Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System

Nathalie Virag

***Abstract*—This paper addresses the problem of single channel speech enhancement at very low signal-to-noise ratios (SNR's) (<10 dB). The proposed approach is based on the introduction of an auditory model in a subtractive-type enhancement process. Single channel subtractive-type algorithms are characterized by a tradeoff between the amount of noise reduction, the speech distortion, and the level of musical residual noise, which can be modified by varying the subtraction parameters. Classical algorithms are usually limited to the use of fixed optimized parameters, which are difficult to choose for all speech and noise conditions. A new computationally efficient algorithm is developed here based on masking properties of the human auditory system. It allows for an automatic adaptation in time and frequency of the parametric enhancement system, and finds the best tradeoff based on a criterion correlated with perception. This leads to a significant reduction of the unnatural structure of the residual noise. Objective and subjective evaluation of the proposed system is performed with several noise types form the Noisex-92 database, having different time–frequency distributions. The application of objective measures, the study of the speech spectrograms, as well as subjective listening tests, confirm that the enhanced speech is more pleasant to a human listener. Finally, the proposed enhancement algorithm is tested as a front-end processor for speech recognition in noise, resulting in improved results over classical subtractive-type algorithms.**

***Index Terms*—Auditory properties, masking, noise reduction, speech recognition, subtractive-type algorithms.**

I. Introduction

AUTOMATIC speech processing systems are employed more and more often in new applications in a variety of real environments. However, in many practical situations, they are confronted with high ambient noise levels, and their performance degrades drastically. Thus, there is a strong need to improve the performance of these systems in noisy conditions by developing enhancement algorithms able to work at very low signal-to-noise ratios (SNR's).

In this paper, we focus on single channel speech enhancement. This is the most difficult task, since the noise and the speech are in the same channel. In this case, noise is usually estimated during speech pauses. The proposed enhancement system is based on a well-known family of enhancement algorithms: *subtractive-type algorithms*. They attempt to estimate the short-time spectral magnitude of speech by subtracting a noise estimation from the noisy speech. The phase of the noisy speech is not processed, based on the assumption that phase distortion is not perceived by the human ear. Short-time spectral magnitude estimation is a basic technique for many speech enhancement algorithms. In addition to the basic approach of spectral magnitude subtraction [1], many variations have been developed [2]. Subtractive-type algorithms constitute a traditional approach for removing stationary background noise in single channel systems. This type of processing has been chosen for of its simplicity of implementation. Additionally, it offers a high flexibility in terms *subtraction parameters* variation. However, this method needs to be improved since it has a major drawback: the introduction in the enhanced speech of a "musical residual noise" with an unnatural structure. This perceptually annoying noise is composed of tones at random frequencies and has an increased variance. Various solutions have been proposed to reduce this effect: magnitude averaging [1], oversubtraction of noise and introduction of a spectral floor [3], soft-decision noise suppression filtering [4], optimal MMSE estimation of the short-time spectral amplitude [5], nonlinear spectral subtraction [6], and introduction of morphological-based spectral constraints [7]. However, experiments performed with these different subtractive-type algorithms show that there is a need for further improvement, especially at very low SNR's. Indeed, in this case, it is difficult to suppress noise without decreasing intelligibility and without introducing speech distortion and residual noise. In very noisy situations, the enhanced speech can be even more disturbing to a human listener than the original corrupted speech signal.

The solution proposed in this paper works toward rendering this residual noise "perceptually white." This is done by introducing knowledge on human perception in the enhancement process. Some methods have been recently developed in this direction, by modeling several aspects of the enhancement mechanism present in the auditory system, resulting in promising results [8]–[10]. However, few existing enhancement algorithms take into account such auditory properties, and still few work at very low SNR's. In this paper, it is proposed to incorporate a human hearing model that is already widely used in wideband audio coding [11]. This model is based on the masking phenomenon. It is related to the concept of critical band analysis, which is a central analysis mechanism in the inner ear. The masking properties are modeled by calculating a *noise masking threshold*. A human listener tolerates additive noise as long as it remains below this threshold. In the enhancement process, the subtraction

Manuscript received November 7, 1996; revised July 9, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

The author is with the Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland (e-mail: nathalie.virag@epfl.ch).

Publisher Item Identifier S 1063-6676(99)01630-2.

parameters are adapted based on this noise masking threshold. This allows one to find the best tradeoff between the amount of noise reduction, the speech distortion and the level of residual noise *in a perceptual sense*.

The objective of the proposed enhancement system, which exploits the masking properties of the human auditory system, is to overcome the limitations of one channel subtractive-type enhancement systems in additive background noise at very low SNR's (<10 dB). The paper is organized as follows. In Section II, the principles of subtractive-type algorithms are described. In Section III, the proposed enhancement algorithm is presented. Finally, an objective and subjective evaluation is performed in Section IV.

II. Subtractive-Type Algorithms

Consider a speech signal $s(n)$ corrupted by additive stationary background noise $d(n)$. The noisy speech can be expressed as follows:

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

The processing is done on a frame-by-frame basis in the frequency domain. Speech and noise are assumed to be uncorrelated. The enhanced speech short-time magnitude $|\hat{S}(\omega)|$ is obtained by subtracting from the noisy speech short-time magnitude $|Y(\omega)|$ a noise spectral magnitude estimate $|\hat{D}(\omega)|$ computed during speech pauses. For the particular case of power spectral subtraction (PSS), the subtraction of the noise estimate is expressed as follows:

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{D}(\omega)|^2, & \text{if } |Y(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $|\hat{D}(\omega)|^2$ represents the noise power spectrum estimate. The phase of the noisy speech is not modified. Therefore, the best result achievable with this method is given by the combination of the clean speech spectral magnitude with the noisy phase. This is called in this paper the *theoretical limit*. Once the subtraction is computed in the spectral domain with (2), the enhanced speech signal is obtained with the following relationship:

$$\hat{s}(n) = \text{IFFT}[|\hat{S}(\omega)| \cdot e^{j \arg Y(\omega)}]. \tag{3}$$

A. Suppression Curves

Subtractive-type algorithms can be studied using a second approach: *filtering of noisy speech* with a time-varying linear filter dependent on the characteristics of the noisy signal spectrum and on the estimated noise spectrum. The noise suppression process becomes a multiplication of the short-time spectral magnitude of the noisy speech $|Y(\omega)|$ by a gain function $G(\omega)$:

$$|\hat{S}(\omega)| = G(\omega) \cdot |Y(\omega)| \quad \text{with} \quad 0 \le G(\omega) \le 1. \tag{4}$$

The filter for PSS corresponding to (2) is given by

$$G(\omega) = \sqrt{1 - \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}}. \tag{5}$$

If $|\hat{D}(\omega)|^2 > |Y(\omega)|^2$ the gain is set to $G(\omega) = 0$, which ensure that the gain is always real. This formulation allows for a better understanding of the procedure of short-time magnitude modification: subtractive-type systems attempt to emphasize the spectral components as a function of the amount by which they exceed noise. Therefore, the gain function $G(\omega)$ changes between noise sections and speech sections: sections containing only speech are unmodified ($G(\omega) = 1$), while sections containing only noise are suppressed ($G(\omega) = 0$). Between these two extreme cases, the gain function takes a value depending on the *a posteriori* SNR:

$$\text{SNR}_{\text{post}}(\omega) = \frac{|Y(\omega)|^2}{|\hat{D}(\omega)|^2}. \tag{6}$$

Each gain function corresponds to a given subtraction rule and can be represented by a *suppression curve*. This representation allows one to observe the attenuation as a function of the $\text{SNR}_{\text{post}}(\omega)$, which can be easily measured on the noisy speech. This attenuation ranges from $-\infty$ dB (maximal attenuation) to 0 dB (no processing). There exist various subtraction rules derived from different criteria [2]. Most existing algorithms have a parametric form, which allows for a greater flexibility in the variation of the suppression curves. These algorithms have been mainly developed to reduce the musical residual noise present in subtractive-type systems. In this paper, the generalized spectral subtraction approach has been chosen due to its flexibility and because it includes most of the basic subtraction rules.

B. Generalized Spectral Subtraction

Spectral subtraction cannot be satisfying without a residual noise reduction process. The most famous algorithm for residual noise reduction has been proposed by Berouti *et al*. in [3]. This method can be combined with the generalized spectral subtraction scheme described in [2], leading to the following gain function:

$$\begin{aligned} G(\omega) &= G[\text{SNR}_{\text{post}}(\omega)] \\ &= \begin{cases} \left(1 - \alpha \cdot \left[\dfrac{|\hat{D}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2}, \\ \qquad \text{if } \left[\dfrac{|\hat{D}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1} < \dfrac{1}{\alpha+\beta} \\ \left(\beta \cdot \left[\dfrac{|\hat{D}(\omega)|}{|Y(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2}, \\ \qquad \text{otherwise.} \end{cases} \end{aligned} \tag{7}$$

This is one of the most flexible form of subtractive-type algorithm. Furthermore, it takes into account the classical subtraction rules. It allows for a variation of the tradeoff between noise reduction, residual noise and speech distortion with the variation of the free parameters of (7):

1) *Oversubtraction factor* $\alpha(\alpha > 1)$: the short-time spectrum is attenuated more than necessary. This leads to a reduction of residual noise peaks but also to an increased

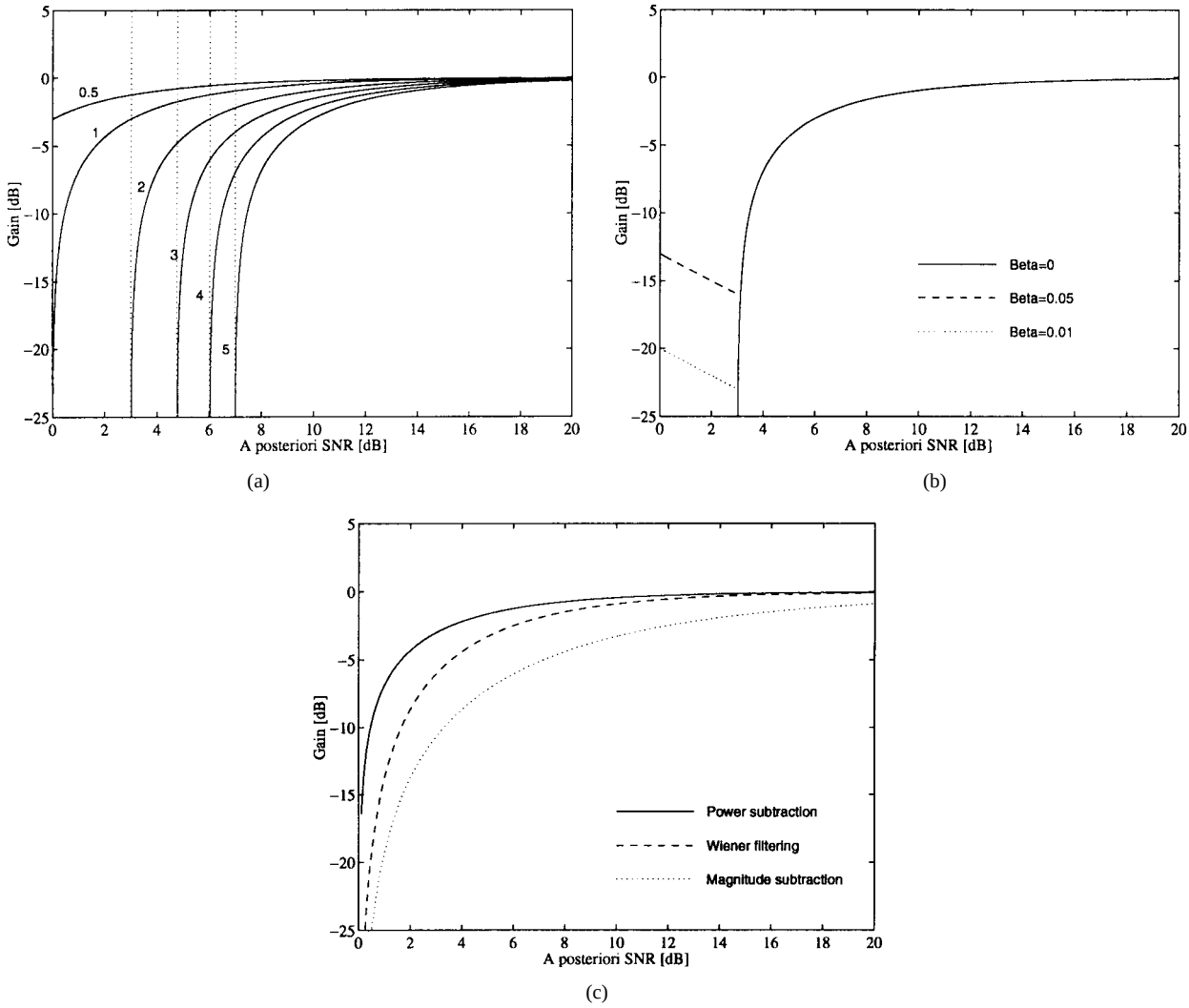


Fig. 1. Suppression curves for generalized spectral subtraction: (a) variations of α (with $\beta = 0$ and $\gamma = 2$); (b) variations of β (with $\alpha = 2$ and $\gamma = 2$); (c) variations of γ (with $\alpha = 1$ and $\beta = 0$). The classical methods are: magnitude subtraction ($\gamma_1 = 1$ and $\gamma_2 = 1$), power spectral subtraction ($\gamma_1 = 2$ and $\gamma_2 = 0.5$) and Wiener filtering ($\gamma_1 = 2$ and $\gamma_2 = 1$).

audible distortion. This suppression factor determines the shift, from left to right, of the suppression curve, as shown in Fig. 1(a).

2) *Spectral flooring* $\beta(0 \leq \beta \ll 1)$: addition of background noise in order to mask the residual noise. This leads to a reduction of residual noise but an increased level of background noise remains in the enhanced speech. This factor determines the minimum value taken by the gain function, as shown in Fig. 1(b).
3) *Exponent* $\gamma = \gamma_1 = 1/\gamma_2$: determines the sharpness of the transition from the $G(\omega) = 1$ (the spectral component is not modified) to the $G(\omega) = 0$ (the spectral component is suppressed). The suppression curves obtained for the classical values of the exponent are shown in Fig. 1(c). However, the choice for the value of γ is not as critical as that of α and β.

The choice of the subtraction parameters α, β, and γ is a central notion in single channel speech enhancement. Indeed, at low SNR's, it is impossible to simultaneously minimize speech distortion and residual noise. In our case, we are concerned with reducing noise and increasing intelligibility, while keeping the residual noise and the distortion acceptable to a human listener. This is done by adapting the subtraction parameters α and β in time and frequency based on masking properties.

III. Speech Enhancement Based on Masking Properties

The proposed enhancement method is based on the main conclusion derived from a comparative study of subtractive-type algorithms: algorithms with fixed subtraction parameters are unable to adapt to varying noise levels and characteristics. Furthermore, the optimization of the parameters is a difficult task. An example of adaptation is however proposed by the nonlinear spectral subtraction algorithm (NSS), which adapts the oversubtraction factor α in time and frequency based on the SNR, leading to improved results [6]. In the present algorithm, an *adaptation of the parameters* is also performed, but based on the concept of *noise masking*.

Masking is present because the auditory system is incapable of distinguishing two signals close in the time or frequency

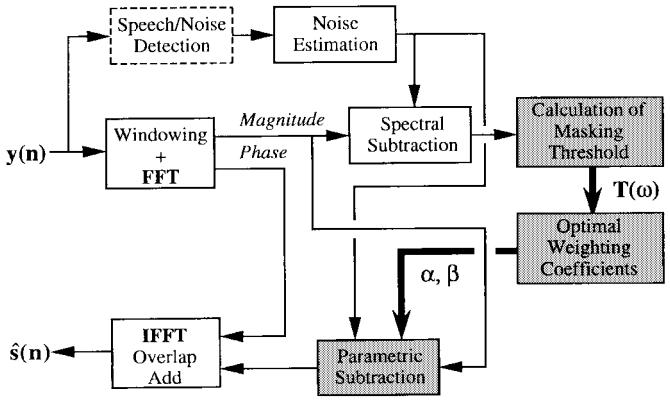


Fig. 2. The proposed perceptual enhancement scheme.

domain. This is manifested by an elevation of the minimum threshold of audibility due to a masker signal. Noise masking is a well-known psychoacoustical property of the auditory system that has already been applied with success to speech and audio coding in order to partially or totally mask the distortion introduced in the coding process [11]. This work only considers the frequency domain masking, or simultaneous masking: a weak signal is made inaudible by a stronger signal occurring simultaneously. This phenomenon is modeled via a noise masking threshold, below which all components are inaudible.

The proposed enhancement scheme is presented in Fig. 2. It is composed of the following main steps.

1) Spectral decomposition.
2) Speech/noise detection and estimation of noise during speech pauses.
3) Calculation of the noise masking threshold $T(\omega)$.
4) Adaptation in time and frequency of the subtraction parameters α and β based on the noise masking threshold $T(\omega)$.
5) Calculation of the enhanced speech spectral magnitude via parametric subtraction with adapted parameters α and β in (7).
6) Inverse transform (3).

After several experiments, the exponent is set to $\gamma = 2$.

A. Noise Masking Threshold Calculation

The noise masking threshold $T(\omega)$ is obtained through modeling the frequency selectivity of the human ear and its masking property. The different calculation steps are summarized in [11] and [12].

1) *Frequency analysis along a critical band scale, or Bark scale [13]:* This critical band analysis is performed on the fast Fourier transform (FFT) power spectrum by adding up the energies in each critical band k, according to the values given in Table I.
2) *Convolution with a spreading function $SF(k)$ to take into account masking between different critical bands*: The function used in this work has been proposed by Schroeder *et al*. in [14] and is represented in Fig. 3(a).
3) *Subtraction of a relative threshold offset $O(k)$ depending on the noise-like or tone-like nature of the masker an the maskee*: A simple method has been proposed by Sinha and Tewfik in [15], which avoids an accurate estimate of the tonality and therefore reduces the computational load: $O(k)$ is given by a simple estimation, based on the fact that the speech signal has a tonelike nature in lower critical bands and a noiselike nature in higher bands. The resulting values for $O(k)$ are represented in Fig. 3(b).
4) *Renormalization and comparison with the absolute threshold of hearing [11].*

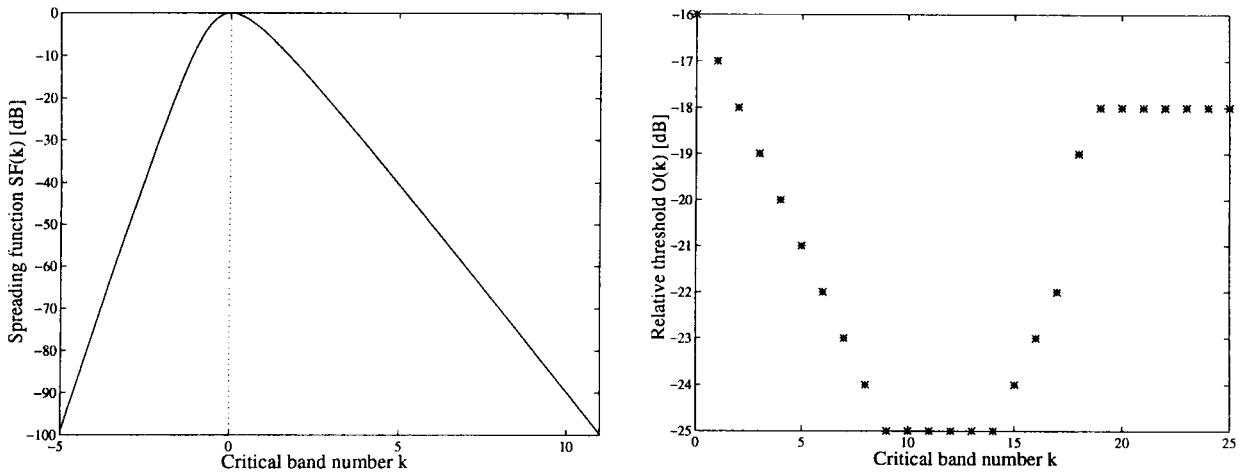
TABLE I
MAPPING FROM FFT BINS TO CRITICAL BANDS AT A SAMPLING FREQUENCY OF 8 kHz AND A FRAME SIZE $N = 256$

Critical band number k	*FFT bins*		*Real frequencies*
	Intervals	*Number of bins*	*[Hz]*
1	1–3	3	0–94
2	4–6	3	94–187
3	7–10	4	187–312
4	11–13	3	312–406
5	14–16	3	406–500
6	17–20	4	500–625
7	21–25	5	625–781
8	26–29	4	781–906
9	30–35	6	906–1094
10	36–41	6	1094–1281
11	42–47	6	1281–1469
12	48–55	8	1469–1719
13	56–64	9	1719–2000
14	65–74	10	2000–2312
15	75–86	12	2312–2687
16	87–100	14	2687–3125
17	101–118	18	3125–3687
18	119–128		3687–4000

An example of noise masking threshold for a given speech frame is represented in Fig. 4, converted to a frequency scale. The sampling frequency for this example is 16 kHz, therefore, the total number of critical bands is $K = 22$. In the method described above, the noise masking threshold is computed from the clean speech signal. However, in the proposed enhancement scheme, only the noisy signal is available. Therefore this threshold has to be estimated in noise. As the threshold calculation is difficult to perform on the noisy speech directly, a rough estimate of the clean speech signal is computed with a simple power spectral subtraction scheme (see block diagram in Fig. 2), reducing background noise but introducing musical residual noise, especially at low SNR's. This residual noise modifies the tonality of the signal and the masking threshold is slightly different from the one obtained from the clean speech, especially for high frequencies (if the critical band number $k > 15$). This is represented in Fig. 4(a). Consequently, the relative threshold offset, represented for clean speech in Fig. 3(b), has to be decreased for $k > 15$ to take into account the tonelike nature of the musical residual noise. Indeed, for tonelike signals, $O(k)$ is lower than for noiselike signals and it is dependent of k as described in [15]. This modification leads to the improved result presented in Fig. 4(b).

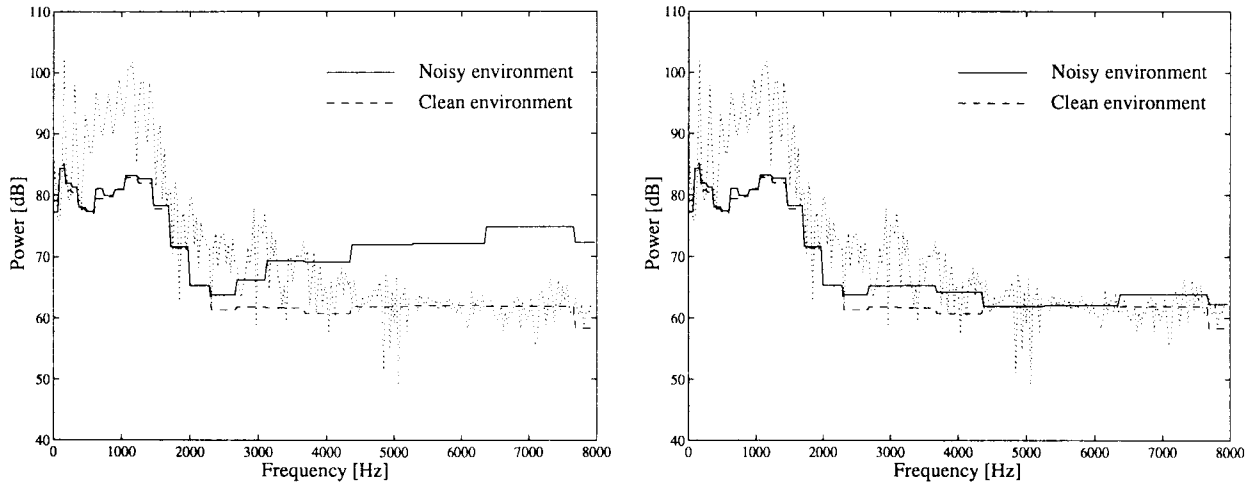
B. Adaptation of the Subtraction Parameters

The adaptation in time (for each frame m) and frequency (for each critical band k) of the subtraction parameters α



(a) (b)

Fig. 3. Functions used for the noise masking threshold calculation: (a) spreading function [14]; (b) relative threshold offset [15].



(a) (b)

Fig. 4. Example of noise masking threshold $T(\omega)$ in white Gaussian noise at a SNR = 0 dB, for a 32 ms section of the English word "one." (a) Original method. (b) Modified method.

and β is based on the masking threshold $T(\omega)$. Residual noise can be reduced by increasing the parameters α and β. However, this leads to more speech distortion and remaining background noise. The best solution would be to choose the enhancement parameters in such a way that the residual noise stays below the masking threshold of the auditory system. This would ensure that the residual noise is masked, and remains inaudible. This approach is well adapted for high input SNR's (>10 dB). However, if the noise level increases, the masking threshold is too low to completely mask the residual noise without increasing the speech distortion, leading to a synthetic sound. Therefore, the proposed adaptation is based on the following consideration: if the masking threshold is high, residual noise will be naturally masked and inaudible. Hence, there is no need to reduce it in order to keep distortion as low as possible. In this case, the subtraction parameters are kept to their minimal values. However, if the masking threshold is low, residual noise will be annoying to the human listener and it is necessary to reduce it. This is done by increasing the subtraction parameters. For each frame m, the minimum of the masking threshold $T_m(\omega)$ corresponds to the maxima of the parameters $\alpha_m(\omega)$ and $\beta_m(\omega)$. The adaptation of the subtraction parameters is performed with the following relations:

$$\alpha_m(\omega) = F_\alpha[\alpha_{\min}, \alpha_{\max}, T(\omega)] \tag{8}$$

$$\beta_m(\omega) = F_\beta[\beta_{\min}, \beta_{\max}, T(\omega)] \tag{9}$$

where $\alpha_{\min}$, $\beta_{\min}$, and $\alpha_{\max}$, $\beta_{\max}$ are the minimal and maximal values of the oversubtraction and spectral flooring. F_α and F_β are functions leading to a maximal residual noise reduction for the minimal values of the masking threshold and a minimal reduction for the maximal values of the threshold: $F_\alpha = \alpha_{\max}$ if $T(\omega) = T(\omega)_{\min}$ and $F_\alpha = \alpha_{\min}$ if $T(\omega) = T(\omega)_{\max}$, where $T(\omega)_{\min}$ and $T(\omega)_{\max}$ are the minimal and maximal values of the masking threshold updated from frame to frame. The values of F_α between these two extreme cases are interpolated based on the value of $T(\omega)$. Similar considerations can be made for F_β. In order to avoid

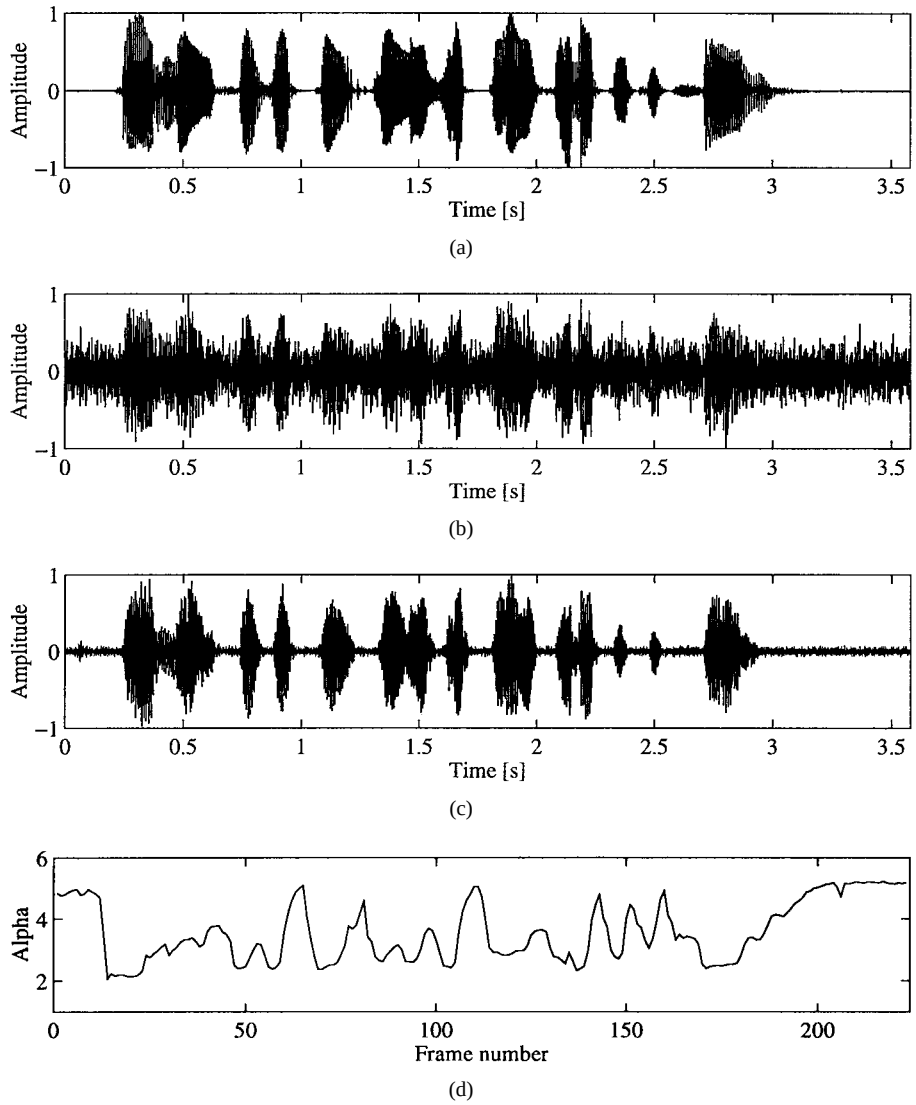


Fig. 5. Example of adaptation of the parameter α in the proposed enhancement algorithm. (a) Original clean speech signal in French: "Un loup s'est jeté immédiatement sur la petite chèvre." (b) Noisy signal (additive speechlike noise at a SNR = 0 dB). (c) Enhanced signal. (d) Mean value of the vector $\alpha(\omega)$ for each frame.

discontinuities in the gain function $G(\omega)$ due to this adaptation, a smoothing operation is applied.

A number of studies and experiments with different noise types and levels have been performed to select the appropriate values for these parameters. The following values have been chosen to obtain a good tradeoff between residual noise and speech distortion for a human listener:

1) $\alpha_{\min} = 1$ and $\alpha_{\max} = 6$;
2) $\beta_{\min} = 0$ and $\beta_{\max} = 0.02$;
3) the exponent is kept fixed to $\gamma = \gamma_1 = 2$ and $\gamma_2 = 1/\gamma_1 = 0.5$.

This tradeoff can easily be changed according to the application. A reduction of $\alpha_{\max}$ increases the residual noise but reduces the speech distortion, while a reduction of $\beta_{\max}$ increases also the residual noise but reduces the background noise which remains in the enhanced speech. An example of adaptation of the parameter α is given in Fig. 5 The adaptation curve for β has a similar shape, with different minimal and maximal values. This adaptation can be compared to the one performed in the NSS algorithm, where β is kept fixed and α is dependent on the SNR [6]. The advantage of the proposed method over NSS lies on the use of the noise masking threshold $T(\omega)$ rather than the SNR.

- The estimation of the noise masking threshold $T(\omega)$ in noisy environments is more accurate than the SNR estimation because $T(\omega)$ is computed from a rough initial estimate of the clean speech obtained using a classical subtractive-type technique.
- $T(\omega)$ has a smoother evolution than the SNR, which is important for residual noise reduction. This is due to the critical band analysis, that performs a perceptually relevant smoothing which is not limited by the nonstationarity of speech.
- The adaptation based on $T(\omega)$ is better correlated with perception than using the SNR.

C. Speech/Noise Detection and Noise Estimation

If the additive noise is stationary or slowly varying, a noise estimate $|\hat{D}(\omega)|^{\gamma}$ can be updated during speech pauses with

TABLE II
COMPARED ALGORITHMS (PARAMETER $\gamma = 2$)

Enhancement method	*Parameter* α		*Parameter* β	
Power spectral subtraction	Fixed	1	Fixed	0
Modified power spectral subtraction	Fixed	Optimized	Fixed	Optimized
Nonlinear spectral subtraction	Adapted	F$[SNR(\omega)]$	Fixed	0.01
Proposed algorithm	Adapted	F$[T(\omega)]$	Adapted	F$[T(\omega)]$

the following averaging rule:

$$|\hat{D}_m(\omega)|^\gamma = \lambda_D \cdot |\hat{D}_{m-1}(\omega)|^\gamma + (1-\lambda_D)\cdot|\hat{Y}_m(\omega)|^\gamma \quad \text{with} \quad 0.5 \le \lambda_D \le 0.9 \tag{10}$$

where m represents the frame index and λ_D is a forgetting factor, which has to be chosen depending on the stationarity of noise. For this work, $\lambda_D = 0.9$, leading to an averaging of about 20 frames (320 ms). Speech pauses are detected with the energy-based adaptive detection algorithm proposed by Lynch *et al*. in [16]. This detector develops speech and noise metrics based on statistical assumptions about the characteristics of speech (production rule) and noise waveforms. It performs an adaptive thresholding based on these metrics. It is easy to implement and has originally been proposed to work on a sample-by-sample basis. However, for the present use in a speech enhancement algorithm, it has been modified in order to indicate whether a given frame contains speech or noise. This detector has been tested in various noisy environments and introduced in the proposed enhancement scheme, showing acceptable results in background noise levels up to SNR $= 0$ dB. Indeed, at low SNR's, the noise estimate is not significantly influenced by the low level speech still present in the frame.

IV. PERFORMANCE EVALUATION

This section presents the performance evaluation of the proposed enhancement algorithm, as well as a comparison with other subtractive-type algorithms. The sampling frequency is 8 kHz. Based on this value, the following parameters have been chosen: 1) frame size $N = 256$ (32 ms) with 50% overlap; 2) Hanning window; 3) total number of critical bands $K = 18$.

Noise signals have different time–frequency distributions, and therefore a different impact on speech. Six different background noises were taken from the Noisex-92 database, designed for speech recognition in noisy environments: white Gaussian noise, car noise, speechlike noise (long term average speech spectrum), aircraft cockpit noise, helicopter cockpit noise, and factory noise. Phonetically balanced speech sentences were extracted from the BDSons database (in French). Both noise and speech databases have been recorded at a sampling frequency of 16 kHz. For our enhancement experiments, the signals were downsampled to 8 kHz. Noise has been added to the clean speech signal with a varying SNR. Several classical methods are compared to the proposed algorithm: 1) power spectral subtraction; 2) modified spectral subtraction (values of α and β are kept fixed, but have been optimized for each noisy speech sentence); 3) nonlinear spectral subtraction. The corresponding subtraction parameters are indicated in Table II. The best performance achievable with subtractive-type algorithms, named theoretical limit, is also taken into account.

Generally, the objective performance evaluation is based on the application of *objective quality or intelligibility measures*. The major drawback of these measures is the fact they are not always well correlated with speech perception [17]. Furthermore, they do not give information about how speech and noise are distributed across frequency. One channel subtractive-type enhancement systems produce generally two main undesirable effects: residual noise and speech distortion. These effects can be annoying to a human listener, but they are difficult to quantify with the help of these objective measures. It is therefore important to analyze the time-frequency distribution of the enhanced speech, in particular the structure of its residual noise. In this work, this is done by observing the *speech spectrograms*, which give a more accurate information about residual noise and speech distortion than the corresponding time waveforms. Finally, in order to validate the objective performance evaluation, *subjective tests* are performed with the different subtractive-type algorithms. The proposed enhancement technique is also tested as a front-end for *speech recognition in noise*.

A. SNR Improvement

The amount of noise reduction is generally measured with the SNR improvement, given by the difference between input and output segmental SNR:

$$G_{\text{SNR}} = \frac{1}{L}\sum_{m=0}^{L-1} 10\cdot \log \frac{\dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1} d^2(n+Nm)}{\dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1}[s(n+Nm)-\hat{s}(n+Nm)]^2} \quad \text{[dB]} \tag{11}$$

where L represents the number of frames in the signal and N the number of samples per frame. This equation takes into account both residual noise and speech distortion. Fig. 6 shows the SNR improvement obtained for various noise types and at various noise levels. The SNR improvements provided by the compared subtractive-type algorithms are similar, although the proposed algorithm produces an increased SNR improvement for low input SNR's. The best noise reduction is obtained in the case of white Gaussian noise, while for colored noise this improvement decreases.

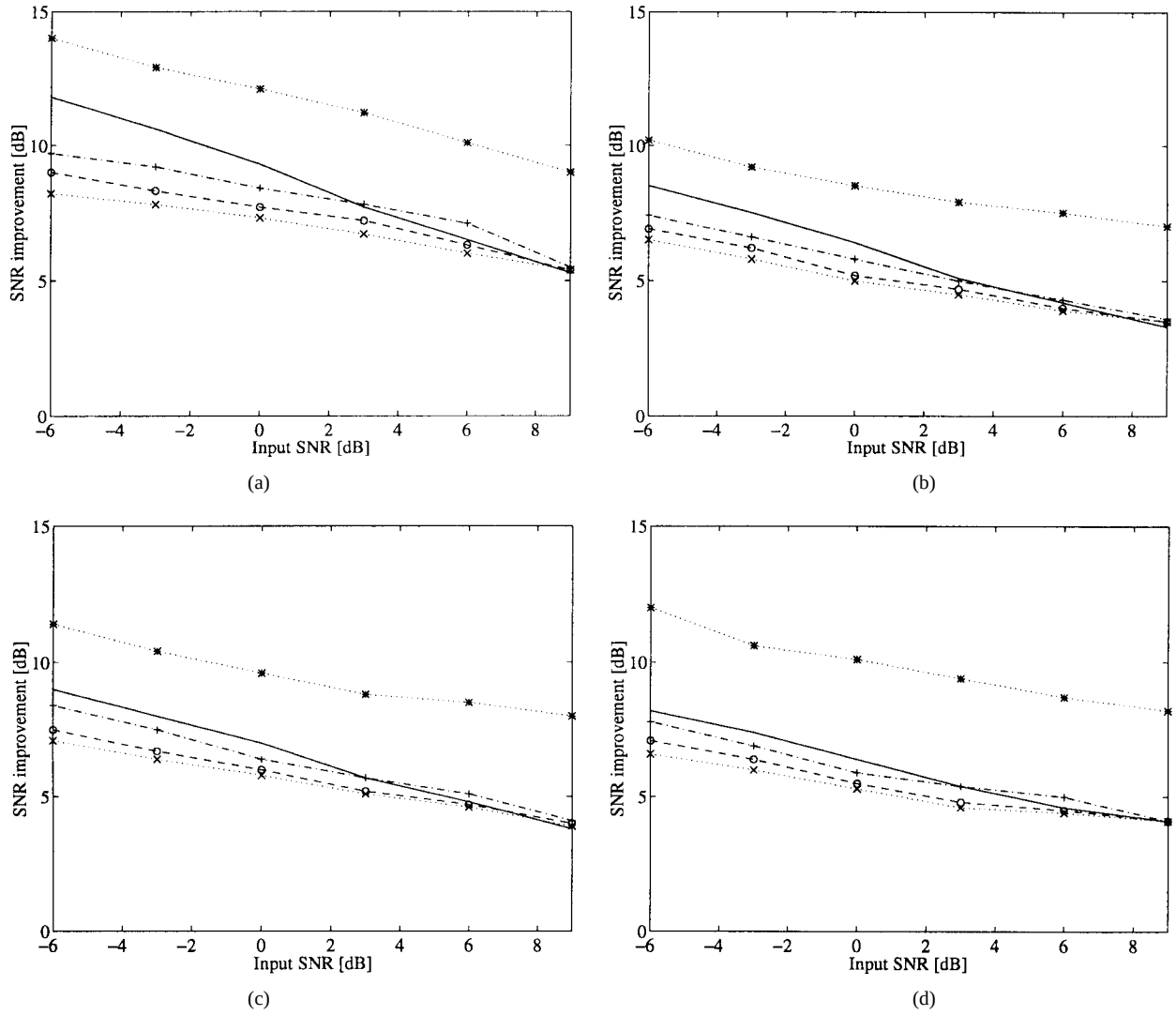


Fig. 6. SNR improvement for various noise types and levels: (a) white Gaussian noise; (b) speechlike noise; (c) aircraft cockpit noise; (d) factory noise. The tested methods are the following: (×) power spectral subtraction, (+) modified spectral subtraction, (∘) nonlinear spectral subtraction, (*) theoretical limit, (solid line) proposed enhancement algorithm.

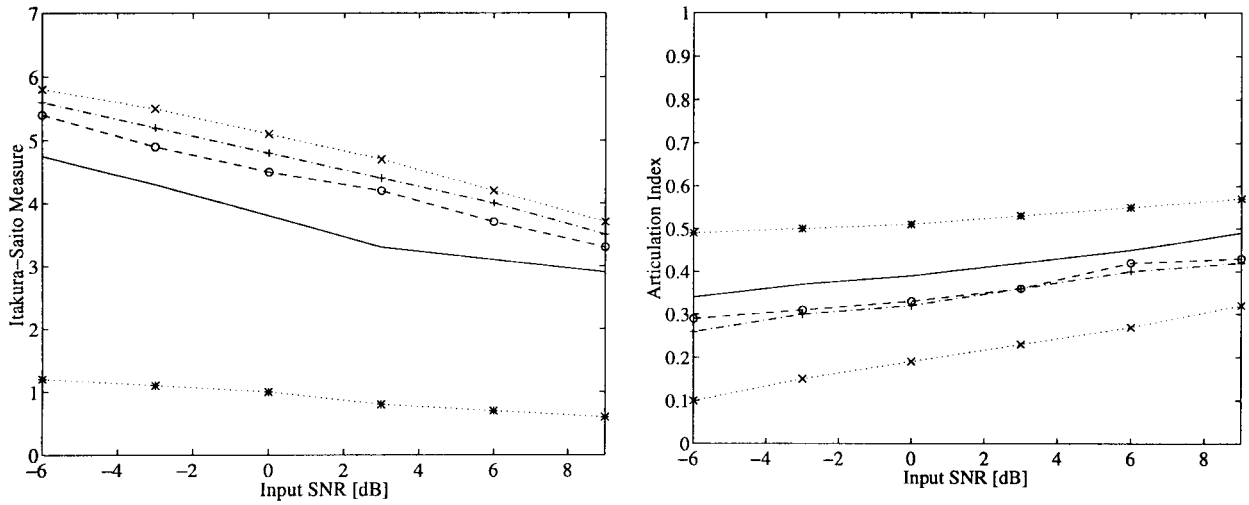
The main drawback of the SNR is the fact that it has a very poor correlation with subjective results [17]. Experiments show that, even though the SNR's are very similar at the output of the enhancement system, the listening test and speech spectrograms can lead to very divergent results (different residual noise and speech distortion). Therefore, the segmental SNR is not a sufficient objective indicator of speech quality and the performance evaluation needs to be completed with further tests.

B. Objective Measures

Various objective measures providing a higher correlation with subjective results than the SNR have been developed to evaluate the performance of speech processing systems. In this paper, the following two measures have been chosen.

1) *Itakura–Saito distortion (IS)*: objective quality measure that performs a comparison between spectral envelopes (all-pole parameters) and that is more influenced by a mismatch in formant location than in spectral valleys [18]. A typical range for the IS measure is 0–10, where the minimal value of IS corresponds to the best speech quality. A correlation between the IS and subjective quality measures is given in [17].
2) *Articulation index (AI)*: objective intelligibility measure based on a decomposition of the speech signal according to a Bark scale. This measure has been standardized and the computation steps are described in [19]. The AI is in the range 0 to 1, where the maximal value of the AI corresponds to the best intelligibility score. A relation between the AI and subjective intelligibility tests is also given in [19].

Table III presents a comparison of IS measure and the AI at a SNR $= 0$ dB for different noises. The worst results are obtained with speechlike noise. Indeed, this noise is particularly difficult to handle, because it has the same frequency distribution as long-term speech. We will therefore focus the study of the enhancement results on this type of noise. Values of the IS measure and AI in various noise levels and for different enhancement algorithms are presented in



(a) (b)

Fig. 7. Example of objective measures for speechlike noise at varying levels: (a) Itakura–Saito measure; (b) Articulation index. The tested methods are the following: (×) noisy signal, (+) modified spectral subtraction, (∘) nonlinear spectral subtraction, (*) theoretical limit, (solid line) proposed enhancement algorithm.

TABLE III
OBJECTIVES MEASURES OBTAINED WITH THE PROPOSED ALGORITHM FOR THE VARIOUS NOISE TYPES AT AN INPUT SNR = 0 dB

Noise type	*Noisy signal*			*Enhanced signal*		
	SNR	*IS*	*AI*	*SNR*	*IS*	*AI*
White Gaussian noise	0 dB	4.2	0.11	9.2 dB	2.3	0.35
Car noise	0 dB	3.8	0.42	9.5 dB	1.9	0.50
Speech-like noise	0 dB	5.1	0.19	6.3 dB	3.9	0.39
Aircraft cockpit noise	0 dB	3.6	0.17	6.9 dB	2.2	0.38
Helicopter cockpit noise	0 dB	2.6	0.21	7.7 dB	1.6	0.41

Fig. 7. For both measures, the proposed algorithm achieves a significant improvement over classical subtractive-type algorithms.

C. Speech Spectrograms

Objective measures do not give indications about the structure of the residual noise. Speech spectrograms constitute a well-suited tool for observing this structure. All the speech spectrograms presented in this paper use a Hanning window of 128 samples with an overlap of 90 samples. An example of spectrogram for clean and noisy speech is presented in Fig. 8 for speechlike noise at SNR = 0 dB. Fig. 8 also presents the spectrogram for the theoretical limit. Experiments show that the noisy phase is not perceived as long as the local SNR is greater than about 6 dB. However, at a SNR = 0 dB, the effect of the noisy phase is audible and is perceived as an *increased roughness*. Fig. 9 shows the spectrograms obtained with the proposed algorithm and other classical subtractive-type algorithms. In the proposed algorithm, the musical structure of the residual noise is better reduced, even compared to NSS, which leads to the best results for classical algorithms. Speech enhanced with the proposed method is more pleasant and the residual noise has a "perceptually white quality" while distortion remains acceptable. This confirms the values of the IS measure and the AI of Fig. 7 and it is validated by informal listening tests.

If we observe the enhanced speech obtained with other noise types at SNR = 0 dB, we can see that the musical structure of residual noise is also greatly reduced, as long as the noise remains stationary. This is shown in Figs. 10–12. As the nonstationarity of noise increases, results become very poor, because in this case, the noise estimate cannot follow the variations of the background noise. This limitation inherent to single channel system is visible for factory noise in Fig. 12, where the musical structure of residual noise is increased. For SNR's greater than 10 dB, the performance of the proposed algorithm becomes comparable to the one provided by classical systems with a residual noise reduction algorithm such as in [3]–[5].

D. Subjective Evaluation

In order to validate the objective performance evaluation, subjective listening tests have been performed with the different subtractive-type algorithms in speechlike noise at SNR = 0 dB. The speech spectrograms of the test signals are presented in Figs. 8 and 9. The listening tests have been realized with ten listeners. Each listener gives for each test signal a score between one and five. This score represent his global appreciation of the residual noise, the background noise still present and the speech distortion. The scale used for these tests correspond to the MOS scale presented in [18]. The test signals have been recorded on a DAT tape, and headphones have been used during experiments. For each speaker, the following procedure has been applied: 1) clean speech and noisy speech are played and repeated twice; 2) each test signal, which is repeated twice for each score, is played three times in a random order. This leads to 30 scores for each test signal. The results are presented in Table IV. The values obtained are well suited for ranking the performance of the different tested methods. Subjective listening tests confirm that the proposed enhancement method leads to the best result for a human listener compared to other subtractive-type algorithms.

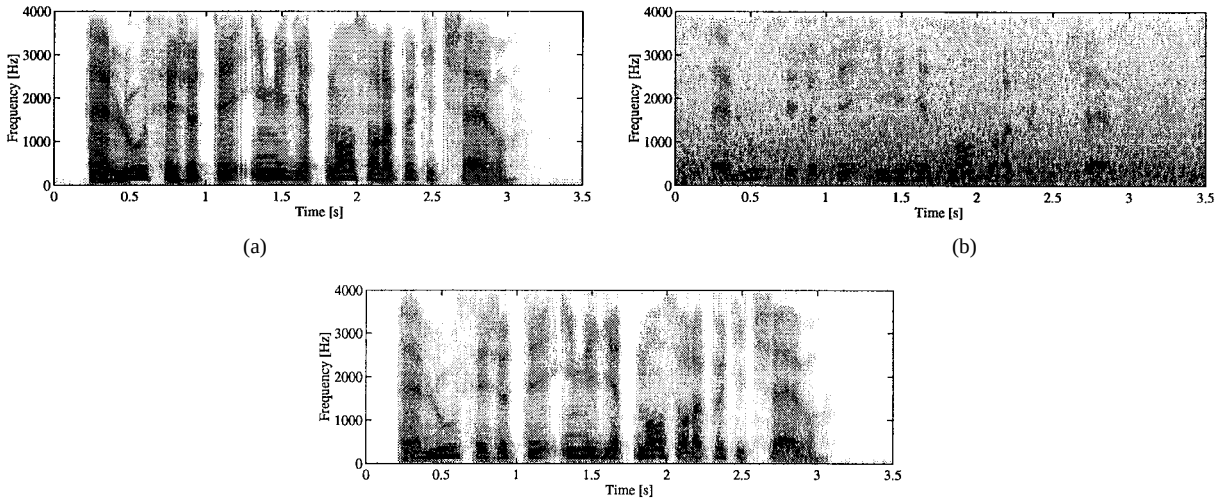


Fig. 8. Speech spectrograms. (a) Original clean speech signal in French: "Un loup s'est jeté immédiatement sur la petite chèvre." (b) Noisy signal (additive speechlike noise at a SNR = 0 dB). (c) Theoretical limit (SNR = 8.54 dB).

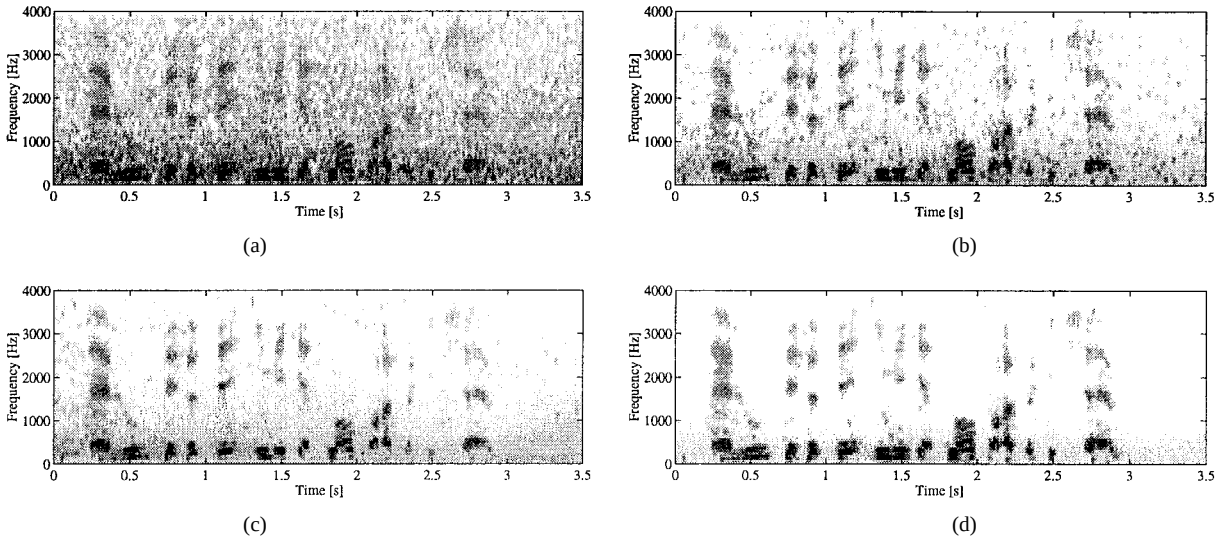


Fig. 9. Speech spectrograms. (a) Power spectral subtraction (SNR = 4.94 dB). (b) Modified spectral subtraction (SNR = 5.81 dB). (c) Nonlinear spectral subtraction (SNR = 5.12 dB). (d) Proposed enhancement algorithm (SNR = 6.36 dB).

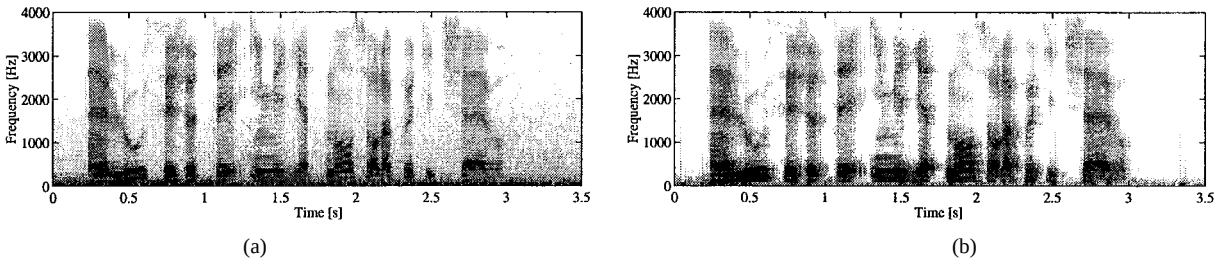


Fig. 10. Speech spectrograms. (a) Noisy speech in the case of additive car noise (SNR = 0 dB, IS = 3.8, AI = 0.42). (b) Speech enhanced with the proposed method (SNR = 9.5 dB, IS = 1.9, AI = 0.50).

E. Recognition Results

The proposed enhancement algorithm has been applied to the problem of speech recognition in adverse environments. Normally, speech recognition gives better results if noise is added to the templates rather than if it is subtracted from the data. However, at very low SNR's, the compensation of the models in noise leads to an increased variance and a reduction of discriminability [20]. In this case, it would therefore be interesting to use our enhancement method as a preprocessing stage in order to improve the recognition rate

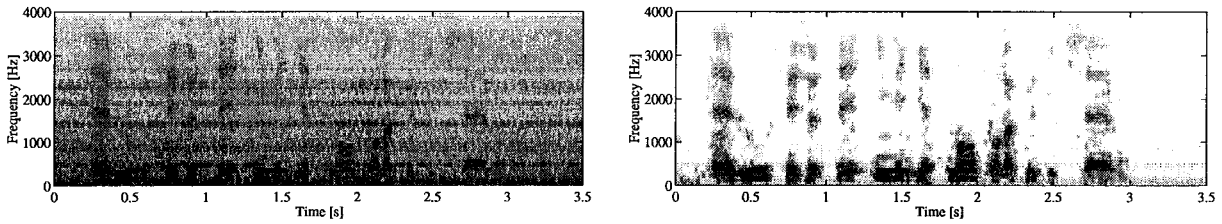


Fig. 11. Speech spectrograms. (a) Noisy speech in the case of additive helicopter cockpit noise (SNR = 0 dB, IS = 2.6, AI = 0.21). (b) Speech enhanced with the proposed method (SNR = 7.7 dB, IS = 1.6, AI = 0.41).

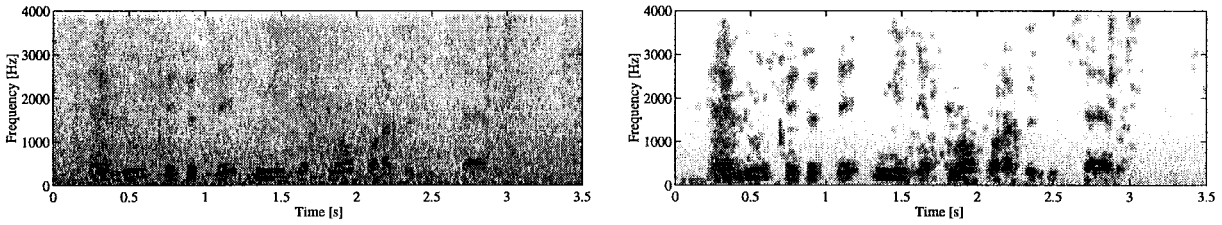


Fig. 12. Speech spectrograms. (a) Noisy speech in the case of additive factory noise at a SNR = 0 dB. (b) Speech enhanced with the proposed method.

TABLE IV
RESULTS OF THE SUBJECTIVE LISTENING TESTS IN THE CASE OF SPEECHLIKE NOISE AT SNR = 0 dB

Enhancement method	*MOS score*	*AI*	*SNR improvement*
Power spectral subtraction	1.18	0.21	4.9 dB
Modified power spectral subtraction	1.83	0.32	5.8 dB
Nonlinear spectral subtraction	1.96	0.33	5.1 dB
Proposed algorithm	2.81	0.39	6.3 dB
Theoretical limit	3.39	0.52	8.5 dB

in noise of systems trained in clean environments. For this particular application, the tradeoff between noise reduction and speech distortion has to be modified compared to the situation with a human listener. Indeed, speech recognizers are more sensitive to speech distortion. In the proposed scheme, this modification can be easily realized in the proposed algorithm by the selection of an appropriate value for $\alpha_{\max}$ (in this case $\alpha_{\max} = 4$).

The proposed algorithm has been tested with a *speaker-dependent isolated digits hidden Markov model-based (HMM-based) recognizer*. The enhancement/recognition scheme has been evaluated using a database created by artificially adding different types of background noises to isolated digits (from 0–9, in English). Therefore, this data does not consider the effect of stress or Lombard effect on the production of speech in noisy environments. Both noise and speech are taken from the Noisex-92 database and the sampling frequency is 16 kHz. Each digit is represented with an HMM containing eight emitting states and one Gaussian distribution per state. The speech was preprocessed using a 32 ms Hanning window every 16 ms, and then parameterized into the first 12 mel frequency cepstral coefficients and the energy, together with the first- and second-order derivatives of these 13 parameters. Recognition uses a standard Viterbi decoder.

The recognition rate improvements obtained with a classical power spectral subtraction and the proposed algorithm are summarized in Table V for various SNR's and noise types. Almost no improvement is provided by the classical subtractive-type algorithm due to the presence of an important residual noise. Nevertheless, the proposed algorithm leads to a significant recognition rate improvement in all cases, except for speechlike noise at SNR $>$ 3 dB, where the recognition rate is better if there is no preprocessing (in this particular case, the proposed algorithm achieves however a significant improvement over the basic power spectral subtraction). This is due to the fact that speech and noise are in the same frequency range, leading to an increased distortion in the noise subtraction process. Therefore, the results presented here show that the introduction of an auditory masking-based processing that modifies the musical structure of the residual noise into a more "white" structure is also a good approach even when the end user is a machine.

V. CONCLUSION

Single channel subtractive-type enhancement systems are efficient in reducing background noise; however, they introduce a perceptually annoying residual noise. In this paper, a simple but efficient way to take into account properties of the auditory system in the enhancement process is proposed. The new algorithm introduces a criterion based on auditory masking. This phenomenon is modeled by the calculation of a noise masking threshold, below which all components are inaudible. The main advantages of the proposed algorithm are the following.

1) It is computationally efficient (the masking threshold computation does not increase too much the computational load because it is based on the FFT of the signal already computed for spectral subtraction and because a simplified method is used).
2) The subtraction parameters are adapted based on a criterion much more correlated with speech perception than the SNR.

TABLE V
IMPROVEMENT OF RECOGNITION RATE OBTAINED USING A CLASSICAL POWER SPECTRAL SUBTRACTION AND THE PROPOSED SCHEME OVER THE CASE WHERE NO PREPROCESSING IS PERFORMED

Input SNR	*White Gaussian noise*		*Speech-like noise*		*F16 cockpit noise*	
	Spectral subtraction	*Proposed algorithm*	*Spectral subtraction*	*Proposed algorithm*	*Spectral subtraction*	*Proposed algorithm*
−3 dB	3 %	18 %	0 %	22 %	0 %	40 %
0 dB	11 %	25 %	0 %	42 %	10 %	56 %
3 dB	0 %	15 %	0 %	5 %	0 %	25 %
6 dB	0 %	20 %	0 %	0 %	0 %	24 %
9 dB	10 %	17 %	0 %	0 %	10 %	20 %

3) It offers the possibility to adjust the tradeoff between noise reduction, residual noise and speech distortion.

The proposed algorithm has been tested and compared to classical subtractive-type algorithms, in various noise types and levels. The objective evaluation has been completed by the description of the speech spectrograms, subjective listening tests, and recognition results. Results show that the background noise is reduced and that the residual noise is less structured than with the classical methods, while the distortion of speech remains acceptable. Hence, we can conclude that the introduction of considerations based on perceptual properties in the enhancement process allows for a significant improvement over classical methods, especially at low SNR's.

ACKNOWLEDGMENT

The author would like to thank P. Renevey for his help in preparing the recognition results, and Prof. J. H. L. Hansen of the Robust Speech Processing Laboratory, Duke University, for his helpful comments and suggestions on this work.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.
[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.
[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, Washington, DC, Apr. 1979, pp. 208–211.
[4] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137–145, Apr. 1980.
[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.
[6] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, June 1992.
[7] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 598–614, Oct. 1994.
[8] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psycho-acoustic criteria," in *Proc. IEEE ICASSP*, Minneapolis, MN, Apr. 1993, pp. 359–361.
[9] T. Usagawa, M. Iwata, and M. Ebata, "Speech parameter extraction in noisy environment using a masking model," in *Proc. IEEE ICASSP*, Adelaide, Australia, Apr. 1994, vol. II, pp. 81–84.
[10] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints on an auditory-based spectrum," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 22–34, Jan. 1995.
[11] J. D. Johnston, "Transform coding of audio signal using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
[12] N. Virag, "Speech enhancement based on masking properties of the auditory system," in *Proc. IEEE ICASSP*, Detroit, MI, May 1995, pp. 796–799.
[13] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models.* Berlin, Germany: Springer-Verlag, 1990.
[14] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1652, Dec. 1979.
[15] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing*, vol. 41, pp. 3463–3479, Dec. 1993.
[16] J. F. Lynch, J. G. Josenhans, and R. E. Crochiere, "Speech/silence segmentation for real-time coding via rule based adaptive endpoint detection," in *Proc. IEEE ICASSP*, Dallas, TX, Apr. 1987, pp. 1348–1351.
[17] S. Quakenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[18] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[19] K. D. Kryter, "Methods for the calculation and the use of the articulation index," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1689–1697, Nov. 1962.
[20] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, Apr. 1995.

Nathalie Virag was born in Leuven, Belgium, on February 21, 1969. She received the M.S. degree in electrical engineering in 1993 and the Ph.D. degree in 1996, both from the Swiss Federal Institute of Technology, Lausanne, Switzerland. Her Ph.D. research involved speech enhancement and recognition in adverse environments based on properties of the human auditory system.

Since 1996, she has been a Research Assistant at the Signal Processing Laboratory, Swiss Federal Institute of Technology, in the field of biomedical signal processing.