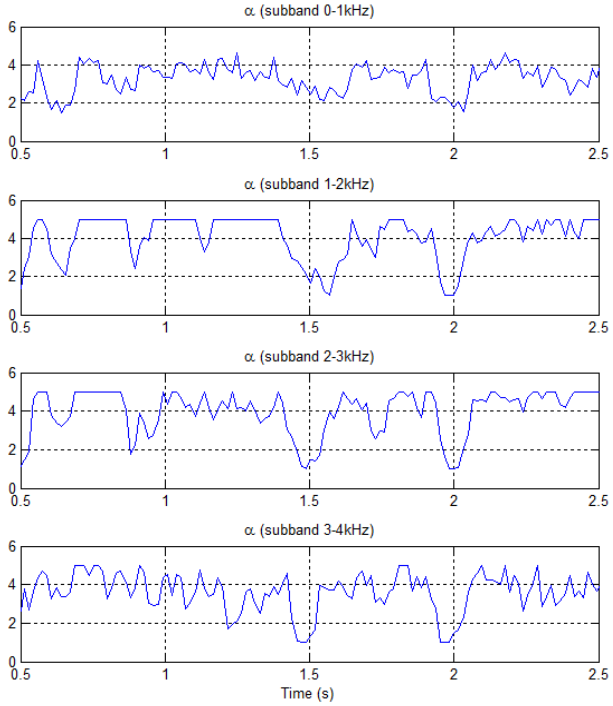

Figure 3. Over-subtraction factor α computed for four linearly-spaced frequency subbands.

In Fig. 2 it is plotted the spectrogram of speech sentence "The sky this morning was clear and light blue" affected by colored noise. Fig. 3 shows the over-subtraction factor computed after dividing the frequency domain into four linearly-spaced frequency subbands. It can be seen that the over-subtraction factor takes different values depending on the SNR in each frequency subband, allowing a better distribution of noise reduction over the entire frequency domain.

Because the human ear sensibility depends nonlinearly on the frequency, a nonlinear frequency spacing approach for multi-band over-subtraction factor estimation is proposed in this paper. The method is presented in Section V.

III. CRITICAL BANDS AND BARK SCALE

The human auditory system is a highly complicated mechanism. During the last decades, a considerable progress has been made within the research of the human hearing. The field of psychoacoustics examines directly the relationships between acoustic stimuli and the associated sensations. The concept of hearing area refers to the ranges of frequency and sound pressure values within which the human ear generally perceives sound.

The absolute threshold of hearing, also known as the threshold in quiet, signifies the minimum sound pressure level of a pure tone that is enough for the tone to be audible in the absence of any interfering voices, i.e., in quiet.

A prominent contributor to the idea of auditory filters was Fletcher who measured the detection threshold of a sinusoidal signal in the presence of a bandpass noise masker. In his experiment, the noise power density was held constant and its centre frequency was always the same as the signal frequency. As the noise bandwidth was increased, the threshold of the signal also increased at first, but after a certain noise bandwidth had been achieved, the signal threshold levelled off.

Basically, the power spectrum model suggests that the peripheral auditory system contains a bank of linear overlapping bandpass filters called auditory filters. It is assumed that when trying to detect a signal in a noisy environment, only one filter whose centre frequency is close to the frequency of the signal is being used. According to the model, this auditory filter blocks out most of the noise and only the part passing through the filter affects the masking of the signal. In reality, the perception of complex signals, e.g. speech, depends on the outputs of several auditory filters and not just one.

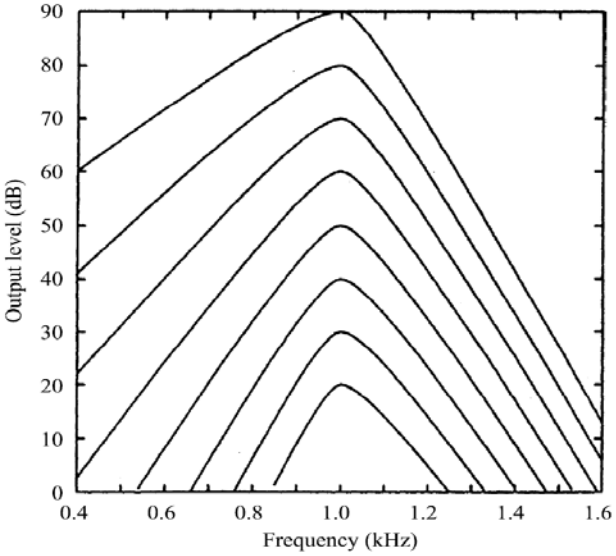

Figure 4. Shape of the auditory filter [12].

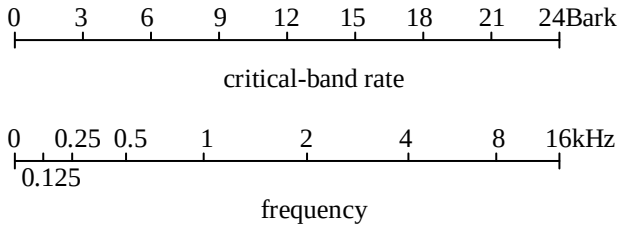

Figure 5. Frequency in Hertz and the critical-band rate scale in Bark [13].

The assumption of linear auditory filters is also incorrect since, strictly speaking, the shape of the filter changes slightly with the input level. As the level of the stimulus is increased, the slope on the low-frequency side of the auditory filter becomes less sharp while the high-frequency skirt becomes steeper. This is illustrated in Fig.4.

In his band-widening experiment, Fletcher introduced the concept of critical bandwidth (CB), denoting the noise bandwidth limit at which the detection threshold of the signal (tone) ceased to increase. For simplicity, he thought that the auditory filter could be approximated as having a rectangular shape and a passband width equal to CB. Fletcher suggested that, with this rectangular model, CB could be evaluated by measuring the threshold of a sinusoidal signal in broadband white noise [12]. In this method, the power of the tone and the power spectral density of the noise masker are first measured. The noise power within the same critical band with the signal is then equal to the product of the measured power spectral density and the CB of the band in question.

The critical bandwidth can also be explained based on the physical structure of the inner ear. Each point on the basilar membrane (BM) responds only to a certain range of frequencies, which leads to the idea that these different points correspond to auditory filters with different centre frequencies [12].

A commonly used scale for specifying the critical bands is the Bark scale which divides the audible frequency range of 16 kHz into 24 bands. Fig. 5 illustrates the relationship between the frequency in Hertz and the critical-band rate in Bark [13].

An approximate analytical expression to describe the conversion from linear frequency, *f*, into the critical band number *z* (in Bark) is [13]:

$$z(f) = 13\arctan\left(0.76 f\right) + 3.5\arctan\left(f / 7.5\right)^2, \quad (19)$$

and the critical bandwidth (in Hz) for a given centre frequency can be evaluated by

$$BW\left(f\right) = 25 + 75\left(1 + 1.4 f^2\right)^{0.69}. \quad (20)$$

In the above equation *f* is given in kHz. By the definition of the Bark scale, each critical band has a width of one Bark.

Table 1 shows the correspondence between Bark scale and the frequency limits for the corresponding CB [13].

Table 1. CRITICAL BANDWIDTH AS A FUNCTION OF CENTER FREQUENCY AND CRITICAL BAND [13].

CB rate	Center frequency	Frequency	CB bandwidth
Bark	***Hz***	***Hz***	***Hz***
0	50	20	80
1	150	100	100
2	250	200	100
3	350	300	100
4	450	400	110
5	570	510	120
6	700	630	140
7	840	770	150
8	1000	920	160
9	1170	1080	190
10	1370	1270	210
11	1600	1480	240
12	1850	1720	280
13	2150	2000	320
14	2500	2320	380
15	2900	2700	450
16	3400	3150	550
17	4000	3700	700
18	4800	4400	900
19	5800	5300	1100
20	7000	6400	1300
21	8500	7700	1800
22	10500	9500	2500
23	13500	12000	3500
24		15500	

IV. NOISE ESTIMATION

The noise signal typically has a nonuniform effect on the spectrum of the speech. Each spectral component will typically have a different effective SNR. Consequently, we can estimate and update individual frequency bands of the noise spectrum whenever the effective SNR at a particular frequency band is extremely low. This observation led to the recursive-averaging type of algorithms [8],[9] in which the noise spectrum is estimated as a weighted average of past noise estimates and the present noisy speech spectrum.

The time-recursive algorithms have the following form:

$$\hat{\sigma}_d^2\left(\lambda,k\right) = \delta\left(\lambda,k\right)\hat{\sigma}_d^2\left(\lambda-1,k\right) + \left(1-\delta\left(\lambda,k\right)\right)\left|Y\left(\lambda,k\right)\right|^2, \quad (21)$$

where $\left|Y\left(\lambda,k\right)\right|^2$ is the speech magnitude spectrum squared (periodogram), $\hat{\sigma}_d^2\left(\lambda,k\right)$ denotes the estimate of the noise power spectral density (PSD) at frame λ and frequency k

and $\delta(\lambda,k)$ is the smoothing factor, which is time and frequency dependent.

In [9], the smoothing factor $\delta(\lambda,k)$ is chosen to be a sigmoid function of the *a posteriori* SNR $\gamma_k(\lambda)$:

$$\delta(\lambda,k)=\frac{1}{1+e^{-\tau(\gamma_k(\lambda)-1.5)}}, \quad (22)$$

where the τ is a parameter with values in the range $15\le\tau\le30$, and $\gamma_k(\lambda)$ is an approximation of the *a posteriori* SNR given by:

$$\gamma_k(\lambda)=\frac{|Y(\lambda,k)|^2}{\frac{1}{10}\sum_{m=1}^{10}\hat{\sigma}_d^2(\lambda-m,k)}. \quad (23)$$

Also, a different function was proposed for computing $\delta(\lambda,k)$:

$$\delta(\lambda,k)=1-\min\left[1,\frac{1}{\gamma_k(\lambda)}\right], \quad (24)$$

used to ensure that $\delta(\lambda,k)$ is in the range of $[0,1]$.

The recursive algorithm given in (21) and (24) can be explained as follows:

- If speech is present, the *a posteriori* estimate $\gamma_k(\lambda)$ will be large and therefore $\delta(\lambda,k)\approx1$. Consequently, we will have $\hat{\sigma}_d^2(\lambda,k)\equiv\hat{\sigma}_d^2(\lambda-1,k)$ according to (21). The noise update will cease and the noise estimate will remain the same as the previous frame's estimate.
- If speech is absent, the *a posteriori* estimate $\gamma_k(\lambda)$ will be small and therefore $\delta(\lambda,k)\approx0$. As a result, $\hat{\sigma}_d^2(\lambda,k)\equiv|Y(\lambda,k)|^2$ and the noise estimate will follow the PSD of the noisy spectrum in the absence of speech.

The main advantage of using the time smoothing factor $\delta(\lambda,k)$ given by (22) or (24), as opposed to using a fixed value for $\delta(\lambda,k)$, is that these factors are time and frequency dependent. This means that the noise PSD will be adapted differently and at different rates in the various frequency bins, depending on the estimate of the *a posteriori* SNR $\gamma_k(\lambda)$ in that bin. This is particularly suited in situations in which the additive noise is colored.

A different and simpler approach [14] to recursive averaging noise estimation is to choose a fixed smoothing factor and to control the update of the noise PSD based on the comparison of the estimated a posteriori SNR to a threshold.

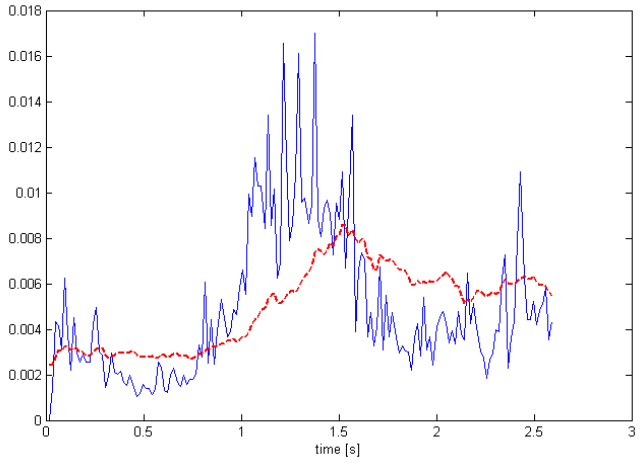

Figure 6. Estimated noise PSD (dashed line) for a frequency bin $k=30$ compared with real noise spectral distribution (continuous line).

$$\hat{\sigma}_v^2(\lambda,k)=\begin{cases}\delta\cdot\hat{\sigma}_v^2(\lambda-1,k)+(1-\delta)|Y(\lambda,k)|^2, & \text{if } \gamma_k(\lambda)<\varepsilon\\ \hat{\sigma}_v^2(\lambda-1,k), & \text{otherwise}\end{cases} \quad (25)$$

If the a posteriori SNR $\gamma_k(\lambda)$ is found to be smaller than a specified threshold ε, suggesting speech absence, then the noise spectrum is updated, else, if the a posteriori SNR is found to be larger than a specified threshold, suggesting speech presence, then the noise spectrum update is stopped.

The threshold ε can have significant effect on the noise spectrum estimation. If ε is chosen too small, then the noise spectrum is not updated often enough and is underestimated. Else, if ε is chosen too large, then the noise spectrum is overestimated. Simulations in [14] showed that choosing $\varepsilon=2.5$ gave a good compromise.

The estimated power of the noise is computed from the noise PSD of each frame:

$$\hat{\sigma}_v^2(\lambda)=\frac{1}{N}\sum_{k=0}^{N-1}\hat{\sigma}_v^2(\lambda,k), \quad (26)$$

where *N* is the number of frequency bins.

In Fig. 6 there are represented the estimated noise PSD (with red-dashed line) for a frequency bin $k=30$ compared with real noise spectral distribution (represented with continuous line) at the same frequency bin.

V. The Proposed Over-subtraction Method

The main drawback of the spectral subtraction method is the residual noise resulted as the difference between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. This residual spectral content manifest themselves in the reconstructed time signal as varying tonal sounds resulting in a musical like noise.

This residual noise can be reduced by a good estimation of the noise PSD and by using the multi-band spectral over-subtraction. Because the human ear sensibility depends nonlinearly on the frequency, a nonlinear frequency spacing approach for multi-band over-subtraction factor estimation is proposed in this paper.

The speech spectrum is divided into *N* non-overlapping bands over the Bark scale of frequency distribution, and spectral subtraction is performed independently in each band. Also, for the noise spectrum estimate, we used the noise PSD estimated $\hat{\sigma}_d^2(\lambda,k)$ at frame λ and frequency k using (21) or (25).

The estimate of the clean speech spectrum in the *i*-th band is obtained by:

$$\left|\hat{S}_i(k)\right|^2 = \left|Y_i(k)\right|^2 - \alpha_i \cdot \hat{\sigma}_d^2(\lambda,k)\,,\ \nu_i < k < \nu_{i+1}\,, \qquad (27)$$

where ν_i and ν_{i+1} are the beginning and ending frequency bins of the *i*-th frequency critical band according to Table 1, and α_i is the over-subtraction factor of the *i*-th critical band.

The over-subtraction factor α_i is a function of the *a posteriori* SNR γ_i of the *i*-th frequency band.

In this paper we propose to use the time-recursive estimation of the *a posteriori* SNR $\gamma_i(\lambda)$ at frame λ for the *i*-th frequency band:

$$\gamma_i(\lambda) = 10\log_{10}\frac{\sum_{k=w_i}^{w_{i+1}}\left|Y(\lambda,k)\right|^2}{\sum_{k=w_i}^{w_{i+1}}\hat{\sigma}_d^2(\lambda,k)}\,. \qquad (28)$$

The over-subtraction factor α_i is calculated for each frequency band as

$$\alpha_i = \begin{cases} 1 & \gamma_i(\lambda) \geq 20\text{dB} \\ 4 - \frac{3}{20}\gamma_i(\lambda) & -6\text{dB} \leq \gamma_i(\lambda) < 20\text{dB}\,. \\ 4.9 & \gamma_i(\lambda) < -6\text{dB} \end{cases} \qquad (29)$$

In the frequency domain of 0-4 kHz, specific for the speech signal, there are 16 critical bands. Experiments showed that there is computational inefficient to separate the spectrum into such a large number of intervals. An equivalent performance is obtained if there are grouped together 3 or 4 critical bands, therefore the spectrum analysis to be performed in a total number of 4 or 5 frequency domains, keeping the nonlinearity frequency spacing given by the human auditory system.

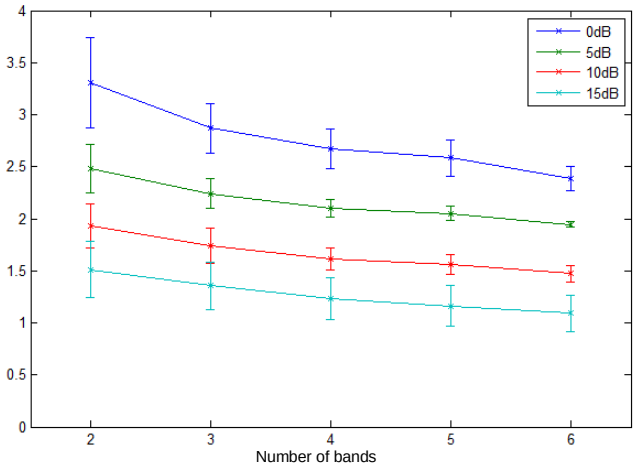

Figure 7. The Itakura-Saito (IS) distance for a variable number of frequency band analysis

VI. Implementation and Experimental Results

The speech signals, sampled at a frequency of 8 kHz, are degraded by the various noise types with segmental SNR's of -5 and 0 dB. Different types of noise taken from the Noisex92 database [15] are added to the speech signal. The noise is chosen with varying spectral distribution simulating colored noise (car engine, factory noise, babble noise).

We used both linear and nonlinear frequency spacing for multi-band spectral over-subtraction, in which the over-subtraction factor α_i is computed as in equation (29) for each frequency band. To determine the number of bands that gives an optimal speech quality, we varied the number of bands from 1 to 8 and examined speech enhancement using both types of frequency spacing.

Objective and subjective quality evaluation methods were applied to establish the performance of the algorithms presented in this study.

The Itakura-Saito (IS) distance method is used as an objective measure to evaluate the performance of the algorithm. The IS measure is based on the similarity or difference between the all-pole model of the clean signal and the corrupted or processed speech signal.

In Fig. 7, the mean IS values are plotted as a function of the number of bands. These values are obtained using the proposed time-recursive estimation for the *a posteriori* SNR used in the over-subtraction factor and a multi-band linear scaled frequency at 0, 5, 10 and 15 dB SNR An improvement in terms of the IS distance can be seen when the number of bands increases from one to six; afterwards, the improvement in quality is no longer perceivable.

Also, we used the ITU-T Recommendation P.862 (PESQ) [16] to obtain a perceptual evaluation of the enhanced speech quality. The Mean Opinion Score (MOS) obtained in the evaluation process is between 0 and 5 where 0 represents a very annoying distortion of the perceived signal and 5 represents imperceptible quality degradation.

Table 2. PESQ MOS EVALUATION FOR THE ENHANCED SPEECH QUALITY

Input SNR	Spectral Subtraction	Standard multi-band spectral subtraction					Time recursive estimation for multi-band spectral subtraction				
No. of bands		*2*	*3*	*4*	*5*	*6*	*2*	*3*	*4*	*5*	*6*
0dB	1.75	1.79	1.82	1.85	1.84	1.84	1.88	1.87	2.00	2.03	2.01
5dB	1.85	1.90	1.99	1.97	1.98	1.97	1.99	2.00	2.07	2.05	2.03
10dB	2.26	2.41	2.44	2.46	2.45	2.40	2.49	2.49	2.55	2.52	2.50
15dB	2.82	2.86	2.88	2.90	2.89	2.84	2.92	2.93	2.95	2.93	2.92

In Table 2 the simulations show that for a single band analysis the results are similar with the standard spectral over-subtraction method. The MOS is increasing when using more than one band, having a maximum when there are four bands, for both methods. Increasing the number of bands more than four bands does not give an increasing of quality since the resolution in frequency analysis is getting worse. A better quality can be noticed for the time recursive estimation of the noise PSD.

Subjective listening tests indicate that, using the multi-band approach and the time-recursive estimation, a very good speech quality with less musical noise and with minimal speech distortion is obtained.

Fig. 8 and Fig. 9 show the spectrograms for speech of speech sentence "The sky this morning was clear and light blue" affected by train noise and car engine noise, at a SNR of 10dB, and the spectrograms of the enhanced speech obtained with standard spectral subtraction, spectral over-subtraction using single-band subtraction factor, multi-band spectral over-subtraction using four linearly-spaced frequency bands, multi-band spectral over-subtraction using four non-linear Bark spaced bands and multi-band spectral subtraction using the time-recursive estimation of the noise PSD.

CONCLUSIONS

This paper presents an improved spectral subtraction method that takes into account the non-uniform effect of colored noise on the speech spectrum. A nonlinear frequency spacing approach for multi-band over-subtraction factor estimation is based on the fact that human ear sensibility varies nonlinear in frequency spectrum. This gives a better perceived quality to the enhanced speech.

Also, time-recursive estimation of the noise PSD is used to compute the multi-band over-subtraction factor in the nonlinear frequency spacing approach. The proposed method reduces the residual musical tones that appear in the case of conventional power spectral subtraction. Simulations with different types of noise show a better quality for the enhanced speech when using time recursive multi-band spectral subtraction.

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

REFERENCES

[1] R. M. Udrea, D. N. Vizireanu, C. C. Oprea, and I. Pirnog, "A time-recursive adaptive algorithm for colored noise reduction in speech enhancement," Sixth Advanced International Conference on Telecommunications AICT 2010, pp.187-190, Barcelona, May 2010.

[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustic, Speech and Signal Processing, vol. 27, Apr. 1979, pp. 113-120.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc., Apr. 1979, pp. 208-211.

[4] C. T. Lin, "Single-channel speech enhancement in variable noise-level environment," Systems, Man and Cybernetics, Part A, IEEE Trans. , Volume: 33 , Issue: 1, Jan. 2003, pp 137–143.

[5] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," Proc. of ICASSP-2002, Orlando, FL, May 2002.

[6] R. M. Udrea and S. Ciochină, "Speech enhancement using spectral oversubtraction and residual noise reduction," Proc. of the Symposium "SCS 2003", Vol II, Iaşi, Romania, July 2003, pp. 165-169.

[7] 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 Issue 5, ISSN: 0165-1684, May 2008, pp. 1299-1303.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. Speech Audio Processing, vol. 11, no 5, pp. 466-475, Sept. 2003.

[9] L. Lin, W.H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," Electronics Letters, vol. 39, no. 9, pp 754–755, May 2003.

[10] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," IEEE Trans. Audio Speech and Language Processing, vol. 16, no. 3, pp. 541–553, March 2008.

[11] D.L. Wang and J.S. Lim, "The unimportance of phase in speech enhancement," IEEE Trans. On Acoustics, Speech, and Signal Processing, vol. 30, no.4, Aug. 1982, pp. 679-681.

[12] B. Moore, "An introduction to the psychology of hearing," 4th ed., London, Academic press,1997.

[13] E. Zwicker and H. Fastl, "Psychoacoustics," 1st ed., Berlin, Springer. 1990.

[14] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," Proc. 20th IEEE Inter. Conf. Acoust. Speech Signal Process., ICASSP-95, Detroit, Michigan, pp. 153-156, 8-12 May 1995.

[15] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Tech. Rep., Speech Research Unit, Defense Research Agency, Malvern, U.K., 1992.

[16] ITU-T, Perceptual evaluation of speech quality PESQ, an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P.862, 2000.

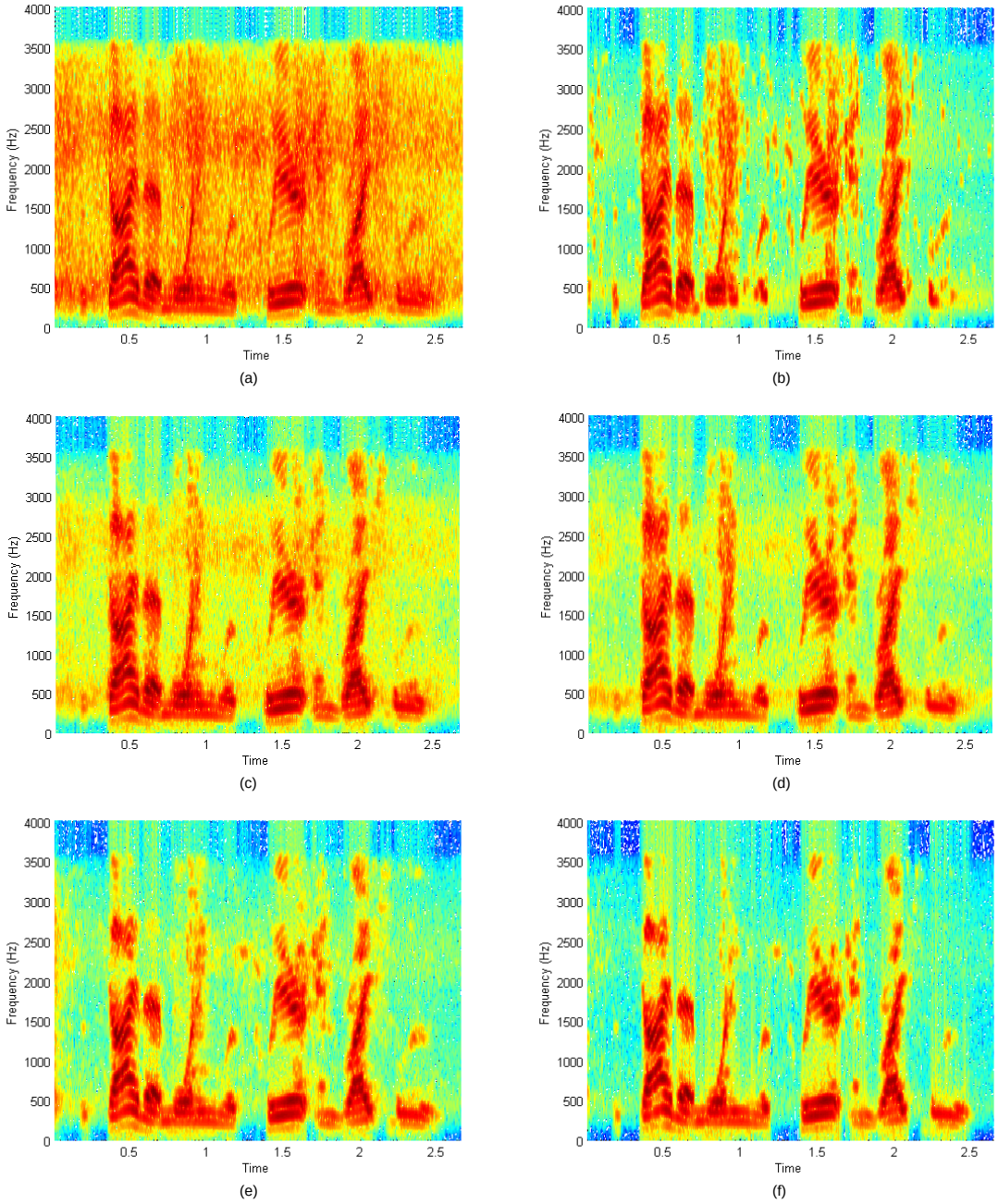

Figure 8. Spectrograms of speech affected by train noise at 10dB SNR (a) and of the enhanced speech obtained with:
standard spectral subtraction (b),
spectral over-subtraction using single-band subtraction factor (c),
multi-band spectral over-subtraction using four linearly-spaced bands (d),
multi-band spectral over-subtraction using four non-linear Bark spaced bands (e),
multi-band spectral subtraction using the time-recursive estimation of the noise PSD (f).

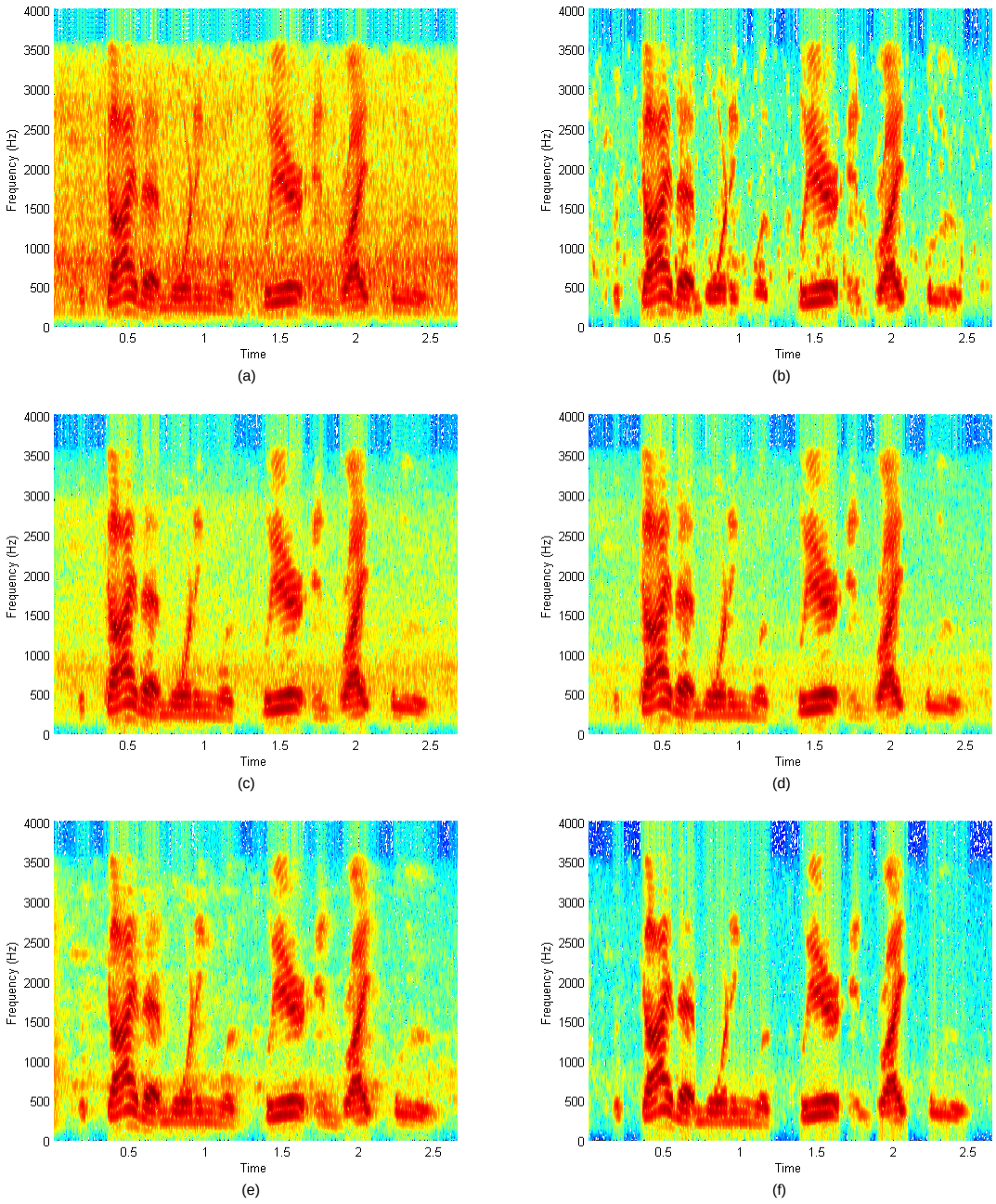

Figure 9. Spectrograms of speech affected by car noise at 10dB SNR (a) and of the enhanced speech obtained with:
standard spectral subtraction (b),
spectral over-subtraction using single-band subtraction factor (c),
multi-band spectral over-subtraction using four linearly-spaced bands (d),
multi-band spectral over-subtraction using four non-linear Bark spaced bands (e),
multi-band spectral subtraction using the time-recursive estimation of the noise PSD (f).