A NOVAL APPROACH TO SPEECH ENHANCEMENT USING ADAPTIVE MULTI-BAND LOGARITHMIC ENVELOPE EXPANSION TECHNIQUE

S.F.Bahgat *, A.Rohim *, S.Ghoniemy *

ABSTRACT

The speech enhancement problem covers a broad spectrum of constraints, applications and issues. The principal application areas of speech enhancement are noise reduction for human listening, preprocessing for recognition systems and preprocessing for linear predictive coding.

This paper introduces a modified technique based on Multi-band envelope expansion technique called adaptive multi-band logarithmic envelope expansion technique. In this technique a threshold level varying according to the change of input signal-to-noise ratio is used, and the nonlinear function used to expand the envelope is a logarithmic function. In addition, the use of non-uniform frequency spacing gives better enhancement since the majority of filters are concentrated in the actual speech band (up to 1.5 KHz) with few filters in the remaining band (1.5 up to 4 KHz).

The simulation results show that the new technique gives a high dynamic range of input signal-to-noise ratio (up to - 8db), and a high improvement factor. The intelligibility of the proposed technique is considered as a subject for further investigation.

1. Introduction :

Over many years speech enhancement has occupied an important area of research in digital signal processing. The objective of speech enhancement is to improve the overall quality of speech, to increase intelligibility, to reduce listener fatigue, and to improve signal-to-noise ratio. Many techniques have been developed for speech enhancement based on : Spectral Subtraction, Non-linear Spectral Processing, Wiener Filtering, Comb Filtering, Adaptive Noise Cancelling, Speech Modelling, All-pole Modelling, or Pole-Zero Modelling of Speech. In this paper, we consider only a subset of possible topics, specially the enhancement of speech degraded by uncorrelated additive noise.

Many studies were performed to establish some cues of envelope of speech waveform. The results of these studies establish that cues for speech recognation are available in time intensity envelope of speech and the envelope processing is sufficient for speech enhancement specially increasing the value of the signal-to-noise ratio as a criterion for enhancing speech signal.

Flangan [1] proved that the envelope at the output of a band limited signal is equal to the magnitude of the short-time Fourier transform of the signal for that channel. Due to the study of Flangan and the problems associated with processing the signal in frequency domain, this research was directed towards the envelope of the signal in the time domain. This is a far more attractive approach for real-time implementation.

* Guidance Department, M.T.C., Cairo, Egypt.

Supporting this approach, Mourjopoulos and Hammond [2] proposed a model to enhance reverberant speech using an envelope convolution technique. Their approach was used to enhance a speech signal degraded by room reverberation. The work of Mourjopoulos and Hammond achieved only limited success due to difficulties in finding the inverse of the room transfer function and computational limitations of their system which restricted the number of bands to four.

Multi-band envelope expansion technique (MBEET) [3] and [4] is the only technique based on this idea. In this technique the input speech signal which is degraded by additive white gaussian noise is band-pass filtered into a number of contiguous frequency bands which cover speech frequency range. The envelope of speech signal is obtained using a full-wave rectifier and low-pass filter. The envelope in each band is then expanded using fractional power coefficients. Finally, the speech signal is reconstructed using the expanded envelope and the output from the band-pass filters.

In this paper we introduce a modified technique based on (MBEET), called adaptive logarithmic envelope expansion technique (ALEET). The basic difference between this technique and MBEET is that :

(1) In ALEET, a variable threshold level is used, which depends on the variation of input signal-to-noise ratio, but in MBEET a fixed threshold is used.

(2) In ALEET, a non-linear logarithmic function is used which gives better smoothing to enhanced signal than expansion function with fractional coefficients used in MBEET.

(3) MBEET uses a fixed number of filters (20 BPF filters) with uniform frequency over speech bandwidth, but in modified technique only 15 BPF filters are used with a non uniform frequency spacing which gives better enhancement because the majority of filters are concentrated in actual speech band-width (up to 1.5 KHz). As a simulation result, the modified technique gives a high dynamic range of input signal-to- noise ratio (up to - 8db) and higher improvement factor . The intelligibility of modified technique is not measured, and is considered as a topic for future ivestigation.

2. Analysis of Adaptive Logarithmic Envelope Expansion Technique (ALEET):

The analysis of this technique, was carried out assuming :

(1) The speech signal and noise are uncorrelated.

(2) The noise is additive white gaussian noise with zero mean and variance σ^2.

The basic concepts of the adaptive logarithmic envelope expansion technique is shown in Fig.1. and consists of the following stages:

Signal filtering (BPF), Envelope extraction (rectifier and LPF), Threshold level calculation (TH), Adaptive envelope expansion by logarithmic function and signal reconstruction.

Initially, the speech waveform is sampled, periodically in time, to produce a sequence of samples $x(n)$. An important point which is often overlooked in discussion of sampling is that even though the signal waveform may have a band-limited spectrum, the signal may be corrupted by wideband noise, prior to analog-to-digital conversion. In such cases, the signal plus noise combination should be filtered with an analog low-pass filter which cuts off sharply above the Nyquist frequency, so that no image of the high frequencies is aliased into the base-band. The digital signal $x(n)$ is the input speech signal to the system degraded by statistically uncorrelated additive white noise.

$$x(n) = s(n) + v(n) \tag{1}$$

where $s(n)$ is the speech signal and $v(n)$ is the additive white noise.

2.1 Signal Filtering:

The first processing step is the decomposition of the degraded speech signal into contiguous frequency bands. This is achieved using infinite impulse response (IIR) band-pass filters.

A non-uniform filter bank based on logarithmic distribution is proposed (as the logarithmic frequency scale is often justified from a human auditory perception point of view). Thus for a set of Q band-pass filters with center frequencies f_i and bandwidth b_i the following distrubution is proposed :

$$b_1 = C \tag{2}$$

$$b_i = \alpha b_{i-1} \tag{3}$$

and

$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{b_i - b_1}{2} \tag{4}$$

where C and f_1 are the arbitrary bandwidth and center frequency of the first band-pass filter. α is the logarithmic factor. Fig.2. shows the specification of 4-channel, 7-channel, and 12-channel filter bank. Also, an alternative criterian for a non uniform filter bank is to use the critical band width scale directly. The general shape of the critical bands as a function of frequency is shown in Fig.3.

Now, we assume that h(n) is the unit sampled response of each band-pass filter. Then, the output can be represented by

$$y_{i1}(n) = h(n) * x(n) \tag{5}$$

The output from ith band-pass filter can be represented by the following difference equation:

$$y_{i1}(n) = \sum_{k=1}^{N} a_{ik} y_{i1}(n-k) + \sum_{r=0}^{M} b_{ir} x(n-r) \quad , i = 1, ..., n \tag{6}$$

To facilitate the operation in real-time using small-scale microprocessor systems, the band-pass filters were restricted to second order. For the final implementation, the second order Butterworth filter was reduced (without appreciable loss of performance) to an even simpler transfer function of the form [5]:

$$H(z) = \frac{1}{1 - 2r(cos\theta)z^{-1} + r^2 z^{-2}} \tag{7}$$

where $\theta = 2\pi f_i T$, using equation (4) we get:

$$\theta = 2\pi(f_1 + \sum_{j=1}^{i-1} b_j + \frac{b_i - b_1}{2})T \tag{8}$$

The impulse response corresponding to the equation (7) is [6]:

$$h(n) = \frac{r^n sin(n+1)\theta}{sin(\theta)} u(n) \tag{9}$$

where $u(n)$ represents the unit step.

Then, the simplified band-pass filter design is implemented as follows:

$$y_{i1}(n) = 2rcos(2\pi f_i T)y_{i1}(n-1) - r^2 y_{i1}(n-2) + x(n) \tag{10}$$

where :

$x(n)$ is the input to the digital band-pass filter,

y_{i1}(n) is the output from ith digital band-pass filter,

r is the coefficient of the band-pass filter whose value depend on the center frequency and the bandwidth of the digital filter.

T is the sampling interval, and

f_i is the central frequency of the ith band-pass filter.

The number of filters used to cover the required speech bandwidth and the center frequencies of these filters and their bandwidths are governed basically by the spectral characteristics of the speech signal. However, special attention has to be paid to the complexity of the system in order to allow real-time implementation. Experimentally, this work started with 20 similar uniformally-spaced filters bank as S.F.Bahgat [7], changed the distribution, decreased the number of filters, and finally changed their bandwidths , taking into consideration the fact that the energy is concentrated from 100 Hz up to 1.5 KHz. The final choice of the filter bank center frequencies and bandwidths is shown in table (1) and the responses of the individual filters is plotted in Fig.4.

Table (1) The center frequencies and bandwidths for filter bank

Filter Number i	Center frequency f_i	Bandwidth b_i
1	250 Hz	100 Hz
2	350 Hz	100 Hz
3	450 Hz	100 Hz
4	550 Hz	100 Hz
5	650 Hz	100 Hz
6	750 Hz	100 Hz
7	850 Hz	100 Hz
8	950 Hz	100 Hz
9	1250 Hz	200 Hz
10	1375 Hz	200 Hz
11	1625 Hz	200 Hz
12	1875 Hz	200 Hz
13	2250 Hz	400 Hz
14	2750 Hz	400 Hz
15	3500 Hz	800 Hz

As a conclusion, we get:

(1) The overall response of the bank of the filters is flat as shown in Fig.4.

(2) The speech band was divided into 4 regions, the first sub-band (below 1 KHz), 8 filters were placed with narrow band width (100 Hz). The second sub-band (between 1 KHz up to 2 KHz), 4 filters were placed with bandwidth equal to 200 Hz . The third sub-band (from 2 KHz up to 3 KHz) two filters were placed with bandwidth equal to 400 Hz. Finally, one filter was used for the fourth sub-band (from 3 KHz up to 3.5 KHz) with bandwidth equal to 800 Hz.

2.2 Envelope Extraction:

There are many methods which can be used to extract the envelope of the speech signal [2]. One method uses the short-time spectral amplitude in the frequency domain as equivalent to the envelope in the time domain. This method is not suitable for real-time application due to the requirement for windowing, overlapping, FFT. etc. Another method uses the following three steps: squaring, low pass filtering, and square root. This method is suitable for real-time application. However, it was modified to include only a full wave rectifier (absolute value) and low pass filtering.

(a) Full-wave rectifier stage:

The signal is full wave rectified and represented as :

$$y_{i2}(n) = |y_{i1}(n)| \tag{11}$$

Due to the fact that the absolute value of the signal does not change its power, the power of the signal and the noise at the output from this stage are equal to their powers at the input.

(b) Low pass filtering stage:

The low-pass filter is implemented using an IIR design. Also, to facilitate the operation in real-time using a single-chip microprocessor system, a first-order low-pass filter is employed. The simplified low-pass filter equation can be written as follows:

$$y_{i3}(n) = By_{i3}(n-1) + y_{i2}(n) \tag{12}$$

where the coefficient B is used to adjust the cutoff frequency of the filter.

2.3 Envelope Expansion:

The output from the low-pass stage is passed through the logarithmic envelope expansion stage. The envelope expansion is based on the following formula:

$$y_{i4} = \frac{y_{i3}/A}{|(1 - log(y_{i3}/A))|} \tag{13}$$

Where A is the adaptive threshold level which is proportional to the noise level.

The key of the above formula is the nonlinearity property of the logarithmic function $log y_{i3}$. The objective of this function is that it compresses the extracted envelope less than the threshold A while expanding the extracted envelope greater than A. There is a smooth transition region in the neighberhood of A and the transition region can be controlled by changing the argument of the logarithmic function.

2.4 Threshold Level Calculation :

To achieve an expression for threshold level used with speech enhancement algorithm, it is assumed that there is only one channel as shown in Fig.1. and the input signal is assumed to be one tone signal of the form:

$$X(t) = Acos\omega_o t \tag{14}$$

The noise is assumed to be white gaussian noise with zero mean and variance η. For improvement of input signal to noise ratio, the threshold level can be written as:

$$\sqrt{\overline{Y^2(t)}}|_{noise\ \ only} \leq TH \leq \sqrt{\overline{Y_n^2(t)}}|_{signal+noise} \tag{15}$$

The noise is white gaussian noise and it can be represented by :

$$n(t) = n_c(t)cos\omega_0 t + n_s(t)sin\omega_0 t \tag{16}$$

To calculate noise power (noise only)

$$\overline{Y^2}(t) = n_c^2(t) + n_s^2(t) \tag{17}$$

power of noise at output of BPF is

$$\overline{Y^2}(t) = 2\eta B \tag{18}$$

where B is band width of BPF.
For signal plus noise

$$\overline{Y_n^2}(t) = A^2 + 2An_c(t) + n_c^2(t) + n_s^2(t) \tag{19}$$

taking the average value of both sides we get:

$$\overline{Y_n^2}(t) = A^2 + 2\eta B \tag{20}$$

The signal to noise ratio (SNR) at input can be calculated as follows:

$$SNR = \frac{\frac{1}{2}A^2}{2\eta B} \tag{21}$$

subsituting by (21) in (18) and (20) we get:

$$\sqrt{2\eta B} \leq TH \leq \sqrt{2\eta B}\sqrt{1 + 2SNR} \tag{22}$$

Then the threshold level has the form :

$$TH = \frac{1}{2}\sqrt{2\eta B}[1 + \sqrt{1 + 2SNR}] \tag{23}$$

From the equation (22) it is clear that the threshold level increases as SNR of the input decreases and decreases as input SNR increases as shown in Fig.5.

2.5 Signal Reconstruction:

The speech signal is reconstructed using the output from the band-pass digital filter and the expanded envelope. The reconstructed signal can be represented by:

$$y_{i5}(n) = y_{i4}(n) y_{i1}(n) \tag{24}$$

The final output from the system is calculated by summing the reconstructed signals from all channels as:

$$y(n) = \sum_{i=1}^{15} y_{i5}(n) \tag{25}$$

where:

$y(n)$ is the final output from the system, and

$y_{i5}(n)$ is the reconstructed signal from ith channel. Figure 6 shows the multi-channel logarithmic envelope expansion technique.

3. Computer Simulation and Discussion:

The proposed system (refer to Fig.6.) has been simulated on a main frame computer (using Fortran-77) under the same simplifications which are needed for real-time realization. These simplifications employ second order digital band-pass filters and first order low-pass filters and restrict the system to 15 channels. Fig.7. shows the flowchart of the algorithms. The objective of this simulation is to achieve the optimum system parameters (coefficient of band-pass filter, coefficient of low-pass filter, and the center frequencies of the band-pass filters). Trials were performed utilizing a segment of speech signal and a group of reconstructed speech signals were obtained. Fig.8. shows the reconstructed speech signal using the proposed technique at different values of input SNR. Fig.9. shows one example of the three dimensional representation of the enhanced speech signal at zero input SNR. Simulation results proved that :

a- The nonlinear distribution of the 15 BPF with their non-uniform critical bandwidths is the best choise to give better enhancement of speech signal.

b- The nonlinear logarithmic envelope expansion gives wide dynamic range for improving output S/N ratio.

c- Adaptive threshold level gives good estimation to the noise canceling without introducing any distortion for the reconstructed speech signal.

d- The logarithmic expansion is very useful at neighbourhood of the threshold level value and linear expansion is used for the speech signals having a value greater than this threshold level.

Conclusions and Future work.

In this paper new algorithms based on adaptive multiband logarithmic envelope expansion have been developed to improve the performance of speech enhancement systems. These algorithms have the potential to improve signal-to-noise ratio, and are simple enough to allow real-time implementation on low-cost microprocessors. Simulation results proved that the proposed technique gives a high dynamic range of input S/N (up to - 8 db) and better improvement factor. However, further work is required to obtain a complete probabilistic analysis of the non-linear expansion scheme and to derive a general form for the output signal-to-noise ratio and improvement factor for all values of the input signal-to-noise ratio.

It would also be interesting to measure the intelligibility using the Diagnostic Rhyme Test to provide further insight into the speech enhancement processes.

A further general question concerns the effects of phase. In this research we have dealt mainly with the envelope and it would be interesting to see the effect of the phase or try to enhance the phase before using it.

References:

[1] Flangan,J.," Parametric Coding of speech Spectra," Journal of the Acoustical Society of America, Vol.68, No.2, Page.412-419, August, 1980.

[2] Mourjopoulos,J. and Hammond,J.,"Modeling and Enhancement of the Reverberant Speech Using an Envelope Covolution Method." Proc. of the International Conference on Acoustics Speech and Signal Processing, Page.1144-1147, 1983.

[3] Bahgat,S.F.,"Real-time Processing Enhancement System Using Envelope Expansion Technique,"Ph.D thesis, Illinois Institute of Technology, Chicago, December 1989.

[4] Bahgat,S.F.,Ghoniemy,S. and Fahmy,M.A.,"An Efficient Approach for Speech Enhancement Using Multi-band Envelope Expansion technique," Ain Shams University, Engineering Bulletin, Vol.28 No.3, Sep.1993, Page.139-150.

[5] Rabiner,L. and Gold,B.,"Theory and Applications of Digital Signal processing." Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[6] Carmichael,R., and Smith, E." Mathematical Tables and Formulas," Dover Publication, Inc.,New York, 1931.

[7] Bahgat,S.F.,Ghoniemy,S. and Fahmy,M.A.,"Real-Time Speech Enhancement System Using TMS320-C25," Ain Shams University, Engineering Bulletin, Vol.28, No.3, Sep.1993, Page.129-138.

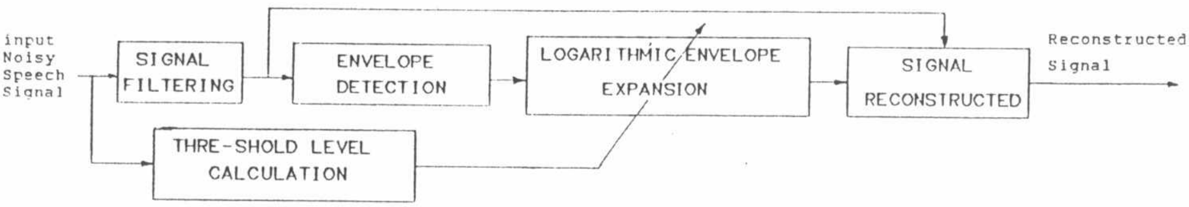

input Noisy Speech Signal
SIGNAL FILTERING
ENVELOPE DETECTION
LOGARITHMIC ENVELOPE EXPANSION
SIGNAL RECONSTRUCTED
Reconstructed Signal
THRE-SHOLD LEVEL CALCULATION

Fig. 1 The basic concepts of the adaptive logarithmic envelope expansion technique for One channel.

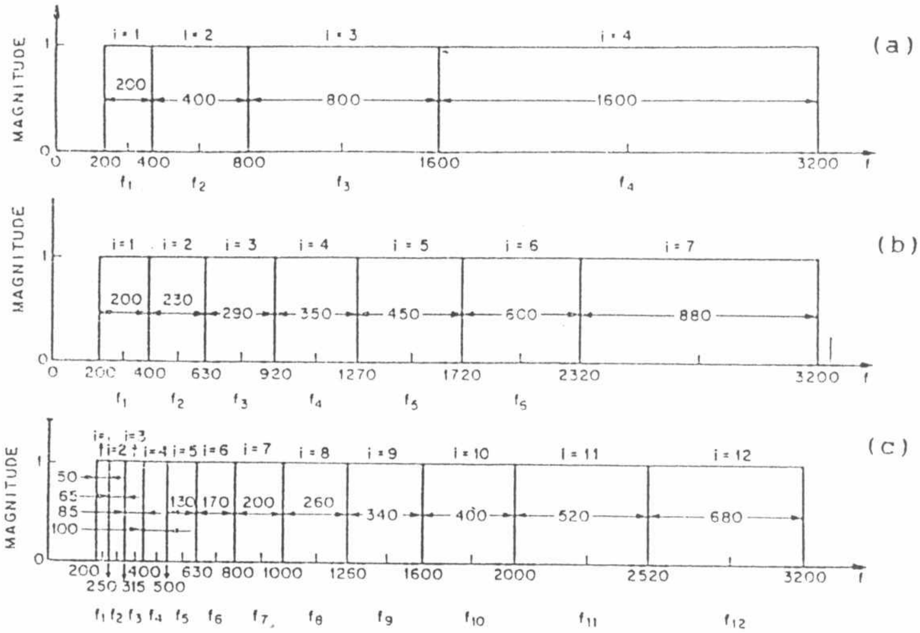

(a)
MAGNITUDE
i = 1
i = 2
i = 3
i = 4
200
400
800
1600
0
200 400 800 1600 3200 f
f1 f2 f3 f4
(b)
MAGNITUDE
i = 1 i = 2 i = 3 i = 4 i = 5 i = 6 i = 7
200 230 290 350 450 600 880
0 200 400 630 920 1270 1720 2320 3200 f
f1 f2 f3 f4 f5 f5
(c)
MAGNITUDE
i=1 i=2 i=3 i=4 i=5 i=6 i=7 i=8 i=9 i=10 i=11 i=12
50
65
85
100
130 170 200 260 340 400 520 680
200 250 315 400 500 630 800 1000 1250 1600 2000 2520 3200 f
f1 f2 f3 f4 f5 f6 f7 f8 f9 f10 f11 f12

Fig. 2 Ideal specification of (a) 4-channel filter bank, (b) 7-channel filter bank, (c) 12-channel filter bank.

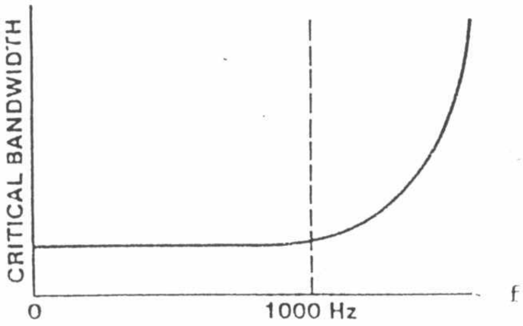


Fig. 3 variation of bandwidth with frequency for the perceptually based critical band scale.

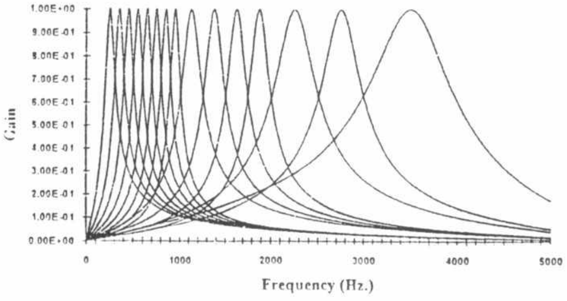


Fig. 4 Response of the filter bank . (number of filters = 15).

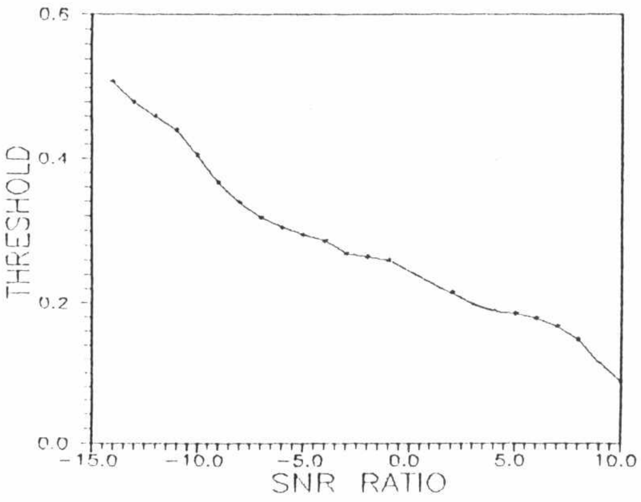


Fig. 5 Threshold level versus input signal-to-noise ratio.

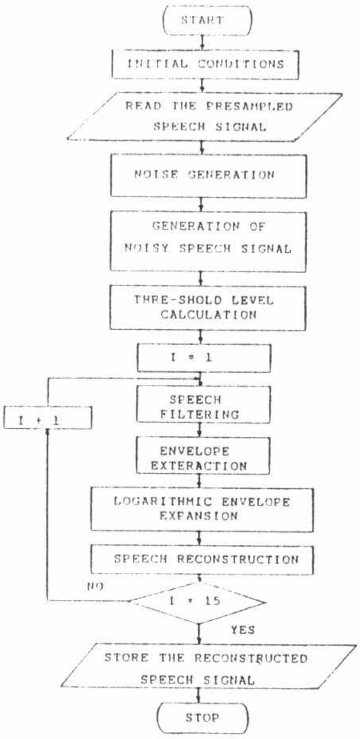

START
INITIAL CONDITIONS
READ THE PRESAMPLED SPEECH SIGNAL
NOISE GENERATION
GENERATION OF NOISY SPEECH SIGNAL
THRE-SHOLD LEVEL CALCULATION
I = 1
I + 1
SPEECH FILTERING
ENVELOPE EXTERACTION
LOGARITHMIC ENVELOPE EXPANSION
SPEECH RECONSTRUCTION
NO
I = 15
YES
STORE THE RECONSTRUCTED SPEECH SIGNAL
STOP

Fig. 7 The flowchart of adaptive multi-channel logarithmic envelope expansion technique.

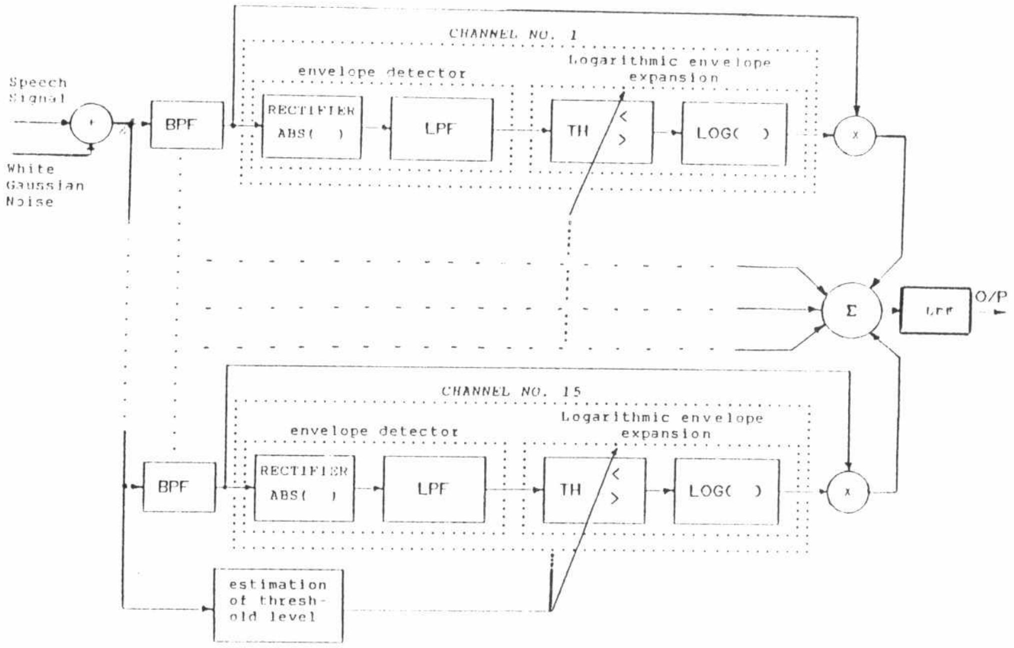
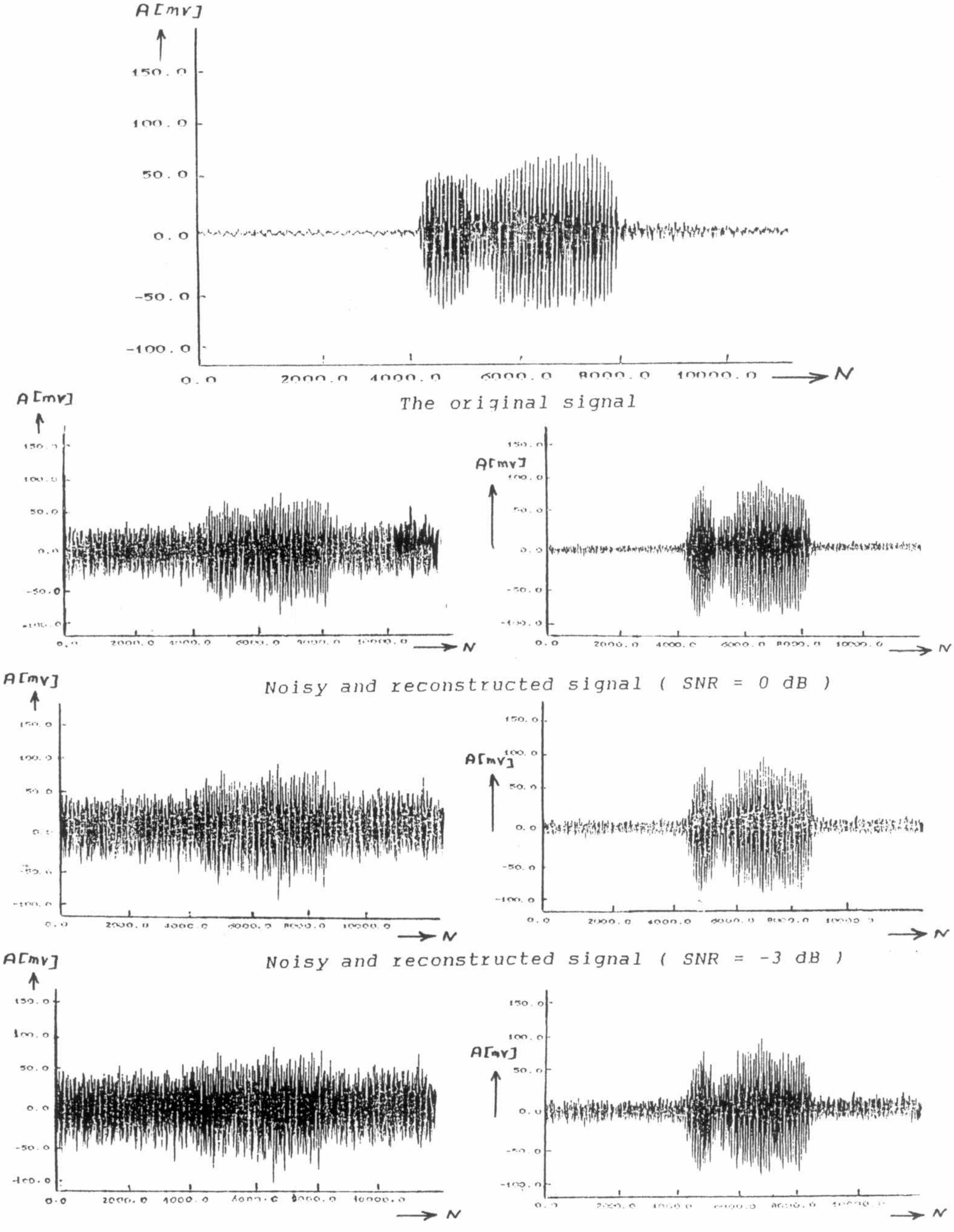


Fig. 6 Block diagram of adaptive multi-channel logarithmic envelope expansion technique.



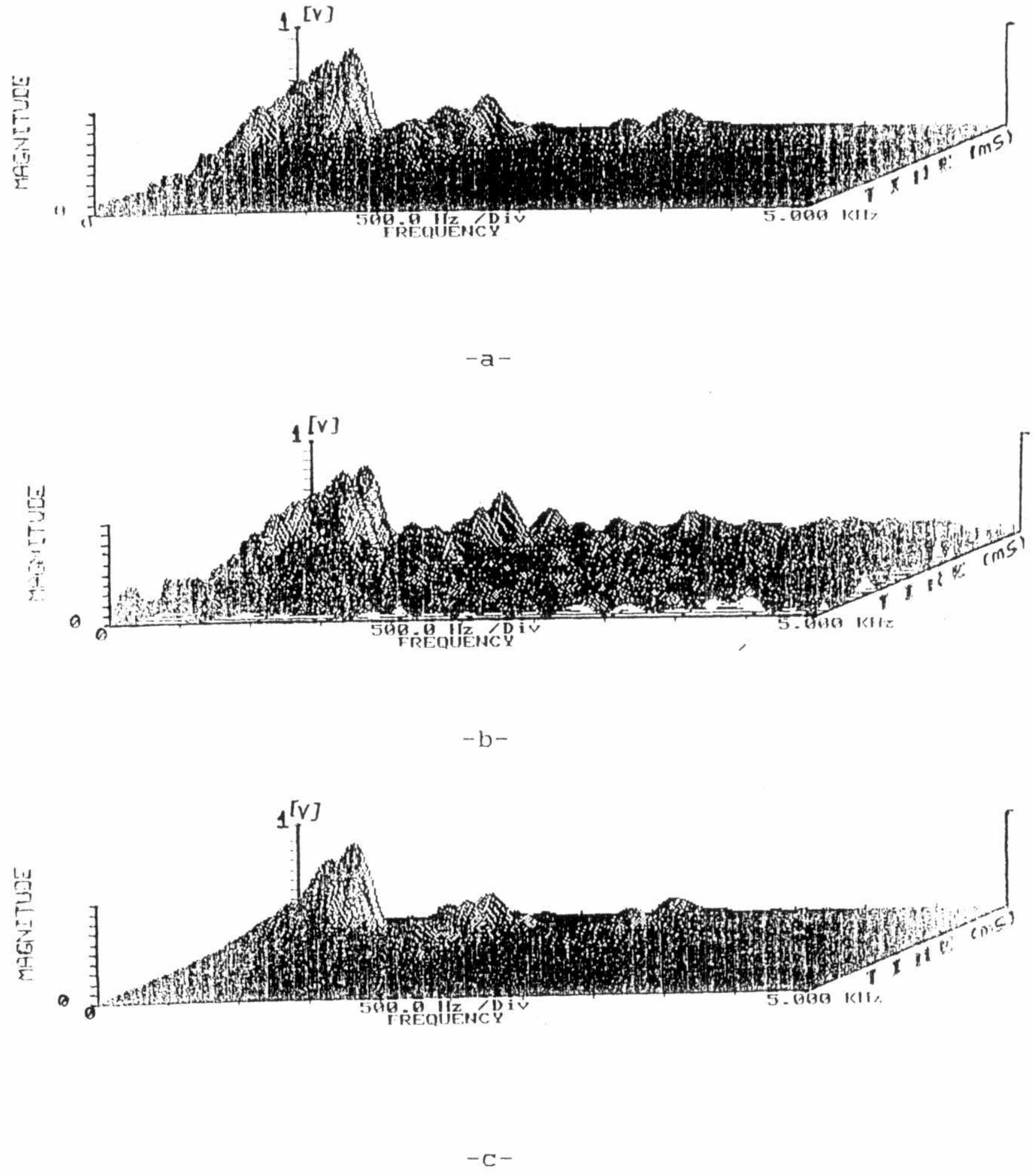
The original signal

Noisy and reconstructed signal (SNR = 0 dB)

Noisy and reconstructed signal (SNR = -3 dB)

Noisy and reconstructed signal (SNR = -5 dB)

Fig. 8 The reconstructed (enhanced) speech signal using the proposed technique at different values of input SNR.



-a-

-b-

-c-

Fig. 9 a) Original speech signal.

b) Noisy speech signal at SNR = 0 dB.

c) Reconstructed speech signal.