Speech Enhancement Based on Audible Noise Suppression

Dionysis E. Tsoukalas, John N. Mourjopoulos, *Member, IEEE,* and George Kokkinakis, *Senior Member, IEEE*

***Abstract*—A novel speech enhancement technique is presented based on the definition of the psychoacoustically derived quantity of audible noise spectrum and its subsequent suppression using optimal nonlinear filtering of the short-time spectral amplitude (STSA) envelope. The filter operates with sparse spectral estimates obtained from the STSA, and, when these parameters are accurately known, significant intelligibility gains, up to 40%, result in the processed speech signal. These parameters can be also estimated from noisy data, resulting into smaller but significant intelligibility gains.**

I. Introduction

THE PROBLEM of enhancing speech degraded by noise remains largely open, even though many significant techniques have been introduced over the past decades. This problem is more severe when no additional information on the nature of noise degradation is available (in the form of an independent measurement, for example), in which case the enhancement technique must utilize only the specific properties of the speech and noise signals.

Existing enhancement methods can be broadly grouped into those aiming at improving speech degraded at low signal-to-noise ratios (SNR's), mainly in order to facilitate communication and intelligibility (either by human or by machine recognizers), and those aiming at improving speech degraded at relatively high SNR's mainly in order to enhance its quality and presentation.

In terms of the methodology adopted by these existing methods, it is evident that although many, usually older approaches were based on specific properties of the speech signal itself, e.g., on speech periodicity [1]–[3], on a model of speech or the production mechanism, etc. [4]–[9], most recent methods are based on the manipulation of the short-time spectral amplitude (STSA) of the degraded signal. Such manipulation schemes are based on the assumption that speech and additive noise degradation are uncorrelated and that it is possible to derive an optimal statistical operator based either on signal spectral variance (e.g., using various spectral subtraction schemes [10]–[14]), or on minimum mean square error (MMSE), e.g., using various forms of Wiener filtering [15]–[17]. All these methods are efficiently implemented on the STSA, and it is also significant that STSA is a relevant signal representation from a perceptual point of view. Given that the human auditory system performs some form of frequency signal analysis and reconstruction under adverse listening conditions, it is also appropriate that enhancement methods are modeled on such procedures. However, hearing models have not been fully exploited by existing enhancement methods apart from [18], where lateral inhibition principles are employed.

Here, an enhancement scheme is presented based on the utilization of a well-known auditory mechanism, noise masking. In addition, estimation procedures are introduced that can optimally or conditionally modify psychoacoustically derived variants of the STSA function. As it is well known from psychoacoustics [19], speech and other signals can mask noise components coexisting with them (in an additive STSA sense). In this sense, the noise degradation perceived by the listener will vary in time according to the time-varying properties of speech STSA, and it is this audible noise component of the degradation that must be removed by the enhancement scheme. Therefore, the enhancement approach adopted here is based on the definition of an audible noise component of the STSA [20], [21], which is extended and used for the derivation of an optimal modifier that achieves audible noise suppression. Furthermore, this modification selectively affects the perceptually significant spectral values, and is therefore more robust than methods that affect the complete STSA and less prone to introduction of unwanted distortions.

Based on the above model, it is shown that optimal psychoacoustic modification can be achieved when only sparse clean signal components (i.e., one spectral value per critical band) are known or have been estimated. Furthermore, it was found that the necessary clean speech data for enhancement are as many as the number of critical bands (CB's) per data window. Apart from this, the only information about the noise required by the technique is restricted to a broad estimate of the noise level per CB.

The performance of the proposed technique was evaluated using objective measures such as the SNR and the noise-to-mask ratio (NMR). Furthermore, the technique was assessed by the diagnostic rhyme test (DRT) and the semantically unpredictable sentences (SUS) test. From these tests, it was found that, at very low SNR's (−5 dB), significant improvements could be achieved by the proposed method. It was also found that the proposed technique could achieve speech reconstruction for arbitrary low SNR's given the correct sparse data. This important result on one hand illustrates the validity of the proposed psychoacoustic model and on the other hand

Manuscript received February 21, 1995; revised November 5, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

The authors are with the Wire Communications Laboratory, University of Patras, 265 00 Patras, Greece (e-mail: mourjop@wcl.ee.upatras.gr).

Publisher Item Identifier S 1063-6676(97)07842-5.

gives an indication of the lower bit rate limits for perceptually significant speech coding.

In terms of speech enhancement now, and assuming that no additional information on the clean signal is known, the proposed technique relies on accurate estimates of these sparse data (which are either the spectral minimum or the masking threshold per CB) from the noisy signal. Although this is a difficult task, two estimation methods are proposed here, the first one based on the statistical distribution of the spectral minima per CB and the second one based on an iterative preprocessing enhancement procedure in conjunction with a rough estimate of the masking threshold. These estimation methods were also evaluated in terms of subjective tests and for several initial SNR conditions, and it was found that in most cases improvements could be achieved of which the most significant were for low initial SNR conditions.

This paper is organized as follows. Section II gives the basic definitions of the proposed psychoacoustic model for speech enhancement as well as the STSA modification scheme. Section III provides methods for practical estimation of the sparse speech data used by the proposed audible noise suppression (ANS) technique. Section IV gives technical details of the processing scheme and describes the implementation and testing of the ANS technique. Section V describes the objective and subjective tests employed for the evaluation of the technique and presents the results. Finally, conclusions are drawn and further work is proposed in Section VI.

II. Psychoacoustic Model for Speech Enhancement

A. Definitions of the Perceptually Significant Spectra

The analysis that follows assumes that the speech and noise signals are discrete-time and finite in duration. In the case of additive noise, the noisy speech signal consists of the sum of the original (clean) speech signal and the noise component, i.e.,

$$y(n) = x(n) + d(n), \qquad 0 \le n \le N-1 \tag{1}$$

where $x(n)$ is the noise-free speech signal, and, $d(n)$ is the noise component.

Equation (1) has an equivalent representation in the frequency domain. Since, in most practical situations, short-time spectra will be required, the Fourier transforms of the windowed noisy and clean speech given by $Y_w(k, i)$ and $X_w(k, i)$, respectively, must be calculated, i.e.,

$$Y_w(k, i) = \sum_{n=0}^{K-1} y(n + \text{off}_i) w(n) I_K^{kn}, \qquad 0 \le k \le K-1 \tag{2}$$

$$X_w(k, i) = \sum_{n=0}^{K-1} x(n + \text{off}_i) w(n) I_K^{kn}, \qquad 0 \le k \le K-1 \tag{3}$$

where $I_K^{kn} = e^{-j(2\pi kn/K)}$, $w(k)$ is a window function [22], K is the length of the Fourier transform, i is the time-domain window index, and, off_i is an offset, assuming that the speech signal is transformed using overlapping time windows.

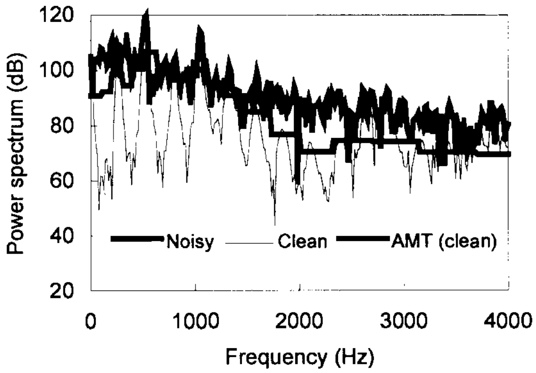


Fig. 1. Power spectra of a short-time speech frame for the noisy, clean speech and its AMT.

The corresponding power spectra are given by $Y_p(k, i)$ and $X_p(k, i)$, respectively, i.e.,

$$Y_p(k, i) = |Y_w(k, i)|^2, \qquad 0 \le k \le K-1 \tag{4}$$

$$X_p(k, i) = |X_w(k, i)|^2, \qquad 0 \le k \le K-1. \tag{5}$$

The basic principle of the psychoacoustic signal enhancement technique is the suppression of spectral components contributing to audible noise. These components can be obtained from an estimate of the auditory masking threshold (AMT), denoted as $T(k, i)$, of the clean signal. The method for the estimation of the AMT is described in Appendix A. As is known [23], the AMT determines the spectral amplitude threshold below which all frequency components are masked in the presence of the masker signal. Consequently, noisy spectral components below this threshold will be inaudible due to the effect of the speech signal.

Typical speech power spectra along with the AMT are shown in Fig. 1. In mathematical terms, the audible spectral components can be expressed using the $\max\{\ \}$ operator, i.e., by taking the maximum between the power spectrum of the speech and the corresponding AMT per frequency component. This function is defined as the audible spectrum of the speech and, in fact, it can be shown that reconstruction of the signal using this function can result in a perceptual equivalent to the original signal, as is also well established in broadband audio coding applications [24]. Now, let us define the audible spectrum of the noisy speech and the audible spectrum of the clean speech as $A_y(k, i)$ and $A_x(k, i)$, respectively, using the expressions

$$\begin{aligned} A_y(k, i) &= \max\{Y_p(k, i), T(k, i)\} \\ &= \begin{cases} Y_p(k, i), & \text{if } Y_p(k, i) \ge T(k, i) \\ T(k, i), & \text{if } Y_p(k, i) < T(k, i), \end{cases} \\ &\quad 0 \le k \le K-1 \end{aligned} \tag{6}$$

$$\begin{aligned} A_x(k, i) &= \max\{X_p(k, i), T(k, i)\} \\ &= \begin{cases} X_p(k, i), & \text{if } X_p(k, i) \ge T(k, i) \\ T(k, i), & \text{if } X_p(k, i) < T(k, i), \end{cases} \\ &\quad 0 \le k \le K-1. \end{aligned} \tag{7}$$

Therefore, the audible spectrum of the additive noise, that is, the spectral components that are perceived as noise, denoted

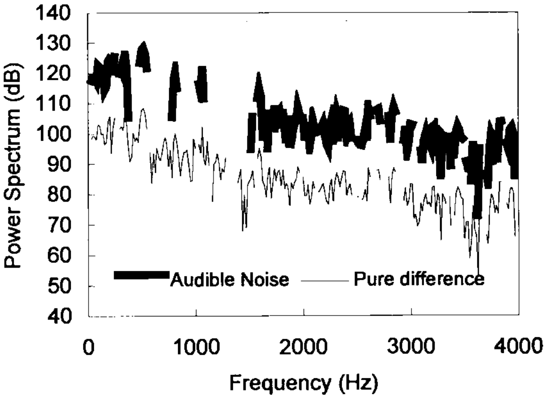


Fig. 2. Power spectra of a short-time speech frame for the audible noise and the pure difference between noisy and clean spectra. Note that resulting negative amplitude noise components are not shown, and that the audible noise was shifted for clarity 20 dB upwards.

as $A_d(k, i)$, can be expressed by the difference between the audible spectra of the noisy and the clean speech. In fact, the main differences between the audible spectrum of noise and the pure difference between noisy and clean spectra, are the reduction in the dynamic range and the order of the estimated noise spectral components. This, in turn, leads to significant processing advantages, since modification of the noisy speech spectrum to suppress the audible noise will introduce less distortion in the speech signal, since only selective frequency components will be modified. Ideally, given a good estimate of the audible noise spectrum, modification of the noisy signal will only affect the audible noise regions and will not distort in an audible manner the underlying speech signal. Therefore, the audible spectrum of the noise is defined as

$$A_d(k, i) = A_y(k, i) - A_x(k, i), \qquad 0 \le k \le K - 1. \quad (8)$$

A typical illustration of the audible noise spectrum and the pure difference between noisy and clean spectrum is shown in Fig. 2, for the short-time spectra of Fig. 1. As can be easily observed in this figure the "pure difference" noise is an overestimation of the audible noise since components of the "pure difference" noise appear in spectral areas in which there is not audible noise.

A more analytic expression for the audible noise can now be found by substituting (6) and (7) for $A_y(k, i)$ and $A_x(k, i)$, respectively, in (8). Then the audible noise can be expressed as shown in (9) [21], at the bottom of the page, which is a four-branched function depending on the relative levels of the power spectra of noisy and clean speech and the corresponding AMT of the clean signal.

B. Psychoacoustic Criteria for Noise Removal

Examination of (9) results in the following observations:

1) Branch (I) may be positive, negative or zero, depending on the relative values of $Y_p(k,i)$ and $X_p(k,i)$.
2) Branch (II) is always positive or zero as indicated by the corresponding conditions. Clearly in this case, there is audible noise that must be removed.
3) Branch (III) is always negative or zero and, consequently, in this case there is not audible noise and no modification is required.
4) Branch (IV) is zero by definition.

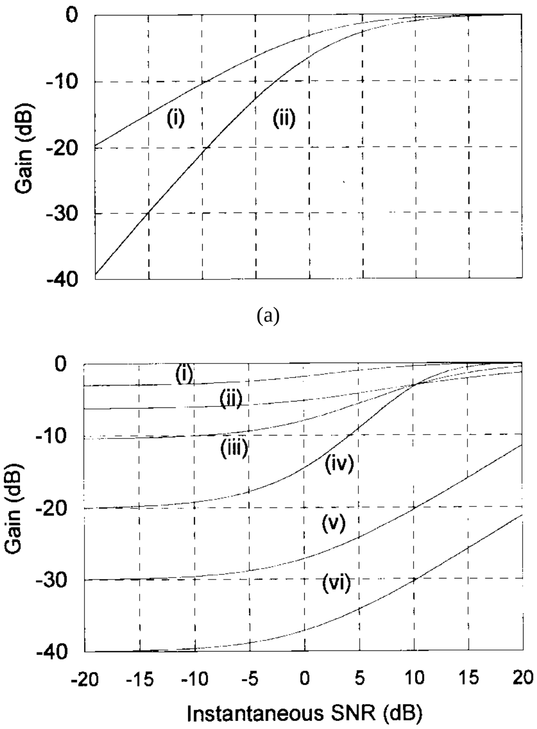
As is also clear from (9), the audible noise spectrum depends on three functions, the noisy speech power spectrum $Y_p(k, i)$, the clean speech power spectrum $X_p(k, i)$, and the AMT $T(k, i)$ of the clean speech. Since only the noisy speech is usually available for processing, this function alone has to be modified for speech enhancement. Therefore, the principle of the proposed ANS technique is to make the audible noise spectrum $A_d(k, i)$ less than or equal to zero by proper modification of the noisy speech power spectrum $Y_p(k, i)$. Consequently, if the noisy speech power spectrum $Y_p(k, i)$ is suitably modified in order to derive the enhanced speech power spectrum, denoted by $\hat{X}_p(k, i)$, then the modified audible noise spectrum, denoted by $\hat{A}_d(k, i)$ must satisfy

$$\hat{A}_d(k, i) \le 0, \qquad 0 \le k \le K - 1. \quad (10)$$

As described in Appendix B, the equality above can be directly obtained from the MMSE estimator, i.e., by considering minimization of $\hat{A}_d(k, i)$ over a specific frequency band. Furthermore, the inequality introduced in (10) was primarily considered in order to give a further degree of freedom in the noise removal process. According to this, a negative value of the $\hat{A}_d(k, i)$ component will mean that: i) either the speech spectrum $\hat{X}_p(k, i)$ was underestimated [Branch I of (9)], in which case a suboptimal solution may be obtained, or ii) the speech spectrum was correctly estimated below the AMT $T(k, i)$ as indicated by the conditions in Branch II of (9) and, hence, by definition is not audible. Note that Branches III and IV of (9) will not be affected by the introduction of the $\hat{X}_p(k, i)$ spectrum. From the above, only case i) may affect the accuracy of the proposed algorithm although, as will be shown from the results in Section V, this effect is rather small.

Efficient spectral modification of the noisy speech power spectrum can be achieved by several methods, as has been shown in the literature (e.g., [10], [16], [25]). Note, however, that for the class of techniques using linear noise suppression, the gain applied to each spectral component is a function of the level of a measurement of the noisy speech and/or the background noise. Such gain curves, for example, the

$$A_d(k, i) = \begin{cases} Y_p(k, i) - X_p(k, i), & \text{if } Y_p(k, i) \ge T(k, i) \text{ and } X_p(k, i) \ge T(k, i) \quad \text{(I)} \\ Y_p(k, i) - T(k, i), & \text{if } Y_p(k, i) \ge T(k, i) \text{ and } X_p(k, i) < T(k, i) \quad \text{(II)} \\ T(k, i) - X_p(k, i), & \text{if } Y_p(k, i) < T(k, i) \text{ and } X_p(k, i) \ge T(k, i) \quad \text{(III)} \\ 0, & \text{if } Y_p(k, i) < T(k, i) \text{ and } X_p(k, i) < T(k, i) \quad \text{(IV)} \end{cases}, \qquad 0 \le k \le K - 1 \quad (9)$$



(a)

(b)

Fig. 3. Gain versus the instantaneous SNR for STSA enhancement methods, for (a) the i) power spectral subtraction and ii) the Wiener filter method, and (b) for ANS (11). i) $\nu(k, i) = 1$, $a(k, i) = D_p$. ii) $\nu(k, i) = 0.5$, $a(k, i) = 10 D_p$. iii) $\nu(k, i) = 1$, $a(k, i) = 10 D_p$. iv) $\upsilon(k, i) = 2$, $a(k, i) = 10 D_p$. v) $\nu(k, i) = 1$, $a(k, i) = 1000 D_p$. vi) $\nu(k, i) = 1$, $a(k, i) = 10\,000\, D_p$, where D_p is the background noise.

power spectral subtraction gain and the Wiener filter gain [16] are shown in Fig. 3(a) as a function of the instantaneous SNR. Given that such gain curves imply constraints in the modification of the noisy speech spectral components, more flexible suppression functions will be required for audible noise spectrum suppression. Therefore, in our case, a parametric nonlinear function was used, which allows greater flexibility in gain control. This function is given by

$$\hat{X}_p(k, i) = \frac{Y_p^{\nu(k,i)}(k, i)}{a^{\nu(k,i)}(k, i) + Y_p^{\nu(k,i)}(k, i)} Y_p(k, i) \qquad (11)$$

where $a(k, i)$ and $\nu(k, i)$ are the time-frequency varying parameters.

As can be observed from (11), the enhanced power spectrum is controlled by the two parameters $a(k, i)$ and $\nu(k, i)$ which are assumed to be both positive. Parameter $a(k, i)$ is a threshold below which all frequency components are highly suppressed. Parameter $\nu(k, i)$ controls the rate of suppression. This rate, however, depends on the ratio $Y_p(k, i)/a(k, i)$, i.e., if this ratio is larger than one, then the larger the $\nu(k, i)$, the smaller the suppression becomes, while if it is smaller than one, the larger the $\nu(k, i)$ the larger the suppression becomes. Typical gain curves obtained by (11) are shown in Fig. 3(b) as a function of the instantaneous SNR. These gain curves imply, in contrast to the gain curves of Fig. 3(a), that suppression remains almost constant for the low-level instantaneous SNR values. This fact may be of significance, since intelligibility degradation [10], [26] after processing is mainly due to exaggerated suppression of low-level speech components, as is the case with the spectral subtraction and the Wiener filter techniques. Note, also, that the ratio $Y_p^{\nu(k,i)}(k, i)/[a^{\nu(k,i)}(k, i) + Y_p^{\nu(k,i)}(k, i)]$ in (11) is always below or equal to one, assuming both $a(k, i)$ and $\nu(k, i)$ are positive.

C. Parameter Estimation for Psychoacoustic Modification

It is now necessary to introduce expressions for optimum modification of the noisy speech spectrum by adjusting the parameters $a(k, i)$ and $\nu(k, i)$ according to the constraints specified by the psychoacoustic model. By combining (9) and (10), substituting $\hat{X}_p(k, i)$ for $Y_p(k, i)$, and taking into account that only Branch (I) and (II) of (9) must be modified, we obtain the set of equations shown in (12), at the bottom of the page, where, as was mentioned, Branches (III) and (IV) of (9) are not involved in the enhancement process, since they do not contribute to audible noise components. By substituting (11) for $\hat{X}_p(k, i)$ into (12), we obtain (13), shown at the bottom of the page, where, hereafter, the common condition $Y_p(k, i) \geq T(k, i)$ in Branches (I) and (II) of (12) will be omitted for simplicity.

By solving (13), and since $\nu(k, i)$ is positive, the following solutions are obtained as shown in (14), shown at the bottom of the next page. Note, however, that it is not desirable to estimate the parameters $a(k, i)$ and $\nu(k, i)$ for every spectral component k, because in this way the estimation will be very

$$\begin{array}{ll} \hat{X}_p(k, i) - X_p(k, i) \leq 0, & \text{if } Y_p(k, i) \geq T(k, i) \text{ and } X_p(k, i) \geq T(k, i) \text{ (I)} \\ \hat{X}_p(k, i) - T(k, i) \leq 0, & \text{if } Y_p(k, i) \geq T(k, i) \text{ and } X_p(k, i) < T(k, i) \text{ (II)} \end{array} \qquad 0 \leq k \leq K - 1 \qquad (12)$$

$$\left. \begin{array}{ll} \dfrac{Y_p^{\nu(k,i)}(k, i)}{a^{\nu(k,i)}(k, i) + Y_p^{\nu(k,i)}(k, i)} Yp(k, i) - X_p(k, i) \leq 0, & \text{if } X_p(k, i) \geq T(k, i) \text{ (I)} \\ \dfrac{Y_p^{\nu(k,i)}(k, i)}{a^{\nu(k,i)}(k, i) + Y_p^{\nu(k,i)}(k, i)} Y_p(k, i) - T(k, i) \leq 0, & \text{if } X_p(k, i) < T(k, i) \text{ (II)} \end{array} \right., \qquad 0 \leq k \leq K - 1 \qquad (13)$$

sensitive to specific spectral values. Apart from this, the CB's are sufficient for the definition of the perceptually significant frequency regions. For these reasons, it is desirable to use a fixed value of $a(k, i)$ and $\nu(k, i)$ over a specific frequency range. Therefore, the above process will be applied to a specific bandwidth of the signal with upper and lower limits k_{lb} and k_{hb}, which correspond to the lower and upper limits of CB b. In this frequency band, the parameters $a(k, i)$ and $\nu(k, i)$ will be constant and denoted by $a_b(i)$ and $\nu_b(i)$. Let also $\nu_b(i)$ take an arbitrary positive value within this band. Clearly, specific frequencies k_j: $k_{lb} \leq k_j \leq k_{hb}$ within this band may correspond to maximum values for both $\alpha_{\rm I}(k, i)$ and $\alpha_{\rm II}(k, i)$ in (14). If $k_{\rm I}$ is such a frequency that produces a maximum in Branch (I) of (14) and $k_{\rm II}$ produces a maximum for Branch (II), then these maximum values, denoted as $a_{{\rm I}b}(i)$ and $a_{{\rm II}b}(i)$ will be given in (15), shown at the bottom of the page. Obviously, the single value $a_b(i)$ within CB b will be given by

$$\begin{aligned} a_b(i) &= \max\{a_{{\rm I}b}(i),\, a_{{\rm II}b}(i)\} \\ &= \begin{cases} a_{{\rm I}b}(i), & \text{if } a_{{\rm I}b}(i) \geq a_{{\rm II}b}(i) \\ a_{{\rm II}b}(i), & \text{if } a_{{\rm I}b}(i) < a_{{\rm II}b}(i) \end{cases}. \end{aligned} \tag{16}$$

This expression describes the optimum psychoacoustic solution that satisfies (10) and relies purely on time-varying model parameters. According to this, enhancement of the noisy signal is performed by applying (11) to noisy signal power spectrum using the value of $a_b(i)$ given by (16) in conjunction with (15) and an arbitrary positive value for $\nu_b(i)$.

D. Parameter Error Analysis and Sensitivity

The effect of parameter $\nu_b(i)$ is only critical to the enhancement procedure in an MMSE sense but not in a psychoacoustic sense, since audible noise suppression can be performed for any positive value of $\nu_b(i)$, and, in an MMSE sense, its value can be obtained by minimization of the spectral difference between the clean and the noisy speech spectral components. Such a spectral distance, however, will highly depend on the clean speech spectral components that will be later shown to be undesirable. Therefore, hereafter, the parameter $\nu_b(i)$will be considered to be constant through the entire enhancement procedure. The effect of parameter $a_b(i)$, however, is crucial to the performance of the ANS technique. An underestimate of this parameter may result in insufficient audible noise suppression, although an overestimate, even when it leads to a suboptimal solution, will still satisfy the condition of audible noise removal given by (10). Nevertheless, it is desirable to estimate the error sensitivity of the ANS with respect to $a_b(i)$. For this reason, let's assume that $\tilde{a}_b(i)$ is an estimate of $a_b(i)$. In this case, the normalized error for $a_b(i)$ will be given by

$$E_a = \frac{a_b(i) - \tilde{a}_b(i)}{a_b(i)}. \tag{17}$$

The normalized error in the approximation of the speech components will be, for $\nu_b(i) = 1$

$$E_X = \frac{\hat{X}_p(k, i) - \hat{\tilde{X}}_p(k, i)}{\hat{X}_p(k, i)} = \frac{-E_a}{1 + \dfrac{Y_p(k, i)}{a_b(i)} - E_a} \tag{18}$$

where the term $Y_p(k, i)/a_b(i) = X_p(k, i)/[Y_p(k, i) - X_p(k, i)]$ at the denominator of (18) can be considered as the instantaneous SNR. Let us now examine the asymptotic behavior of (18). At high SNR's, i.e., $Y_p(k, i)/a_b(i) \gg 1$, and since E_a will be significantly smaller than $Y_p(k, i)/a_b(i)$, it may be concluded that $E_X \rightarrow 0$. This means that at high SNR's, errors in $a_b(i)$ will generate insignificant errors in the approximation of the speech signal. At low SNR's, i.e., $Y_p(k, i)/a_b(i) \ll 1$, (18) becomes

$$E_X = \frac{-E_a}{1 - E_a} = -E_a \frac{a_b(i)}{\tilde{a}_b(i)} \tag{19}$$

which means that an overestimation of $a_b(i)$ will produce an underestimation in the speech signal attenuated by $a_b(i)/\tilde{a}_b(i)$, although an underestimation of $a_b(i)$ will be amplified by $a_b(i)/\tilde{a}_b(i)$. Illustration of the speech error E_X for typical values of the error E_a versus the instantaneous SNR $= Y_p(k, i)/a_b(i)$ is shown in Fig. 4. Furthermore, it must be noted that the speech approximation error cannot be arbitrarily large due to the $E_a/(1-E_a)$ factor in (19). If E_a is very large, then E_X tends to one. Therefore, it may be concluded that the ANS is very sensitive to underestimation of $a_b(i)$, which anyway does not satisfy the target of audible noise removal, but is less sensitive to overestimation of $a_b(i)$, since even in

$$\left.\begin{aligned} a_{\rm I}(k, i) &\geq Y_p(k, i)\left[\frac{Y_p(k, i)}{X_p(k, i)} - 1\right]^{1/\nu(k, i)}, \quad \text{if } X_p(k, i) \geq T(k, i) \text{ (I)} \\ a_{\rm II}(k, i) &\geq Y_p(k, i)\left[\frac{Y_p(k, i)}{T(k, i)} - 1\right]^{1/\nu(k, i)}, \quad \text{if } X_p(k, i) < T(k, i) \text{ (II)} \end{aligned}\right., \quad 0 \leq k \leq K - 1 \tag{14}$$

$$\begin{aligned} a_{{\rm I}b}(i) &= Y_p(k_{\rm I}, i)\left[\frac{Y_p(k_{\rm I}, i)}{X_p(k_{\rm I}, i)} - 1\right]^{1/\nu_b(i)}, \quad \text{if } X_p(k, i) \geq T(k, i) \text{ (I)} \\ a_{{\rm II}b}(i) &= Y_p(k_{\rm II}, i)\left[\frac{Y_p(k_{\rm II}, i)}{T(k_{\rm II}, i)} - 1\right]^{1/\nu_b(i)}, \quad \text{if } X_p(k, i) < T(k, i) \text{ (II)} \end{aligned} \qquad k_{lb} \leq k \leq k_{hb} \tag{15}$$

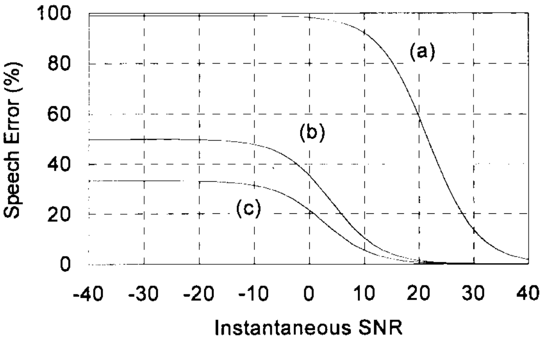


Fig. 4. Speech error E_X (%) versus the instantaneous SNR $[Y_p(k, i)/a_b(i)]$ for typical overestimates of the error E_a. (a) -100. (b) -1. (c) -0.5.

the worst case, i.e., an arbitrary overestimation of $a_b(i)$, the speech signal error will be less than or equal to one.

E. Psychoacoustic Speech Enhancement and Reconstruction Based on Sparse Speech Data

The previously described parametric speech enhancement approach has the disadvantage of relying on a good estimate of the clean speech spectrum, per data window, which is not easily estimated, especially at low SNR's. For this reason, it will be now shown that a relaxation in the requirement of estimating the complete speech spectrum [i.e., $X_p(k, i)$] can be introduced, which will only rely on a single value of the $X_p(k, i)$ components per CB, referred to, thereafter, as sparse speech estimation. This approach, which optimizes the clean speech spectrum estimation within subband regions, has the advantage that such sparse speech components can be more easily detected in noisy signals, so that further enhancement will only rely on these data and not on the exact estimation of the complete speech spectrum. Furthermore, the enhancement parameters $a_b(i)$ are only estimated (and updated) per subband region allowing flexible modification of the noisy signal.

By definition [23], the AMT $T(k, i)$ of the speech signal within each critical frequency band is constant, i.e.,

$$T(k, i) = T_b(i), \qquad k_{lb} \le k \le k_{hb}. \tag{20}$$

Let us now assume, as is approximately true in most practical cases, that

1) the noise $d(n)$ has zero mean and is uncorrelated with the speech $x(n)$, so that [25]

$$Y_p(k, i) = X_p(k, i) + D_p(k) \tag{21}$$

where, $D_p(k)$ is the mean power spectrum of the noise;

2) the power spectrum of the noise remains constant within the same CB, i.e.,

$$D_p(k) = D_{pb}, \qquad k_{lb} \le k \le k_{hb}. \tag{22}$$

Under these assumptions, by substituting (20)–(22) in (14), and assuming again that the maximum values for $a_{\mathrm{I}\,b}(i)$ and $a_{\mathrm{II}\,b}(i)$ correspond to the frequencies k_{I} and k_{II}, respectively, we obtain

$$\begin{aligned} a_{\mathrm{I}\,b}(i) &= [D_{pb} + X_p(k_{\mathrm{I}}, i)]\left[\frac{D_{pb}}{X_p(k_{\mathrm{I}}, i)}\right]^{1/\nu_b(i)} \\ &\qquad X_p(k_{\mathrm{I}}, i) \ge T_b(i) \text{ (I)} \\ a_{\mathrm{II}\,b}(i) &= [D_{pb} + X_p(k_{\mathrm{II}}, i)]\left[\frac{D_{pb} + X_p(k_{\mathrm{II}}, i)}{T_b(i)} - 1\right]^{1/\nu_b(i)} \\ &\qquad X_p(k_{\mathrm{II}}, i) < T_b(i) \text{ (II).} \end{aligned} \tag{23}$$

Note, however, that $a_{\mathrm{I}\,b}(i)$, $a_{\mathrm{II}\,b}(i)$, and k_{I}, k_{II} are not necessarily the same as those implied in (15). In (23), $a_{\mathrm{I}b}(i)$ and $a_{\mathrm{II}b}(i)$ depend only on $X_p(k_{\mathrm{I}}, i)$ and $X_p(k_{\mathrm{II}}, i)$ (which, in turn, depend on the frequencies k_{I} and k_{II}), and on $T_b(i)$ and D_{pb}, which are independent of frequency within the same CB. Therefore, it can be shown (see Appendix C) that frequency k_{I} will now correspond to the minimum value of $X_p(k, i)$ for all $X_p(k, i)$: $X_p(k, i) \ge T_b(i)$, if $0 < \nu_b(i) < 1$, and k_{II} to the maximum value of $X_p(k, i)$ for all $X_p(k, i)$: $X_p(k, i) < T_b(i)$. Therefore, the number of parameters required for speech enhancement has been reduced to the minimum and maximum spectral components $X_p(k_{\mathrm{I}}, i)$ and $X_p(k_{\mathrm{II}}, i)$, the AMT $T_b(i)$ and the broad noise level D_{pb} per CB.

Application of the nonlinear law given by (11) to the noisy speech spectrum, for this value of $a_b(i)$ [obtained by (16) and (23) and $\nu_b(i) \le 1$] per CB, will give an enhanced speech spectrum $\hat{X}_p(k)$ that satisfies (10), i.e., in such a case, the audible noise spectrum $\hat{A}_d(k, i)$ will be ≤ 0 for all frequency components.

Note, however, that the solution given by (16) and (23) is not unique due to the inequality implied by (16). In fact, if $a_b'(i)$ has such a value that $a_b'(i) \ge a_b(i)$, then $a_b'(i)$ will be also a solution that satisfies (10). However, $a_b'(i)$ cannot be arbitrary large, since the enhanced speech spectrum will be finally reduced to zero as can be easily observed in (11). Apart from this, it is desirable to obtain such a solution for $a_b'(i)$, so that dependence on the clean speech frequency components is minimized, i.e., only a few speech components are required for the evaluation of $a_b'(i)$. Two classes of sparse spectral data were derived in this way: one containing the minima of the spectrum and the other containing the AMT. Both approaches require the same number of *a priori* known data, i.e., one spectral value per CB.

1) Audible Noise Suppression Using Spectral Minima: One way to obtain the required sparse data is to estimate $a_b'(i)$ from the first branch of (23) using the minimum speech power spectrum component, denoted by $X_{pb,\,\mathrm{min}}(i)$, in the specific CB instead of the partial minimum component $X_p(k_{\mathrm{I}}, i)$ (from those components above the AMT). However, in such a case, it must be shown that the new parameter $a_b'(i)$ is larger than the corresponding $a_b(i)$ implied by (16) and (23). Therefore, if $a_b'(i)$ is given by

$$\begin{aligned} a_b'(i) &= [D_{pb} + X_{pb,\,\mathrm{min}}(i)]\left[\frac{D_{pb}}{X_{pb,\,\mathrm{min}}(i)}\right]^{1/\nu_b(i)} \\ &\qquad X_{pb,\,\mathrm{min}}(i) = \min_k\{X_p(k, i),\, k_{lb} \le k \le k_{hb}\} \end{aligned} \tag{24}$$

then it can be shown (Appendix D) that

$$\begin{aligned} a'_b(i) &\geq a_{\mathrm{I}\,b}(i) \text{ (I)} \\ a'_b(i) &\geq a_{\mathrm{II}\,b}(i) \text{ (II)} \end{aligned} \tag{25}$$

and, hence, $a'_b(i)$ is also a solution that satisfies (10). In such a way, the amount of the clean speech data required for audible noise suppression has been reduced to one minimum spectral component per CB.

2) Audible Noise Suppression Using the AMT Values: The second way to reduce the speech data *a priori* required for the enhancement is to estimate $a''_b(i)$ from the first branch of (23) using the AMT $T_b(i)$ instead of the partial minimum component $X_p(k_{\mathrm{I}}, i)$ (from those components above the AMT). In this case, $a''_b(i)$ will be given by

$$a''_b(i) = [D_{pb} + T_b(i)] \left[\frac{D_{pb}}{T_b(i)} \right]^{1/\nu_b(i)}. \tag{26}$$

Using this estimate, it can be shown (see Appendix E) that

$$\begin{aligned} a''_b(i) &\geq a_{\mathrm{I}\,b}(i) \text{ (I)} \\ a''_b(i) &\geq a_{\mathrm{II}\,b}(i) \text{ (II)} \end{aligned} \tag{27}$$

and, hence, $a''_b(i)$ is also a solution that satisfies (10). Furthermore, the number of the clean speech data has been reduced to one AMT value $T_b(i)$ per CB.

The solutions given by (24) and (26) indicate that enhancement of the noisy speech is possible using one value per critical band, either the spectral minimum or the AMT of the clean speech, and the broad noise level. This result is of great importance, since the problem of speech enhancement has been now reduced to that of determining only a few components per data window, i.e., selective minima of the speech signal or its AMT values. Given that the number of these data is equal to or less than the number of CB's B, there are, therefore, up to $B = 22$ data values for a 16 kHz sampling rate speech signal (or $B = 18$ for an 8 kHz sampling rate speech signal) [19, ch. 6].

3) The ANS as a Speech Reconstruction Technique: Apart from this, and as will be shown in Section V, the proposed method can theoretically [i.e., when the speech spectrum minima or the AMT are accurately known, using (24) or (26)] improve speech intelligibility irrespective of initial SNR, indicating the correctness of the psychoacoustic model principles. Furthermore, the technique can theoretically work for very low SNR's, since the preceding theory did not make any assumptions for the input SNR. In fact, the proposed method can work even for input SNR $= -\infty$, i.e., when the noisy signal consists only of the noise component given that the sparse speech parameters are known. As will be shown in Section V, intelligible speech will be reconstructed from such a noisy input. This, in turn, suggests a finding of importance, i.e., that a lowest limit of psychoacoustically valid bit rate of the speech can be determined, which will be given by a finite set of frequency speech components, e.g., one per CB, sufficient for resynthesis of the speech signal. In this context, it was also found that the sparse data for reconstruction can be described by 4-bits numbers. In this case, the ANS can achieve a bit rate of 2750 b/s instead of the 256 000 b/s for a 16 KHz, 16-b resolution speech signal.

III. Methods for the Estimation of the Sparse Data for ANS

A. A Statistical Estimator for the Minimum Spectral Value per Critical Band

In order to model the minima of the speech spectrum, it is possible to express them as a function of the mean value of the speech spectrum per critical band, i.e.,

$$X_{b,\,\min}(i) = f\{\overline{X}_b(i)\} \tag{28}$$

where $\overline{X}_b(i)$ is the mean spectral value in band b and time window i, given by

$$\overline{X}_b(i) = \frac{1}{k_{hb} - k_{lb} + 1} \sum_{k=k_{lb}}^{k_{hb}} |X_w(k,\, i)|. \tag{29}$$

In order to use a statistical model for the estimation of the unknown function $f\{\ \}$, it is desirable to measure the probability distribution of the minimum spectral component per CB and that of the mean spectral values per CB. Such measurements were made during this work using speech material from the ESPRIT PROJECT 6819 (SAM-A) speech data base. According to these measurements, the probability distribution of the minimum spectral component follows a Rayleigh distribution for most of the CB's, as shown in Fig. 5(a). The distribution of the mean spectral value on the other hand, was found to approach a normal distribution for all bands, as shown in Fig. 5(b). As can be easily observed in this plot, the conditional mean spectral value distributions, given the minimum value, are shifted versions of the mean spectral value distribution. This suggests that the minimum component per CB can be modeled as linear combination of the mean spectral values per CB which, in turn, can be more easily estimated in noisy conditions.

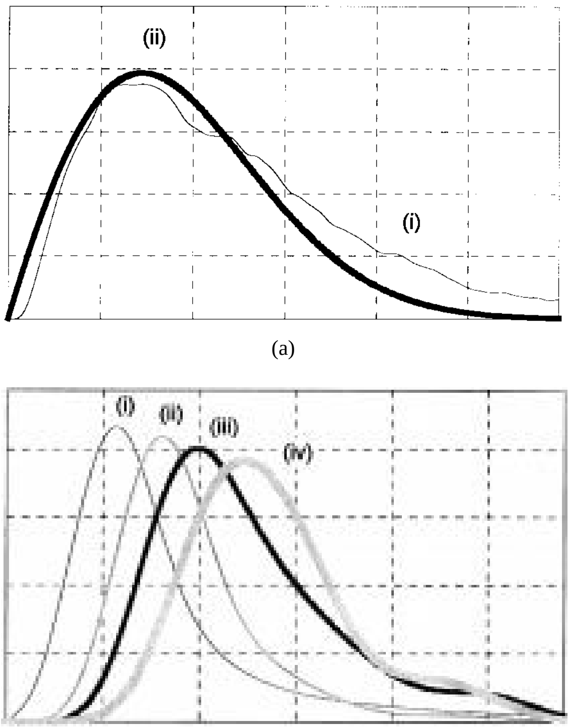
Following the above statistical measurements, let us now define the probability density function (pdf) of the minimum power spectrum component per CB as

$$p(X_{b,\,\min}) = \frac{2X_{b,\,\min}}{\lambda_{b,\,\min}} \exp\left\{ -\frac{X^2_{b,\,\min}}{\lambda_{b,\,\min}} \right\} \tag{30}$$

and the probability of the mean spectral value given the minimum component as

$$p(\overline{X}_b|X_{b,\,\min}) = \frac{2}{\sqrt{p\lambda_{b,\overline{X}}}} \exp\left\{ -\frac{|\overline{X}_b - X_{b,\,\min}|^2}{\lambda_{b,\overline{X}}} \right\} \tag{31}$$

where $\lambda_{b,\,\min} \equiv E\{|X_{b,\,\min}|^2\}$ and $\lambda_{b,\overline{X}} \equiv E\{|\overline{X}_b - X_{b,\,\min}|^2\}$ are the variances of the minimum and the mean power spectrum for critical band b, respectively. Then, in



(a)

(b)

Fig. 5. Experimental distributions of speech spectral parameters for a typical critical band. (a) i) Minimum power spectrum component and ii) corresponding Rayleigh pdf. (b) i) Mean power spectral amplitude and conditionals (ii)–(iv), given the minimum spectral component.

an MMSE sense, the estimator for the minimum spectral component will be given by

$$\hat{X}_{b,\min} = E\{X_{b,\min}|\overline{X}_b\} = \frac{\int_0^\infty X_{b,\min} p(\overline{X}_b|X_{b,\min})p(X_{b,\min})\, dX_{b,\min}}{\int_0^\infty p(\overline{X}_b|X_{b,\min})p(X_{b,\min})\, dX_{b,\min}}. \tag{32}$$

By substituting (30) and (31) for $p(X_{b,\min})$ and $p(\overline{X}_b|X_{b,\min})$, respectively, in (32) the following solution is obtained (see Appendix F):

$$\hat{X}_{b,\min} = \frac{1}{\sqrt{2}}\sqrt{\left(\frac{1}{1+Z_{b,\text{post}}}\right)\left(\frac{Z_{b,\text{prio}}}{1+Z_{b,\text{prio}}}\right)} \cdot M\left[-\sqrt{2}\sqrt{(1+Z_{b,\text{post}})\left(\frac{Z_{b,\text{prio}}}{1+Z_{b,\text{prio}}}\right)}\right]\overline{X}_b. \tag{33}$$

In the above expression, there are several terms to be explained. First, $M[\ ]$ is the function

$$M[z] = \frac{1-\Phi\left(\frac{z}{\sqrt{2}}\right)}{\sqrt{\frac{\pi}{2}}\exp\left\{-\frac{z^2}{2}\right\} - z\left[1-\Phi\left(\frac{z}{\sqrt{2}}\right)\right]} - z \tag{34}$$

where $\Phi(\ )$ is the error function [27, Eq. 8.250.1].

The terms $Z_{b,\text{post}}$ and $Z_{b,\text{prio}}$ are defined as

$$Z_{b,\text{post}} \equiv \frac{\overline{X}_b^2}{\lambda_{b,\overline{X}}} - 1, \quad Z_{b,\text{prio}} \equiv \frac{\lambda_{b,\min}}{\lambda_{b,\overline{X}}}. \tag{35}$$

A similar result was obtained by Ephraim in an earlier work [15] in which estimation of the STSA of the speech was achieved by an MMSE estimator. Although, in that work, the estimator was obtained by the mean probability of the spectral component given the noisy observation, it is believed that similar principles also apply here, so that finally the $Z_{b,\text{prio}}$ term, although here cannot be interpreted as the *a priori* SNR, can be estimated using

$$Z_{b,\text{prio}}(i) = (1-\xi)P[Z_{b,\text{post}}(i)] + \xi\,\frac{|\hat{X}_{b,\min}(i-1)|^2}{\lambda_{b,\overline{X}}} \tag{36}$$

where

$$P[z] = \begin{cases} z, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0 \end{cases} \qquad \text{and } \xi = 0.98.$$

Since the variance of the mean spectrum is also generally unknown, this parameter was adaptively estimated during processing according to the expression

$$\lambda_{b,\overline{X}}(i) = \frac{i \cdot \lambda_{b,\overline{X}}(i-1) + [\overline{X}_b(i) - \hat{X}(i)]^2}{i+1}. \tag{37}$$

In practice, it was found that this parameter after a few windows reached a constant value. Furthermore, the mean spectral value $\overline{X}_b(i)$ was obtained after application of the spectral subtraction method.

B. A Clean Speech AMT Estimator in the Presence of Noise

In this section, it is shown that a satisfactory estimate of the clean speech AMT can be also obtained from the noisy data using an iterative procedure at some expense of computational efficiency. Specifically, this procedure consists of passing the noisy signal through the nonlinear filter given by (11) several times. As will be shown, each time the signal passes through such process, a better approximation of the noise-free speech can be obtained and, consequently, a more accurate AMT estimate can be derived. In some respect, this process of iterative updating of the AMT values resembles a similar procedure by Lim [4] for updating the noisy speech AR parameters.

Let us consider the case when the AMT $T_b(i)$ of the clean speech is known. Then the parameter $a_b(i)$ of the nonlinear function will be given by $a_b''(i)$ of (26). The enhanced speech power spectrum for $\nu_b(i) = 1$[1] will be

$$\hat{X}_p(k,i) = \frac{Y_p(k,i)}{a_b''(i) + Y_p(k,i)}\,Y_p(k,i), \qquad k_{lb} \leq k \leq k_{hb}. \tag{38}$$

[1] As will be shown in Section V, the best performance is obtained by this value of $\nu_b(i)$.

TABLE I
SIMULATION RESULTS FOR THE CLEAN SPEECH AMT ESTIMATOR

Iterations	1	2	3	4	5	6	7	8	9
SNR (dB)	3.36	4.12	4.18	4.21	4.21	4.20	4.20	4.20	4.20
NMR (dB)	17.31	11.34	9.30	8.59	8.16	7.83	7.60	7.43	7.6

Let us now assume that the AMT $T_b(i)$ is not known but an approximation denoted by ${}^1T_b(i)$ is known, which satisfies the constraint

$$ {}^1T_b(i) = T_b(i) + dT_b(i) > T_b(i) \tag{39}$$

where $dT_b(i)$ has a small value, i.e., ${}^1T_b(i)$ is an overestimation of $T(b, i)$.

Then, the j iteration of the enhancement procedure will produce the enhanced power spectrum given by

$$ {}^j\hat{X}_p(k, i) = \frac{{}^{j-1}\hat{X}_p(k, i)}{{}^ja_b''(i) + {}^{j-1}\hat{X}_p(k, i)} \, {}^{j-1}\hat{X}_p(k, i), \quad k_{lb} \le k \le k_{hb}, \quad j = 1, 2, 3, \cdots \tag{40}$$

where ${}^ja_b''(i)$ is given by

$$ {}^ja_b''(i) = {}^{j-1}D_{pb} + \frac{{}^{j-1}D_{pb}^2}{{}^jT_b(i)} \tag{41}$$

and the initial conditions are given by ${}^0\hat{X}_p(k, i) = Y_p(k, i)$ and ${}^0D_{pb} = D_{pb}$.

Apparently, since ${}^ja_b''(i) \ge 0$, from (40) it can be easily shown that

$$ {}^1\hat{X}_p(k, i) \ge {}^2\hat{X}_p(k, i) \ge {}^3X_p(k, i) \ge \cdots \ge {}^j\hat{X}_p(k, i). \tag{42}$$

Furthermore, from (39) it is easy to show that ${}^1\hat{X}_p(k, i) > \hat{X}_p(k, i)$. Note also that parameter ${}^ja_b''(i)$ will be decreasing with the number of iterations, because it is proportional to the amount of background noise measured during nonspeech activity intervals. This ensures that the above process will practically converge to a finite state when ${}^ja_b''(i)$ reaches zero, which means that no more suppression is needed. Therefore, the amount of suppression is larger for small values of j and smaller for large values of j. Since, however, the dynamics of the iterative process are very complicated due to the nonlinear suppression law, simulation was performed to validate the proposed iterative procedure, and results are presented in terms of the SNR and NMR measures (described in Section V) in Table I.

To initialize this iterative process, the first approximation of the AMT ${}^1T_b(i)$ of the speech signal can be easily obtained by the power spectral subtraction technique, which was experimentally found to satisfy the condition implied by (39), although it was also found that even the noisy signal can be used, in which case more iterations must be performed.

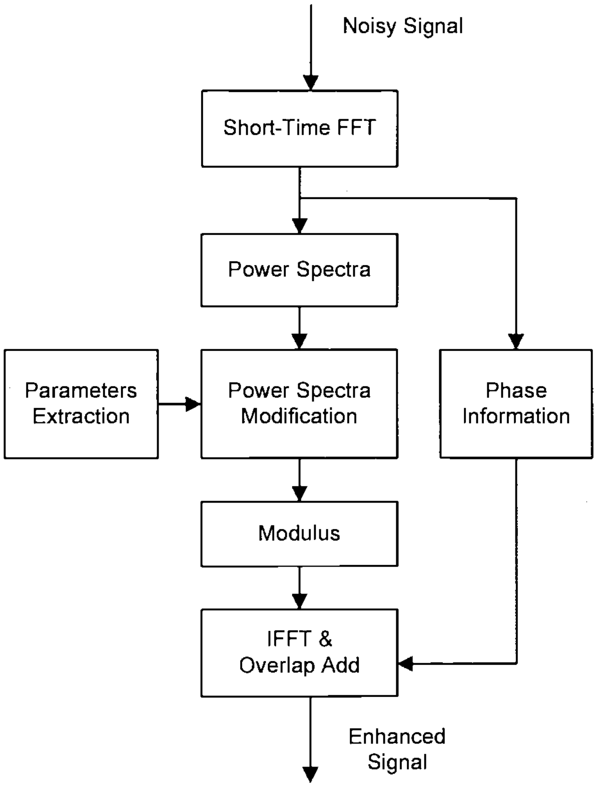


Fig. 6. General block diagram for the ANS technique.

IV. IMPLEMENTATION

A. Algorithm Description

The proposed technique was simulated on a general purpose computer. The speech material was digitized using 16 kHz sampling rate and 16-b resolution, and was stored into files. Noise, also stored in files, was added to the speech signal to produce noisy signals at specific SNR's. After processing, the speech material was also stored into files for further evaluation using objective and subjective measures. The general block diagram of the proposed ANS method is shown in Fig. 6. The steps of the algorithm are summarized below.

1) Short-time windows of the noisy speech are transformed into the frequency domain using the short-time fast Fourier transform (STFFT), as implied by (2).
2) The power spectrum of the noisy speech is obtained using (4), and the phase information is extracted.
3) The power spectrum of the noisy speech is processed using the nonlinear law given by (11) in conjunction with the previously estimated parameters $a_b(i)$ and $\nu_b(i)$ per CB.

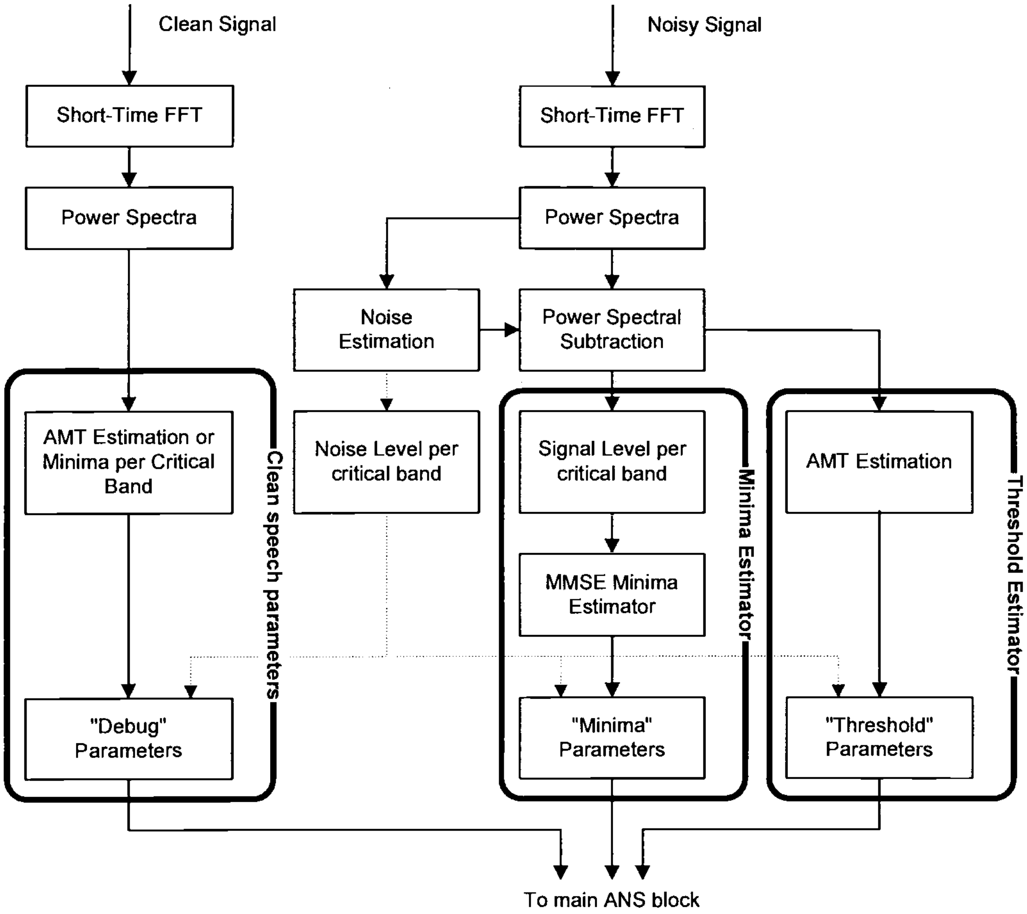


Fig. 7. Parameter extraction block diagram for the ANS technique.

The modulus of the modified power spectrum is transformed back into the time domain using the short-time inverse fast Fourier transform (FFT) and the original (noisy signal) phase information. The enhanced speech is reconstructed using the overlap-add method.

B. Parameter Estimation

The parameter extraction procedure is shown in Fig. 7. This diagram describes three different approaches, one for validation of the technique and two based on the proposed sparse data estimators.

1) The first approach tested was to use the AMT of the noise-free signal in conjunction with (26). Although this method has no meaning in terms of enhancement, it was used in order to show the validity of the proposed method. Apart from this, it is worth it to evaluate the performance of the ANS technique in performing a data compression task, i.e., when the algorithm is fed with the noise signal (SNR $= -\infty$) and only B parameters of a speech signal per data window are known. This method will hereafter be called the *debug* method and will be denoted by "D."
2) The second method tested was based on the statistical model for the estimation of the minimum spectral component in conjunction with (24). This method will be referred to as the *minima* method and will be denoted by "M."
3) The third method tested was based on the clean speech AMT estimator in conjunction with (26). This method will be called the *threshold*, and will be denoted by "T." In utilizing this method, it was found that up to three iterations were necessary for sufficient noise suppression. This is also validated by the results in Table I, where it is shown that after the third iteration there are only negligible changes in the objective SNR and NMR measures.

C. The Noise Data

In order to simulate the proposed technique in a real environment, the type of noise used in the tests should be of practical importance. For these tests, the noise data were drawn from the NOISEX-92 CD-ROM's [28]. From the noise data in these CD-ROM's, and for the tests described in the following sections, the noise denoted as "6-Speech Noise" was chosen. This noise is stationary and has a mean slope of 8 dB/octave, while its main energy is concentrated toward the lower frequencies or, in other words, toward significant frequencies of the speech signal and is therefore, more immune to the application of enhancement.

V. Tests and Results

A. ANS Performance Limit Evaluation

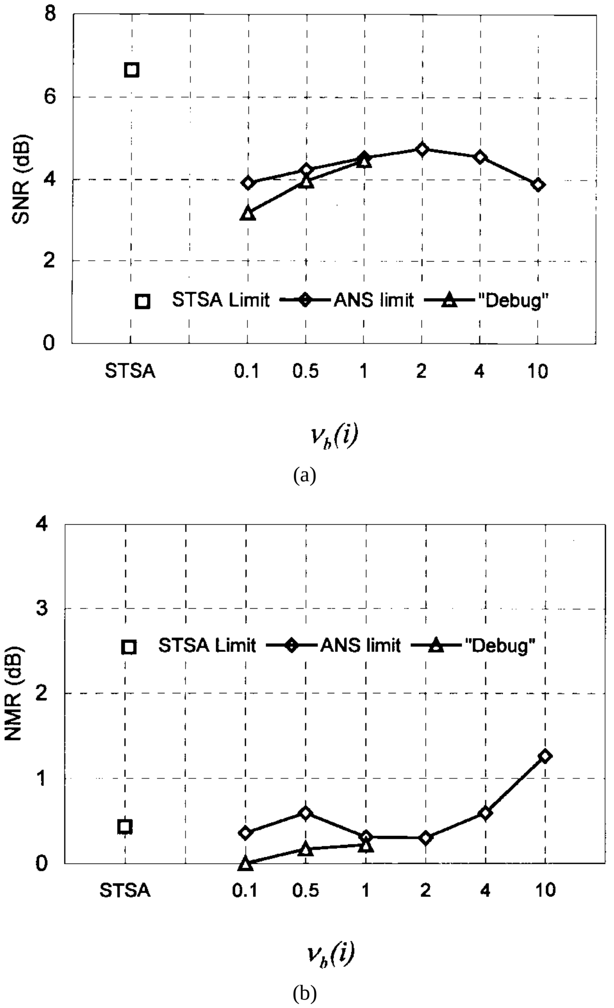
The performance limit of the ANS technique was evaluated by means of objective measures. This evaluation was mainly performed in order: 1) to show the negligible influence of $\nu_b(i)$, 2) to compare the performance of the technique with the theoretical STSA limit, and 3) to compare the ANS technique [(15) and (16)] to the sparse data approach (debug method). The STSA theoretical limit was obtained by reconstructing the speech signal using the clean signal spectral amplitude components combined with the phase of the noisy signal, and indicates the maximum theoretical SNR improvement for STSA-based enhancement methods. The ANS limit was obtained from (11), (15), and (16) by using all the spectral components of the noisy and noise-free speech. The debug method was obtained from (11) and (26). Experiments were performed using approximately 400 s of speech signal from 20 speakers drawn from the ESPRIT PROJECT 6819 (SAM-A) speech data base. Results are presented in Fig. 8 (for the SNR and the NMR measures, described in detail in the next paragraph). As can be observed in this figure, the ANS technique is less sensitive to the influence of the parameter $\nu_b(i)$, although best results were obtained for $\nu_b(i) = 2$ for the ANS limit and for $\nu_b(i) = 1$ for the debug method. Note also that, in terms of the SNR, the ANS technique can achieve an SNR improvement of up to 9.7 dB (for input SNR $= -5$ dB), which is about 2 dB lower than the theoretical STSA enhancement limit (11.6 dB). In terms of the NMR, the ANS technique can achieve slightly better performance compared to the theoretical STSA enhancement limit. This important result, it is believed, is mainly due to the fact that the target of the ANS technique is suppression of the audible noise, which can be more appropriately measured using the NMR than the SNR criterion. Furthermore, results for the debug method have very small differences compared to the ANS limit, which shows that the ANS is less sensitive to the assumptions made by (21) and (22). Therefore, for the subsequent experiments, the value of parameter $\nu_b(i)$ will be equal to one.

B. Objective and Subjective Evaluation

1) Objective Evaluation Tests: Objective evaluation of the proposed method was performed using the classical SNR method and the NMR method. The SNR was measured using [29]:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1}[\phi(n) - x(n)]^2}\ [\mathrm{dB}] \qquad (43)$$

where $x(n)$ is the noise-free speech signal, and $\phi(n)$ is the signal under test, i.e., the noisy or enhanced speech. The NMR method is an objective method based on subjective quantities, and indicates the occurrences of audible noise components (i.e., noise components above the signal's AMT). This method was found by researchers to have a high degree of correlation with subjective tests [30]. For the NMR method, the following expression was used:



(a)

(b)

Fig. 8. Enhancement performance for different values of $\nu_b(i)$, obtained for the ANS method [enhancement limit by (16), the debug condition (26)], and the theoretical limit for STSA methods. The noisy signal SNR was -5 dB and the corresponding NMR 16.5 dB. (a) SNR performance. (b) NMR performance.

$$\mathrm{NMR} = \frac{10}{M}\sum_{i=0}^{M-1}\log_{10}\frac{1}{B}\sum_{b=0}^{B-1}\frac{1}{C_b}\frac{\sum_{k=k_{lb}}^{k_{lh}}|D(k,\, i)|^2}{T_b(i)}\ [\mathrm{dB}] \qquad (44)$$

where M is the total number of windows, B is the number of CB's, C_b is the number of frequency components for CB b, and $|D(k,\, i)|^2$ is the power spectrum of the noise at frequency bin k and time window i, estimated by the difference between the noisy and clean signals in the time domain.

Note that (44) is in accordance with the time-domain segmental SNR [29].

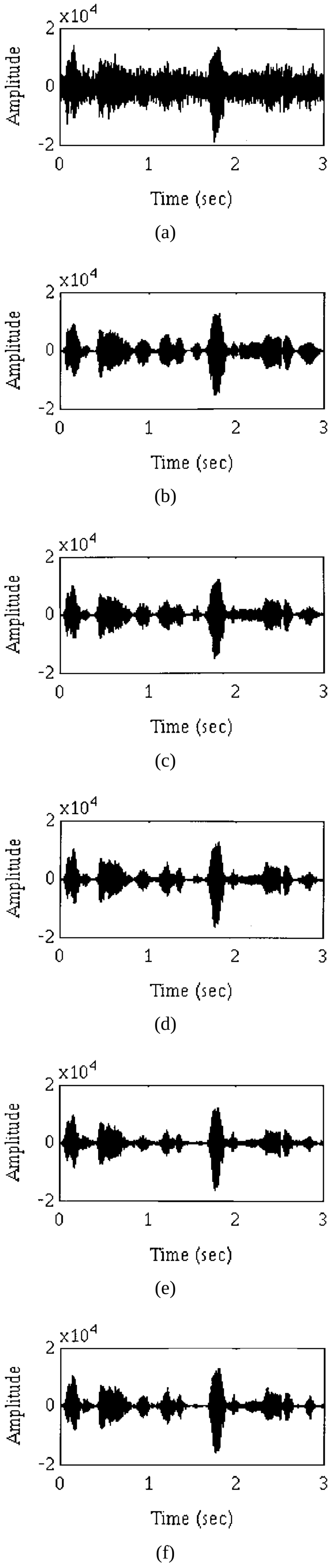
2) Subjective Evaluation Tests: For the subjective evaluation, two tests were performed. The first test, at word level, was the diagnostic rhyme test (DRT) [31], whereas the second test, at sentence level, was the semantically unpredictable sentences (SUS) test [32]. From those, the DRT was performed on Greek and English-language speech data, while the SUS test was performed only on Greek-language speech data. Note that both the DRT and a restricted form of the SUS test have been used for the evaluation of many speech enhancement techniques [10], [13], [25], [26]. A limited two-speaker (one male and one female) DRT test in English was performed using six listeners and 96 word-pairs. The speakers were native English speakers, while all listeners were either native English speakers or had extensive knowledge of the English language. This test was mainly performed in order to be able to compare its results with the corresponding Greek-language DRT test. For the Greek-language DRT, the word-pair material was created from two-syllable words drawn out of two Greek lexicons and by converting all material to phonetic form. A total of 192 word-pairs (384 words) were finally used. This material was spoken by four speakers (two male and two female) having normal Greek accents. A total of 20 subjects participated in the test. For the SUS, test sentences based on five syntactical structures were created using a corpus of over 10 million words. Finally, a total of 80 sentences were used for the training and the evaluation session. All sentences were spoken by four speakers (two male and two female) and a total of 20 subjects participated in the test.

C. Results

Typical time-domain plots for the ANS technique are shown in Fig. 9, which illustrates the significant noise suppression effect of the method.

Objective results were obtained for the complete test data base created for the described intelligibility tests and are presented in Fig. 10. These results are plotted for the Greek-language speech data DRT (G-DRT), the English-language speech data DRT (E-DRT), and the SUS test (SUS), for various initial SNR conditions (i.e., $-\infty$, -5, 0, 5 dB). At each initial SNR condition, the following processing categories are included: "D" for the "debug" approach, "N" for the noisy signal, "T" for the "threshold" approach, and "M" for the "minima" approach. From these results, the following observations can be made.

1) There are no significant differences with respect to the type of speech material used for the objective tests (i.e., DRT or SUS).
2) As expected, the best results were obtained for the debug condition, indicating also the validity of the proposed psychoacoustic and sparse data model. This is also obvious from the SNR $= -\infty$ dB results.
3) In all cases, improvements were measured by the use of the two types of sparse-data estimators, with the threshold approach having a small advantage over the minima approach for most conditions, and particularly for the NMR tests.
4) For most cases, the proposed estimation methods achieved results close to the debug "D" method, with typical SNR improvement of 10 dB and typical NMR improvement of 20 dB.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 9. Time domain plots for a typical sentence. (a) Noisy speech (SNR = 0 dB). (b) Noise-free speech. (c) ANS limit (16). (d) ANS by "debug" parameters. (e) ANS by "minima" parameters. (f) ANS by "threshold" parameters.

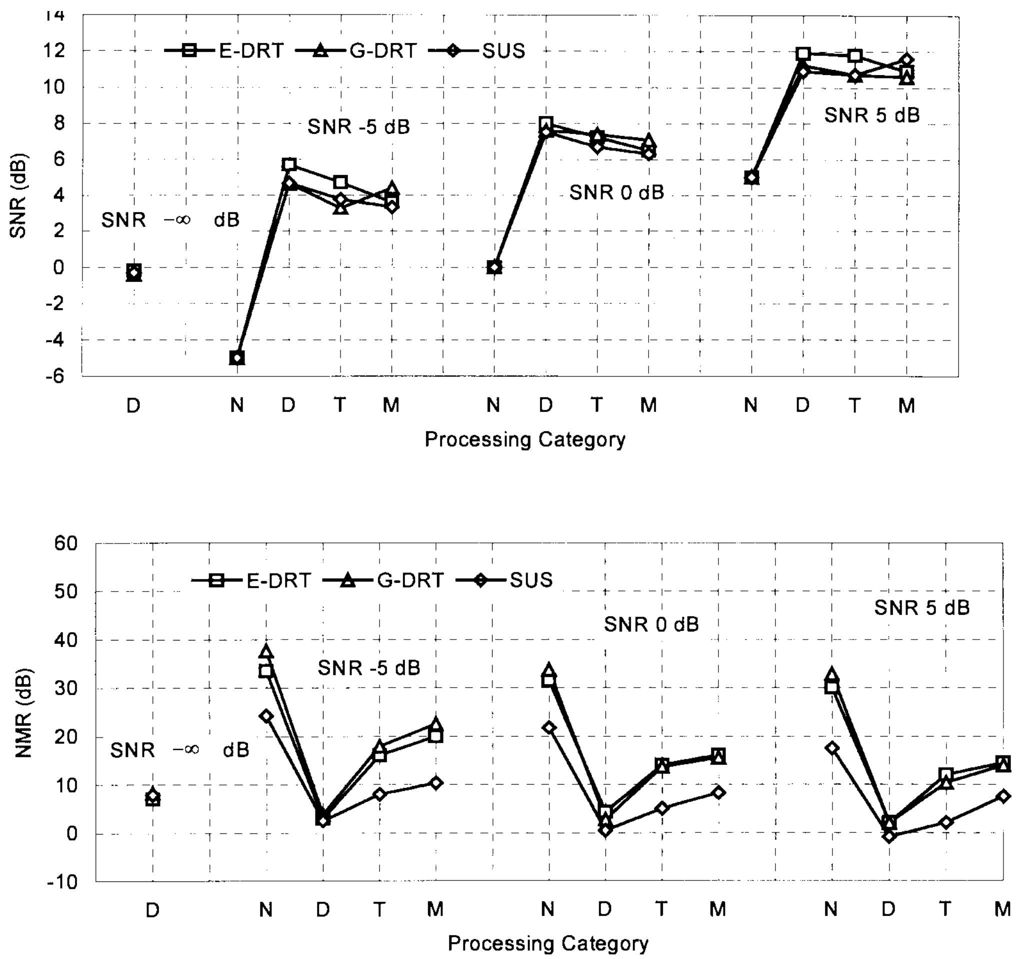


Fig. 10. Objective ANS method performance for the English language speech data DRT (E-DRT), the Greek language speech data DRT (G-DRT), and the SUS test. Initial SNR condition is also indicated for each curve. The horizontal axis denotes the processing category, where "N" stands for the noisy signal, "D" for the "debug" method, "T" for the "threshold" approach, and "M" for the "minima" approach (see text). (a) SNR performance. (b) NMR performance.

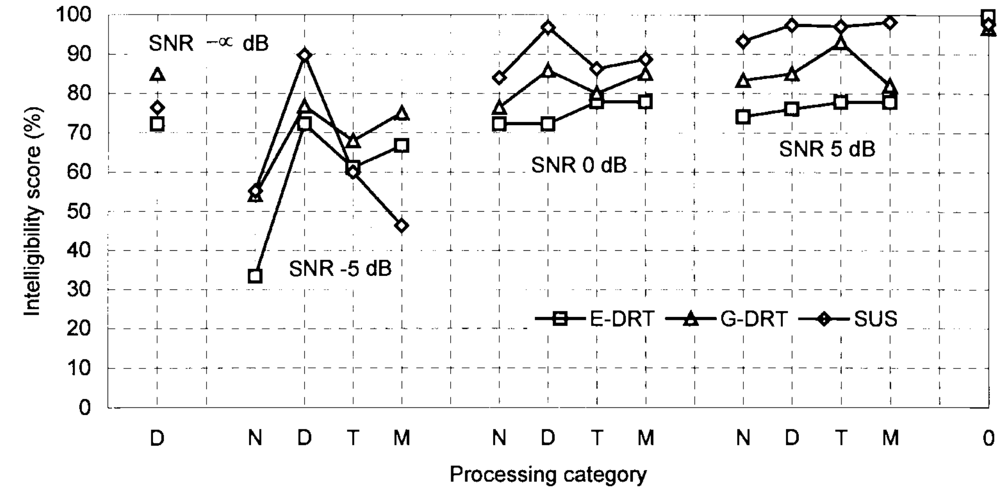
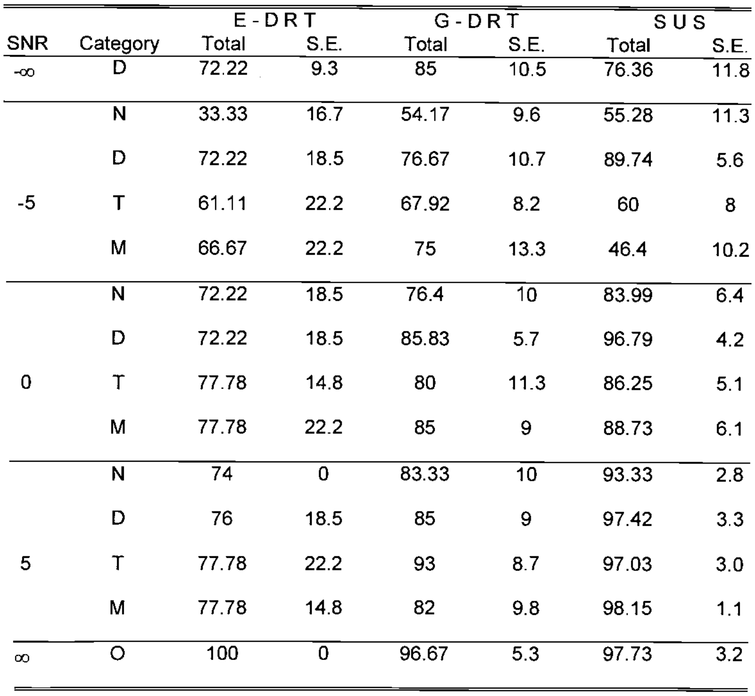


Fig. 11. Intelligibility scores for the English language speech data DRT (E-DRT), the Greek language speech data DRT (G-DRT), and the SUS test. Initial SNR condition is also indicated for each curve. The horizontal axis denotes the processing category, where "N" stands for the noisy signal, and "O" for the noise-free signal.

These objective improvements were also confirmed to a large extent by the subjective tests, as is shown by the results of Fig. 11 and Table II, where the standard error (SE) among the individual listeners scores is also included. For all the above results, an additional category is also included, that of the noise-free speech signal, denoted by "O." From these results, the following observations can be made.

1) The debug method, for initial SNR $= -\infty$ dB, achieved scores of 72.22% (for E-DRT), 85% (for G-DRT), and 73.36% (for SUS), indicating again the validity of the

TABLE II
INTELLIGIBILITY SCORES AND LISTENER STANDARD ERROR (SE) FOR THE ENGLISH LANGUAGE SPEECH DATA DRT (E-DRT), THE GREEK LANGUAGE SPEECH DATA DRT (G-DRT), AND THE GREEK LANGUAGE SPEECH DATA SUS TEST PER INITIAL SNR VALUE AND PROCESSING CATEGORY



		E-DRT		G-DRT		SUS	
SNR	Category	Total	S.E.	Total	S.E.	Total	S.E.
-∞	D	72.22	9.3	85	10.5	76.36	11.8
-5	N	33.33	16.7	54.17	9.6	55.28	11.3
	D	72.22	18.5	76.67	10.7	89.74	5.6
	T	61.11	22.2	67.92	8.2	60	8
	M	66.67	22.2	75	13.3	46.4	10.2
0	N	72.22	18.5	76.4	10	83.99	6.4
	D	72.22	18.5	85.83	5.7	96.79	4.2
	T	77.78	14.8	80	11.3	86.25	5.1
	M	77.78	22.2	85	9	88.73	6.1
5	N	74	0	83.33	10	93.33	2.8
	D	76	18.5	85	9	97.42	3.3
	T	77.78	22.2	93	8.7	97.03	3.0
	M	77.78	14.8	82	9.8	98.15	1.1
∞	O	100	0	96.67	5.3	97.73	3.2

proposed ANS model and also that the method can also be used for speech reconstruction (e.g., for data compression applications), using noise excitation and the proposed nonlinear enhancement filter fed by sparse data parameters derived from noise-free speech. This result indicates that the intelligible, psychoacoustically significant bit rate of speech can be very low, but it is also believed that the above scores can be further improved by the use of additional voicing (pitch) information and by minimization of the spectral difference between reconstructed and source speech, adjusting the parameter $\nu_b(i)$ per data window and critical band.

2) The debug method achieved also intelligibility improvement for all other SNR conditions, although these improvements were smaller for the better initial SNR's. Specifically, at SNR $= -5$ dB, the debug method improvements were 22% (for G-DRT), 38.89% (for E-DRT), and 34.46% (for SUS). The smaller improvements at SNR $=$ 0 and 5 dB were somewhat expected, given the satisfactory initial (noisy speech) intelligibility.

3) The proposed estimators achieved intelligibility improvements for most conditions and tests. These improvements were larger for lower initial SNR's (mainly for the previously explained reasons), and were lower than those achieved by the debug method, indicating that there is further scope for improving the parameter estimation process of the ANS method. Specifically, at SNR $= -5$ dB, the DRT intelligibility improvement was better for the minima method with 33.34% (for E-DRT) and 20.83% (for G-DRT), the threshold method achieved improvements of 13.75% (for G-DRT) and 27.78% (for E-DRT). At this condition, the SUS test was less successful, with a small 4.72% improvement for the threshold method and an intelligibility degradation for the minima method. At higher SNR's, some intelligibility improvements were also measured, except for the case of SNR $=$ 5 dB, where intelligibility degradation was measured for G-DRT. Nevertheless, it is believed that these results have smaller significance due to the already fair signal presentation combined with the possibility of statistical errors, due to the relatively small scale of the tests.

VI. CONCLUSIONS

A novel speech enhancement technique was developed, analyzed, and tested. The technique relies on the definition of the psychoacoustic quantity of audible noise, derived from the signal's STSA. This quantity describes the amount of noise perceived as degradation by the auditory mechanism (inner ear) and it is shown that its suppression can lead to objectively and subjectively enhanced speech.

The main advantages of the proposed approach over previously developed enhancement methods, are derived from the selective and limited number of spectral regions specified for processing. At one hand, this minimizes the processing artifacts and at the other hand, as was shown, this approach leads to reduced requirements for the *a priori* known or estimated clean speech data. The required audible noise suppression was achieved by the introduction of a flexible frequency-domain nonlinear filter, whose time-varying parameters were derived from such sparse data estimates. These estimates were shown to be as many as the number of CB's (per data window), and were found to be either the spectral minima, or alternatively, the masking threshold value. For each approach, a suitable estimation procedure was also derived, allowing parameter extraction from noisy data.

The most significant result that has emerged from the above analytic and experimental procedure is that only a limited and small number of psychoacoustically derived spectral data (per data window) is required to reconstruct intelligible speech, irrespective of the initial SNR condition. It is then up to the development of suitable estimators that can extract these sparse-data from the noisy signal. A secondary finding of this work was the definition of the lower, psychoacoustically derived intelligible speech reconstruction bit rate, which can be achieved when the ANS technique is driven by noise excitation and clean-speech sparse data.

The objective and subjective tests described support the above statements. Specifically, a general agreement was found between objective and subjective tests, and in all cases significant improvements were achieved by the ANS technique, given correct sparse data (debug method). These were larger for low initial SNR's (e.g., -5 dB), where intelligibility improvements approaching 40% were measured, although these were smaller for better initial SNR conditions. Smaller but significant improvements were also measured when the noisy speech signal alone was used for the extraction of the enhancement parameters, with intelligibility improvement of up to 33% for the DRT and initial SNR $= -5$ dB.

In terms of computational complexity, the ANS technique requires calculation of two FFT's, estimation of the AMT (or alternatively, estimation of the spectral minimum per CB), and some simple arithmetic operations. This computational load was found to be approximately 1.5 times the real duration of the speech data when implemented on a PC-486 type computer. Therefore, implementation of the ANS method may be possible in real-time on a general purpose DSP board.

Nevertheless, the significantly lower performance of the ANS method for estimated parameters (compared to the debug condition) indicates that there is further scope for development in the parameter estimation procedure. Furthermore, it is believed that the ANS technique would be improved if a suitable model existed for estimation of the clean signal's masking threshold from the noisy properties and the noisy speech signal, given that the current technique relies on a rather heuristic AMT estimator. Furthermore, the speech reconstruction technique that has emerged from the ANS method can be further improved by further investigations into the form of nonlinear filter and also in the excitation input signal properties. Finally, another possible area of improvement would be for applications when the statistics of the speech (i.e., after analysis of the speaker's data) and/or the noise are known in advance and used for optimal adjustment of the ANS estimators.

Appendix A

The algorithm for the estimation of the AMT $T_b(i)$ is briefly described here, although a more detailed description can be found in [23]. First, the total power of the spectrum of the signal per CB is found as follows:

$$Q_b(i) = \sum_{k=k_{lb}}^{k_{hb}} X_p(k, i), \qquad 0 \le b \le B-1 \tag{A.1}$$

where, k_{lb} and k_{hb} are the lower and upper limits of CB b, B is the total number of CB's, and $X_p(k, i)$ is the power spectrum of the speech signal. The total power spectrum per CB is then convolved with the basilar membrane spreading function $\mathrm{Sp}(\;)$, which provides information on masking of signals by signals in the bark domain, as follows:

$$C_b(i) = \sum_{m=1}^{B} \mathrm{Sp}\,(b-m+25)Q_m(i), \qquad 0 \le b \le B-1. \tag{A.2}$$

The noiselike or tonelike nature of the signal is determined by the statistical characteristics of the power spectrum and is mathematically given by the spectral flatness measure (SFM):

$$\mathrm{SFM} = \frac{G(i)}{A(i)}, \qquad \mathrm{SFM}_{\mathrm{dB}} = 10 \log_{10} \mathrm{SFM} \tag{A.3}$$

where $G(i)$ and $A(i)$ are the respective geometric and arithmetic means of the signal's power spectrum. From this measure, the tonality of the signal is found using

$$\mathrm{ton}\,(i) = \min\left\{\frac{\mathrm{SFM}_{\mathrm{dB}}}{\mathrm{SFM}_{\max}}, 1\right\} \tag{A.4}$$

where $\mathrm{SFM}_{\max} = -60$ is defined as the SFM value of a sine wave. Therefore, $\mathrm{ton}\,(i) = 1$ for $\mathrm{SFM} = \mathrm{SFM}_{\max}$ (sine wave input), whereas $\mathrm{ton}\,(i) = 0$ for $\mathrm{SFM} = 0$ (white noise input).

An offset is then estimated by which the threshold has to be reduced in order to take into account the signal tonality

$$O_b(i) = \mathrm{ton}\,(i)(14.5 + b) + (1 - b)5.5, \qquad 0 \le b \le B-1. \tag{A.5}$$

The auditory masking threshold can now be calculated using

$$T_b(i) = 10^{\log_{10} C_b(i) - O_b(i)/10}, \qquad 0 \le b \le B-1. \tag{A.6}$$

Finally, normalization and comparison to the absolute auditory threshold is performed.

Appendix B

Consider minimization of the MSE of the audible noise spectrum $\hat{A}_d(k, i)$ over some constant parameter $a(i)$, i.e.,

$$\min_{a(i)}\left\{\sum_k \hat{A}_d^2(k, i)\right\} \tag{B.1}$$

where, it is assumed that the enhanced speech power spectrum $\hat{X}_p(k, i)$ depends on $Y_p(k, i)$ and $a(i)$. From (B.1), it follows that the MMSE solution is given by

$$\frac{\partial}{\partial a(i)} \sum_k \hat{A}_d^2(k, i) = 0. \tag{B.2}$$

By substituting $[\hat{X}_p(k, i) - \mathbf{X}_p(k, i)]$ for $\hat{A}_d(k, i)$, where $\mathbf{X}_p(k, i)$ is equal to either $X_p(k, i)$ [Branch I of (9)] or $T(k, i)$ [Branch II of (9)], (B.2) becomes

$$\frac{\partial}{\partial a(i)} \sum_k [\hat{X}_p(k, i) - \mathbf{X}_p(k, i)]^2 = 0$$

or

$$\sum_k [\hat{X}_p(k, i) - \mathbf{X}_p(k, i)] \frac{\partial \hat{X}_p(k, i)}{\partial a(i)} = 0. \tag{B.3}$$

Given that in general $\partial \hat{X}_p(k, i)/\partial a(i)$ cannot be zero, one solution will require that

$$\hat{X}_p(k, i) - \mathbf{X}_p(k, i) = 0, \qquad \text{for } 0 \leq k \leq M$$

or

$$\hat{A}_d(k, i) = 0, \qquad \text{for } 0 \leq k \leq M \tag{B.4}$$

where M is an arbitrary spectral region.

Consequently, (B.4) is also an MMSE solution for the audible noise removal problem.

Appendix C

Branch (I) of (23) can be also written as

$$\begin{aligned} a_{\mathrm{I}\,b}(i) = & D_{pb}^{1+1/\nu_b(i)} X_p^{-1/\nu_b(i)}(k_{\mathrm{I}}, i) \\ & + D_{pb}^{1/\nu_b(i)} X_p^{1-1/\nu_b(i)}(k_{\mathrm{I}}, i) \end{aligned} \tag{C.1}$$

from where it is clear that if $0 < \nu_b(i) \leq 1$, then $1 - 1/\nu_b(i) \leq 0$, so that $a_{\mathrm{I}\,b}(i)$ is inversely proportional to $X_p(k, i)$ and, hence, the maximum $a_{\mathrm{I}\,b}(i)$ corresponds to the minimum $X_p(k_{\mathrm{I}}, i)$ for those speech components above the AMT. If, however, $\nu_b(i) > 1$, then $X_p(k_{\mathrm{I}}, i)$ is not necessarily the minimum spectral component in CB b.

From Branch (II) of (23) it is clear that $a_{\mathrm{II}\,b}(i)$ is proportional to $X_p(k_{\mathrm{II}}, i)$, and, therefore, $X_p(k_{\mathrm{II}}, i)$ corresponds to the maximum spectral component in CB b for those components below the AMT.

Appendix D

It will be shown that $a'_b(i)$ given by (24) is greater than or equal to $a_{\mathrm{I}\,b}(i)$ and $a_{\mathrm{II}\,b}(i)$ given by (23).

At first, it is easy to notice that since Branch (I) of (23) takes its maximum value by the minimum $X_p(k, i)$, for those values of $X_p(k, i)$ above the AMT $T_b(i)$, as was shown in Appendix C, it takes a larger value by the minimum $X_{pb,\,\min}(i)$ for all components within the CB, irrespective of whether these components fall above or below the AMT $T_b(i)$. Therefore, provided that

$$X_{pb,\,\min}(i) \leq X_p(k_{\mathrm{I}}, i) \tag{D.1}$$

and also that $0 < \nu_b(i) \leq 1$, then

$$\begin{aligned} & [D_{pb} + X_{pb,\,\min}(i)] \left[\frac{D_{pb}}{X_{pb,\,\min}(i)}\right]^{1/\nu_b(i)} \\ & \quad \geq [D_{pb} + X_p(k_{\mathrm{I}}, i)] \left[\frac{D_{pb}}{X_p(k_{\mathrm{I}}, i)}\right]^{1/\nu_b(i)} \end{aligned} \tag{D.2}$$

i.e., Branch (I) of (25) is satisfied. Assume now that there exist frequency components $X_p(k, i)$ within CB b below the AMT $T_b(i)$, i.e.,

$$X_p(k_{\mathrm{II}}, i) < T_b(i) \tag{D.3}$$

so that Branch (II) of (23) has to be taken into account. Consider now the expression $a'_b(i) - a_{\mathrm{II}\,b}(i)$, which by using (D.3) can be also written as

$$\begin{aligned} A' = & [D_{pb} + X_{pb,\,\min}(i)] \left[\frac{D_{pb}}{X_{pb,\,\min}(i)}\right]^{1/\nu_b(i)} \\ & - [D_{pb} + X_p(k_{\mathrm{II}}, i)] \left[\frac{D_{pb} + X_p(k_{\mathrm{II}}, i)}{T_b(i)} - 1\right]^{1/\nu_b(i)} \\ > & [D_{pb} + X_{pb,\,\min}(i)] \left[\frac{D_{pb}}{X_{pb,\,\min}(i)}\right]^{1/\nu_b(i)} \\ & - [D_{pb} + T_b(i)] \left[\frac{D_{pb}}{T_b(i)}\right]^{1/\nu_b(i)}. \end{aligned} \tag{D.4}$$

From (D.1) it is clear that $X_{pb,\,\min}(i) < T_b(i)$ and consequently (as shown in Appendix C) $A' = a'_b(i) - a_{\mathrm{II}\,b}(i) > 0$. Therefore, the second branch of (25) is satisfied.

Appendix E

It will be shown that $a''_b(i)$ given by (26) satisfies the conditions implied by (27). At first, it is easy to notice that since Branch (I) of (23) is satisfied by the minimum $X_p(k_{\mathrm{I}}, i)$, for those values of $X_p(k, i)$ above the AMT $T_b(i)$, it is also satisfied by the AMT $T_b(i)$, so that Branch (I) of (27) is satisfied.

Consider now that there are frequency components below the AMT so that (D.3) is valid. The quantity $a''_b(i) - a_{\mathrm{II}\,b}(i)$, which can be written as

$$\begin{aligned} A'' = & [D_{pb} + T_b(i)] \left[\frac{D_{pb}}{T_b(i)}\right]^{1/\nu_b(i)} \\ & - [D_{pb} + X_p(k_{\mathrm{II}}, i)] \left[\frac{D_{pb} + X_p(k_{\mathrm{II}}, i)}{T_b(i)} - 1\right]^{1/\nu_b(i)} \\ > & [D_{pb} + X_{pb,\,\min}(i)] \left[\frac{D_{pb}}{X_{pb,\,\min}(i)}\right]^{1/\nu_b(i)} \end{aligned}$$

$$- [D_{pb} + X_p(k_{\mathrm{II}}, i)] \left[\frac{D_{pb} + X_p(k_{\mathrm{II}}, i)}{T_b(i)} - 1 \right]^{1/\nu_b(i)} \tag{E.1}$$

Since, however, $X_{pb,\,\min}(i) < X_p(k_{\mathrm{II}}, i)$, it is concluded that $A'' = a_b''(i) - a_{\mathrm{II}\,b}(i) > 0$ and, therefore, the second branch of (27) is satisfied.

Appendix F

By substituting (30) and (31) into (32) and using [27, Eq. 3.462.1], we obtain

$$\hat{X}_{b,\,\min} = \sqrt{2\lambda_b}\, \frac{\mathbf{D}_{-3}\left(-\sqrt{2\lambda}\, \dfrac{\overline{X}_b}{\lambda_{b,\,\overline{X}}}\right)}{\mathbf{D}_{-2}\left(-\sqrt{2\lambda}\, \dfrac{\overline{X}_b}{\lambda_{b,\,\overline{X}}}\right)} \tag{F.1}$$

where $\mathbf{D}_p(\ )$ are parabolic cylinder functions [27, Eq. 9.240], and, λ_b is given by

$$\frac{1}{\lambda_b} = \frac{1}{\lambda_{b,\,\min}} + \frac{1}{\lambda_{b,\,\overline{X}}}. \tag{F.2}$$

By using [27, Eqs. 9.247.1, 9.254.1, and 9.254.2], (F.1) can be written as

$$\hat{X}_{b,\,\min} = \sqrt{\frac{\lambda_b}{2}} \left(\frac{1 - \Phi\left(\dfrac{z}{\sqrt{2}}\right)}{\sqrt{\dfrac{\pi}{2}} \exp\left\{\dfrac{-z^2}{2}\right\} - z\left[1 - \Phi\left(\dfrac{z}{\sqrt{2}}\right)\right]} - z \right) \tag{F.3}$$

where $z = -\sqrt{2\lambda_b}(\overline{X}_b/\lambda_{b,\,\overline{X}})$.

Then, by using (34) and (35), (F.3) can be written as

$$\hat{X}_{b,\,\min} = \frac{1}{\sqrt{2}} \sqrt{\left(\frac{1}{1 + Z_{b,\,\mathrm{post}}}\right)\left(\frac{Z_{b,\,\mathrm{prio}}}{1 + Z_{b,\,\mathrm{prio}}}\right)} \cdot \mathbf{M}\left[-\sqrt{2}\sqrt{(1 + Z_{b,\,\mathrm{post}})\left(\frac{Z_{b,\,\mathrm{prio}}}{1 + Z_{b,\,\mathrm{prio}}}\right)}\right] \overline{X}_b. \tag{F.4}$$

References

[1] H. Frazier, S. Samsam, L. D. Braida, and V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. IEEE ICASSP,* Apr. 1976, pp. 251–253.

[2] W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.,* vol. 60, pp. 911–918, Oct. 1976.

[3] R. Sambur, "Adaptive noise canceling for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-26, pp. 419–423, Oct. 1978.

[4] S. Lim, "All pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-26, pp. 197–210, June 1978.

[5] J. Done and C. K. Rushforth, "Estimating the parameters of a noisy all-pole process using pole-zero modeling," in *Proc. IEEE ICASSP,* Apr. 1979, pp. 228–231.

[6] B. R. Musicus and J. S. Lim, "Maximum likelihood parameters estimation of noisy data," in *Proc. IEEE ICASSP,* Apr. 1979, pp. 224–227.

[7] H. Kobatake, J. Inari, and S. Kakuta, "Linear predictive coding of speech signals in a high ambient noise environment," in *Proc. IEEE ICASSP,* June 1978, pp. 472–475.

[8] M. R. Sambur, "A preprocessing filter for enhancing LPC analysis/synthesis of noisy speech," in *Proc. IEEE ICASSP,* Apr. 1979, pp. 971–974.

[9] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing,* vol. 39, pp. 795–805, Apr. 1991.

[10] 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.

[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP,* Apr. 1979, pp. 208–211.

[12] 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.

[13] R. D. Preuss, "A frequency domain noise canceling preprocessor for narrowband speech communications systems," in *Proc. IEEE ICASSP,* Apr. 1979, pp. 212–215.

[14] J. H. L. Hansen, "Speech enhancement employing adaptive boundary detection and morphological based spectral constraints," in *Proc. IEEE ICASSP,* 1991, pp. 901–904.

[15] 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.

[16] P. Vary, "Noise suppression by spectral amplitude estimation—Mechanism and theoretical limits," *Signal Process.,* vol. 8, pp. 387–400, July 1985.

[17] T. Langhans and W. Strube, "Speech enhancement by nonlinear multiband envelope expansion," in *Proc. IEEE ICASSP,* 1982, pp. 156–159.

[18] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Processing,* vol. 39, pp. 1943–1954, Sept. 1991.

[19] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models.* New York: Springer-Verlag, 1990.

[20] J. Mourjopoulos, G. Kokkinakis, and M. Paraskevas, "Noisy audio signal enhancement using subjective spectra," in *Proc. 92nd AES Conv.,* Mar. 1992.

[21] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psychoacoustic criteria," in *Proc. IEEE ICASSP,* Apr. 1993, pp. 359–362.

[22] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE,* vol. 66, pp. 51–83, Jan. 1978.

[23] J. D. Johnston, "Transform coding of audio signal using perceptual noise criteria," *IEEE J. Select. Areas Commun.,* vol. 6, pp. 314–323, Feb. 1988.

[24] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.,* vol. 42, pp. 780–792, 1994.

[25] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE,* vol. 67, pp. 1586–1604, Dec. 1979.

[26] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-26, pp. 471–472, Oct. 1978.

[27] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* New York: Academic, 1994.

[28] NOISEX-92 CD-ROM's, NATO: AC243/(Panel 3)/RSG.10, ESPRIT Project 2589 (SAM), Speech Research Unit, Defence Research Agency, U.K., 1992.

[29] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

[30] J. Herre, E. Eberlein, H. Schott, and K. Brandenburg, "Advanced audio measurement system using psychoacoustic properties," in *Proc. 92nd AES Conv.,* Mar. 1992.

[31] S. Meister, "The diagnostic rhyme test (DRT): An Air Force implementation," Rep. RADC-TR-78-129, AD-A060917, U.S. Air Force, 1978.

[32] Benoit, M. Grice, and V. Hazan, "A manual for the SUS test: A unified methodology for multilingual text-to-speech synthesis assessment at the sentence level," ESPRIT Project 2589 (SAM), Ref. no. SAM-ICP-UCL-001, Apr. 1991.

Dionysis E. Tsoukalas was born in Athens, Greece, in 1966. He received the B.Sc. degree in electrical engineering from the University of Patras, Patras, Greece.

Since then, he has been a postgraduate research student, working at the Wire Communications Laboratory, University of Patras, where he is currently pursuing the Ph.D. degree. His current research interests include noise removal and reduction from speech and audio signals, adaptive signal processing, and neural networks in audio applications. He has worked on a number of research projects focusing on the development of speech production workstations, audio workstations for broadcast applications, and audio system analysis tools.

Mr. Tsoukalas is a member of the Technical Chamber of Greece.

John N. Mourjopoulos (M'90) was born in Greece in 1954. He received the B.Sc. degree from Lanchester Polytechnic (now Coventry University), Coventry, U.K., in 1978, and the the M.Sc. and Ph.D. degrees from the Institute of Sound and Vibration Research (ISVR), University of Southampton, Southampton, U.K., in 1980 and 1985, respectively. His dissertation centered on the areas of acoustics and digital signal processing.

From 1984 to 1985, he was a Research Assistant at ISVR, and from 1986 to 1987, he was a Research Assistant at the Wire Communications Laboratory, Electrical Engineering Department, University of Patras, Patras, Greece, where he is currently an Assistant Professor in electroacoustics. His research interests are in the areas of digital audio and acoustic system equalization, audio signal processing, coding and enhancement, as well as speech processing and recognition. In these fields, he has published more than 50 technical papers, articles, and reports.

Dr. Mourjopoulos was co-organizer of seminars in England and Greece, and is a Member of the AES and the Hellenic Acoustical Society.

George Kokkinakis was born in Chios, Greece, on March 17, 1937. He received the Dipl.-Ing. in 1961, the Dr.-Ing. in 1966, and the Diploma in engineering economics (Dipl. Wirt.-Ing.), all from the Technical University of Munich, Munich, Germany.

During 1968 and 1969, he served at the Ministry of Coordination in Athens, Greece. Since 1969, he has been with the Department of Electrical Engineering at the University of Patras, Patras, Greece, where he has organized and is directing the Wire Communications Laboratory (WCL). His current activity in research and development, which coincides with activity of WCL, includes the design and optimization of telecommunication networks, and the analysis, synthesis, recognition, and linguistic processing of the Greek language. He has published several books and more than 100 technical papers, articles, and reports on telecommunication, electrotechnology, and speech technology.

Prof. Kokkinakis is a Member of the Technical Chamber of Greece, the Verein Deutscher Elektrotechniker, the European Speech Communication Association, the European Association for Signal Processing, the Societé Europeenne pour la Formation des Ingenieurs, the Greek Operations Research Society, and the Linguistics Society of America.