IMPROVING THE INTELLIGIBILITY OF NOISY SPEECH USING AN AUDIBLE NOISE SUPPRESSION TECHNIQUE

D. E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis
Wire Communications Laboratory
Electrical & Computer Engineering Dept.
University of Patras, 261 10, Greece
Tel: +30 61 991722, FAX: +30 61 991855, E-mail: tsoukala@wcl.ee.upatras.gr

ABSTRACT

This paper presents some novel results concerning the problem of enhancing speech degraded by wideband additive noise. The enhancement scheme proposed in this work is based on the utilisation of the Auditory Masking mechanism as a measure for the definition and subsequent suppression of the frequency audible noise components. Accordingly, the enhancement technique minimises only those noise components responsible for audible signal degradations, so that the underlying speech signal quality is only minimally degraded. Extensive subjective and objective tests have shown that, after enhancement, the intelligibility of the processed signal can be improved even at very low S/N ratios.

1. INTRODUCTION

Although during the past decades several methods have been proposed [1], the area of Speech Enhancement remains largely open, since there are still many potential applications for real-environment human-machine communication, telecommunications, etc., in need for techniques with improved performance.

Such speech enhancement techniques can be viewed as procedures used to improve different perceptual aspects of the speech signal (i.e. quality, intelligibility, S/N ratio). Here, as will be shown, such a technique is presented which can improve the intelligibility of noisy speech.

The enhancement scheme used in this case is based on the utilisation of the Auditory Masking mechanism [2], which has successfully been used in a number of coding applications [3], and in earlier works by the authors [4],[5]. According to the proposed approach, only the audible noise components are estimated and suppressed, in such a way so that degradations in the underlying speech signal are minimised.

Extensive objective and subjective tests, are discussed in Section 3, and show that the proposed enhancement method is capable of improving Speech Intelligibility, especially at very low Signal to Noise ratios.

2. SPEECH ENHANCEMENT USING AUDIBLE NOISE SUPPRESSION

In this Section, the enhancement method is briefly described, while more details can be found in [5].

Let, a speech signal $x(n)$ be contaminated by additive wideband stationary noise $d(n)$, to produce the noisy speech signal $y(n)$:

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

Let, also the Short Time Fourier transforms of the noisy and original speech signals be given by $Y_w(k,i)$ and $X_w(k,i)$ respectively, i.e.:

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

$$X_w(k,i) = \sum_{n=0}^{K\text{-}1} x(n+off_i)\, w(n)\, I_K^{kn}\ ,\ 0 \le k \le K\text{-}1 \qquad (2.b)$$

where $I_K^{kn}=e^{-j(2\pi kn/K)}$, $w(n)$ is a window function, k,i are the frequency and time domain indexes respectively, and, off_i is an offset.

The corresponding power spectra are given by

$$Y_p(k,i)=\left|Y_w(k,i)\right|^2 \qquad (3.a)$$

$$X_p(k,i)=\left|X_w(k,i)\right|^2 \qquad (3.b)$$

Let, also, the Auditory Masking Threshold (AMT) [2] of the clean signal be given by $T(k,i)$ as can be easily calculated according to [3]. This threshold is responsible for masking of signals by other stronger signals in the frequency domain and has been extensively used in signal compression [3].

According to these definitions, let us define the Audible signal spectrum of the noisy and clean speech, i.e. $A_y(k,i)$ and $A_x(k,i)$ respectively:

$$A_y(k,i)=\max\{Y_p(k,i),T(k,i)\} \qquad (4.a)$$

$$A_x(k,i)=\max\{X_p(k,i),T(k,i)\} \qquad (4.b)$$

where, $\max\{a,b\}=\begin{cases} a, & if\ a \ge b \\ b, & if\ a < b \end{cases}$

From eq. (4), it is clear that only frequencies above the AMT contribute to the Audible signal components.

The Audible Noise Components can, in turn, be defined as the difference between the Audible noisy and clean speech spectra:

$$A_d(k,i)=A_y(k,i)\text{-}A_x(k,i) \qquad (5)$$

$A_d(k,i)$ is a four branch function, depending on the relative levels of the Noisy and clean speech power spectra and the AMT, i.e.:

$$A_d(k,i)=\begin{cases} Y_p(k,i)\text{-}X_p(k,i),\ if \left\{\begin{matrix} Y_p(k,i)\geq T(k,i) \\ X_p(k,i)\geq T(k,i)\end{matrix}\right\} & (I) \\ Y_p(k,i)\text{-}T(k,i),\ if \left\{\begin{matrix} Y_p(k,i)\geq T(k,i) \\ X_p(k,i)< T(k,i)\end{matrix}\right\} & (II) \\ T(k,i)\text{-}X_p(k,i),\ if \left\{\begin{matrix} Y_p(k,i)< T(k,i) \\ X_p(k,i)\geq T(k,i)\end{matrix}\right\} & (III) \\ 0,\ if \left\{\begin{matrix} Y_p(k,i)< T(k,i) \\ X_p(k,i)< T(k,i)\end{matrix}\right\} & (IV)\end{cases} \quad (6)$$

This quantity is responsible for Audible degradations in the noisy speech signal. Therefore, appropriate modifications of the noisy speech power spectrum according to this quantity may lead to psychoacoustically - optimised signal enhancement.

Let us propose that the psychoacoustic enhancement method is based on minimisation of the Audible Noise function, according to the expression:

$$\hat{A}_d(k,i)=A_y(k,i)\text{-}A_{\hat{x}}(k,i)\leq 0 \quad (7)$$

where $A_{\hat{x}}(k,i)$ is the audible spectrum calculated after considering the enhanced (i.e. after modification) speech power spectrum $\hat{X}_p(k,i)$.

From eq. (6), only branches (I) and (II) must be taken into account in the enhancement process, since branches (III) and (IV) are by definition below or equal to zero. To modify the noisy speech power spectrum, a non-linear time-frequency variant filter has been chosen, given by the following expression:

$$\hat{X}_p(k,i)=\frac{Y_p^{v(k,i)}(k,i)}{a^{v(k,i)}(k,i)+Y_p^{v(k,i)}(k,i)}\ Y_p(k,i) \quad (8)$$

where the time-frequency varying parameters $a(k,i)$ and $v(k,i)$ must be appropriately estimated so that eq. (7) is valid.

In order to simplify eq. (8), it will be assumed that $v(k,i)=1$. Then by using eq. (6), (7) and (8) it can be shown [5] that two solutions for estimating parameter $a(k,i)$ can exist, which are based on sparse spectral speech data, specifically (a) the spectral minima, and (b) the values of the AMT per critical band.

(a) Speech Enhancement based on spectral minima

Given that $X_{pb,\min}(i)$ is the minimum power spectral component of the speech signal in critical band b, then a proper estimate of $a(k,i)$ for eq. (8) which satisfies eq. (7) is [5]:

$$a'_b(i)=(D_{pb}+X_{pb,\min}(i))\left(\frac{D_{pb}}{X_{pb,\min}(i)}\right) \quad (9)$$

where, D_{pb} is the mean noise power per critical band b, and , $a'_b(i)$ is the value of $a(k,i)$ in critical band b.

(b) Speech Enhancement based on AMT values

Another way to estimate $a(k,i)$ is to use the AMT values $T_b(i)$ per critical band which can lead to the following expression [5]:

$$a''_b(i)=(\ D_{pb}+T_b(i)\ )\left(\frac{D_{pb}}{T_b(i)}\right) \quad (10)$$

3. IMPLEMENTATION AND EVALUATION

3.1. Technical Considerations

The described Audible Noise Suppression (ANS) technique was simulated on a general purpose computer using pre-recorded speech and noise data at several S/N ratios, 16 kHz sampling frequency and 16 bit A/D conversion.

The general block diagram of the ANS technique is shown in Fig. 1.

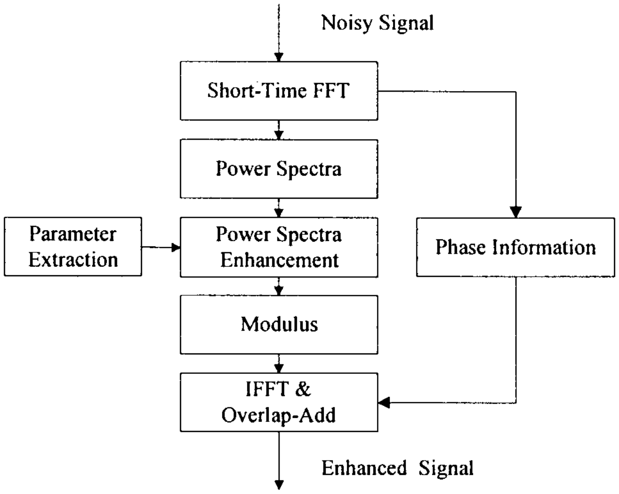

Fig. 1. General Block Diagram of the ANS technique

For the parameter extraction procedure three different approaches were used, one for validation of the technique and two based on the proposed sparse data estimators [5]:

(a) The first approach used the AMT of the noise-free signal in eq. (10). Although this method has no meaning in terms of speech enhancement, since the noise-free signal in unknown), it was used in order to show the validity of the proposed method. This method will hereafter been mentioned as the "Debug" method and denoted by 'D'.

(b) The second method was based on a statistical model for the estimation of the minimum spectral component in conjunction with eq. (9) [5]. This method will be referred to as the "Minima" method and denoted by 'M'.

(c) The third method tested was based on a clean speech AMT estimator in conjunction with eq. (10) [5]. This method will be referred to as the "Threshold" and will be denoted by 'T'.

3.2. Evaluation and Results

The ANS technique was evaluated using objective and subjective tests. In terms of the objective tests the Signal to Noise Ratio (SNR) and the Noise to Mask Ratio (NMR) [6] were used. Two subjective tests were also used. The first, at word level, was the Diagnostic Rhyme Test (DRT) [7] performed on Greek and English language speech data. For the Greek language tests 192 word-pairs were used uttered by 4 speakers and a total of 20 subjects were employed in the listening tests. For the English language tests 96 word-pairs, 2 speakers and 6 listeners were used. The second test, at the sentence level, was the Semantically Unpredictable Sentences (SUS) test [8] which was performed on Greek language speech data using 80 sentences based on 5 syntactical structures, 4 speakers and 20 listeners.

Results in terms of the objective SNR and NMR measures are presented in Fig. 2, for all the above tests, i.e. English DRT (E-DRT), Greek DRT (G-DRT) and SUS test, for initial SNR conditions -∝, -5, 0, and 5 dB, for the three approaches described in Section 3.1, i.e. 'D', 'T', 'M', while results for the noisy signal are also shown denoted by 'N'.

Intelligibility results are shown in Fig. 3 for using the same databases, initial conditions and enhancement approaches.

From the results in Fig. 2 and 3, several observations can be made:

(a) The results between the subjective and objective tests are generally in agreement.
(b) The best results are given by the "Debug" method, which was used to demonstrate the validity of the proposed enhancement technique. This method works also for initial SNR condition -∝ dB which shows that intelligible speech can be obtained by appropriate modulation of a noise signal
(c) Improvement was measured by the use of the two types of sparse-data estimators, with the "Threshold" one having a small advantage over the "Minima" for most conditions, especially for the NMR tests.
(d) In most cases the proposed ANS estimation methods achieved results close to the "Debug" method, with typical SNR improvement of 10 dB and NMR improvement of 20 dB.

In terms of the subjective tests the following observations can be made:

(a) For initial condition -∝ dB, the "Debug" method achieved scores of 72.22 % (for E-DRT), 85 % (for G-DRT) and 73.36 % (for SUS), indicating the validity of the proposed ANS enhancement technique.
(b) The "Debug" method achieved also intelligibility improvement for all other initial conditions, although, this improvements was smaller for the better initial SNRs. At SNR=-5 dB, this method produced improvement of up to 22 % for G-DRT, 38.89 % for E-DRT and 34.46 % for SUS.
(c) The proposed estimators "Threshold" and "Minima" achieved intelligibility improvement in almost all initial conditions and tests. This improvement was, however, lower than that achieved by the "Debug" method, indicating that there is further space for improving the parameter estimation procedures of the ANS technique. 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 improvement 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.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. of the IEEE, vol. 67, no. 12, pp. 1586-1604, Dec. 1979*

[2] E. Zwicker and H. Fastl, Psychoacoustics, Facts and Models, Springer-Verlag Berlin Heidelberg 1990

[3] J. D. Johnston, "Transform Coding of Audio Signal using Perceptual Noise Criteria", *IEEE Jour. Select. Areas in Communications, vol. 6, No. 2, pp. 314-323, Feb. 1988*

[4] Tsoukalas D., J. Mourjopoulos, and G. Kokkinakis, "Low bit-rate speech coding by perceptually optimized noise excitation modulation", *Signal Processing, 56 (1997), pp. 77-89*

[5] D. E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech Enhancement based on Audible Noise Suppression (ANS)", *IEEE Trans. Speech, Audio Proc, acceptedfor publication, 1997*

[6] J. Herre, E. Eberlein, H. Schott and K. Brandenburg, "Advanced Audio Measurement System using Psychoacoustic Properties", *in Proc. 92nd AES Convention*, Mar. 1992

[7] S. Meister, "The Diagnostic Rhyme Test (DRT): An Air Force Implementation", *RADC-TR-78-129, AD-A060917, 1978*

[8] 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, April 1991*

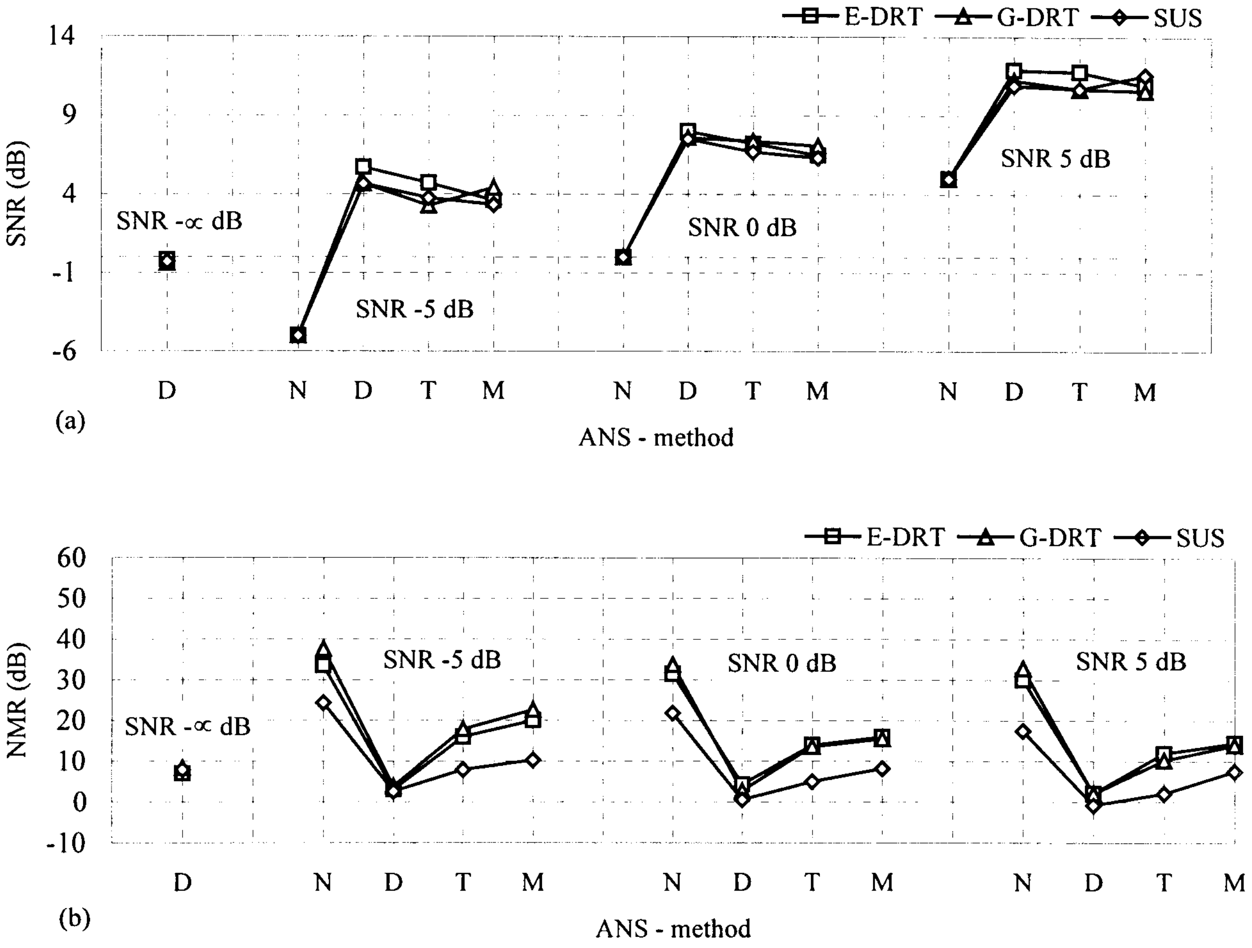

Fig. 2. Objective ANS method evaluation for the English language speech data DRT (E-DRT), the Greek language speech data DRT (G-DRT), and the SUS test. Initial SNR condition is also indicated for each curve. The horizontal axis denotes the ANS - method, where 'N' stands for the noisy signal, 'D' for the "Debug" method, 'T' for the "Threshold" approach and 'M' for the "Minima" approach (see also the text). (a) SNR performance, (b) NMR performance.

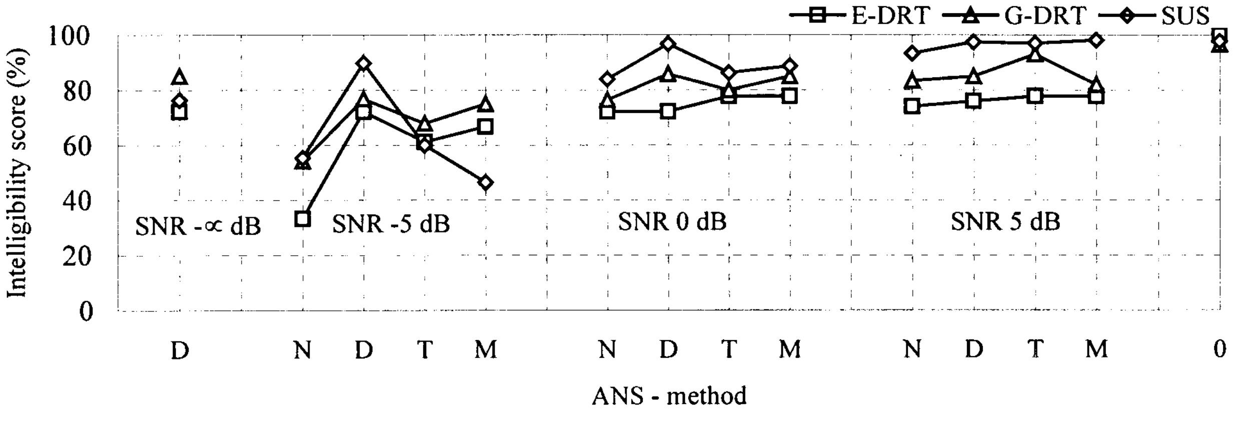

Fig. 3. Intelligibility scores for the English language speech data DRT (E-DRT), the Greek language speech data DRT (G-DRT) and the SUS test. Initial SNR condition is also indicated for each curve. The horizontal axis denotes the processing category, where 'N' stands for the noisy signal, and 'O' for the noise-free signal.